TESI DOCTORAL

Títol Modelado de la cualidad de la voz para la síntesis del habla expresiva

Realitzada per Carlos Manuel Monzo Sánchez

en el Centre Escola Tècnica Superior d'Enginyeria Electrònica i Informàtica La Salle

i en el Departament Comunicacions i Teoria del Senyal

Dirigida per Dr. Joan Claudi Socoró Carrié
Dr. Ignasi Iriondo Sanz

C. Claravall, 1-3
08022 Barcelona
Tel. 936 022 200
Fax 936 022 249
E-mail: urlsc@sec.url.es
www.url.es

A Alba.

Resumen

Esta tesis se realiza dentro del marco de trabajo existente en el grupo de investigación *Grup de Recerca en Tecnologies Mèdia* (GTM) de *Enginyeria i Arquitectura La Salle*, con el objetivo de dotar de mayor naturalidad a la interacción hombre-máquina. Para ello nos basamos en las limitaciones de la tecnología empleada hasta el momento, detectando puntos de mejora en los que poder aportar soluciones. Debido a que la naturalidad del habla está íntimamente relacionada con la expresividad que esta puede transmitir, estos puntos de mejora se centran en la capacidad de trabajar con emociones o estilos de habla expresivos en general.

El objetivo último de esta tesis es la generación de estilos de habla expresivos en el ámbito de sistemas de Conversión de Texto en Habla (CTH) orientados a la Síntesis del Habla Expresiva (SHE), siendo posible transmitir un mensaje oral con una cierta expresividad que el oyente sea capaz de percibir e interpretar correctamente. No obstante, este objetivo implica diferentes metas intermedias: conocer las opciones de parametrización existentes, entender cada uno de los parámetros, detectar los pros y contras de su utilización, descubrir las relaciones existentes entre ellos y los estilos de habla expresivos y, finalmente, llevar a cabo la síntesis del habla expresiva. El propio proceso de síntesis implica un trabajo previo en reconocimiento de emociones, que en sí mismo podría ser una línea completa de investigación, ya que muestra la viabilidad de usar los parámetros seleccionados en la discriminación de estos y aporta el conocimiento necesario para extraer los modelos que pueden ser usados durante el proceso de síntesis.

La búsqueda del incremento de la naturalidad ha implicado una mejor caracterización del habla emocional o expresiva, con lo que para ello se ha investigado en parametrizaciones que pudieran llevar a cabo este cometido. Estos son los parámetros de Cualidad de la Voz —*Voice Quality*— (VoQ), que presentan como característica principal que son capaces de caracterizar individualmente el habla, identificando cada uno de los factores que hacen que sea única. Los beneficios potenciales, que este tipo de parametrización puede aportar a la interacción natural, son de dos clases: el

reconocimiento y la síntesis de estilos de habla expresivos. La propuesta de la parametrización de VoQ no pretende sustituir a la ya empleada prosodia, sino todo lo contrario, trabajar conjuntamente con ella para mejorar los resultados obtenidos hasta el momento.

Una vez realizada la selección de los parámetros se plantea el modelado de la VoQ, es decir, la metodología de análisis y de modificación de forma que cada uno de ellos pueda ser extraído a partir de la señal de voz y posteriormente modificado durante la síntesis. Asimismo, se proponen variaciones para los parámetros implicados y tradicionalmente utilizados, adaptando su definición al contexto del habla expresiva. A partir de aquí se pasa a trabajar en las relaciones existentes con los estilos de habla expresivos, presentando finalmente la metodología de transformación de estos últimos, mediante la modificación conjunta de VoQ y prosodia, para la SHE en un sistema de CTH.

PALABRAS CLAVE: Cualidad de la voz, síntesis del habla expresiva, conversión de texto en habla, reconocimiento de emociones, tecnologías del habla.

Resum

Aquesta tesi es realitza dins del marc de treball existent en el grup d'investigació *Grup de Recerca en Tecnologies Mèdia* (GTM) d'*Enginyeria i Arquitectura La Salle*, amb l'objectiu de dotar de major naturalitat a la interacció home-màquina. Per això ens basem en les limitacions de la tecnologia emprada fins al moment, detectant punts de millora en els que poder aportar solucions. Donat que la naturalitat de la parla està íntimament relacionada amb l'expressivitat que aquesta pot transmetre, aquests punts de millora es centren en la capacitat de treballar amb emocions o estils de parla expressius en general.

L'objectiu últim d'aquesta tesi és la generació d'estils de parla expressius en l'àmbit de sistemes de Conversió de Text a Parla (CTP) orientats a la Síntesi de la Parla Expressiva (SPE), essent possible transmetre un missatge oral amb una certa expressivitat que l'oient sigui capaç de percebre i interpretar correctament. No obstant, aquest objectiu implica diferents metes intermitges: conèixer les opcions de parametrització existents, entendre cadascun dels paràmetres, detectar els pros i contres de la seva utilització, descobrir les relacions existents entre ells i els estils de parla expressius i, finalment, portar a terme la síntesi de la parla expressiva. Donat això, el propi procés de síntesi implica un treball previ en reconeixement d'emocions, que en si mateix podria ser una línia complerta d'investigació, ja que aporta el coneixement necessari per extreure models que poden ser usats durant el procés de síntesi.

La cerca de l'increment de la naturalitat ha implicat una millor caracterització de la parla emocional o expressiva, raó per la qual s'ha investigat en parametritzacions que poguessin portar a terme aquesta comesa. Aquests són els paràmetres de Qualitat de la Veu —*Voice Quality*— (VoQ), que presenten com a característica principal que són capaços de caracteritzar individualment la parla, identificant cadascun dels factors que fan que sigui única. Els beneficis potencials, que aquest tipus de parametrització pot aportar a la interacció natural, són de dos classes: el reconeixement i la síntesi d'estils de parla expressius. La proposta de la parametrització de VoQ no

pretén substituir a la ja emprada prosòdia, sinó tot el contrari, treballar conjuntament amb ella per tal de millorar els resultats obtinguts fins al moment.

Un cop realitzada la selecció de paràmetres es planteja el modelat de la VoQ, és a dir la metodologia d'anàlisi i de modificació, de forma que cadascun d'ells pugui ser extret a partir de la senyal de veu i posteriorment modificat durant la síntesi. Així mateix, es proposen variacions pels paràmetres implicats i tradicionalment utilitzats, adaptant la seva definició al context de la parla expressiva. A partir d'aquí es passa a treballar en les relacions existents amb els estils de parla expressius, presentant finalment la metodologia de transformació d'aquests últims, mitjançant la modificació conjunta de la VoQ y la prosòdia, per a la SPE en un sistema de CTP.

PARAULES CLAU: Qualitat de la veu, síntesi de la parla expressiva, conversió de text a parla, reconeixement d'emocions, tecnologies de la parla.

Abstract

This thesis is conducted on the existing working framework in the *Grup de Recerca en Tecnologies Mèdia* (GTM) research group of the *Enginyeria i Arquitectura La Salle*, with the aim of providing the man-machine interaction with more naturalness. To do this, we are based on the limitations of the technology used up to now, detecting the improvement points where we could contribute solutions. Given that the speech naturalness is closely linked with the expressivity communication, these improvement points are focused on the ability of working with emotions or expressive speech styles in general.

The final goal of this thesis is the expressive speech styles generation in the field of Text-to-Speech (TTS) systems aimed at Expressive Speech Synthesis (ESS), with the possibility of communicating an oral message with a certain expressivity that the listener will be able to correctly perceive and interpret. Nevertheless, this goal involves different intermediate aims: to know the existing parameterization options, to understand each of the parameters, to find out the existing relations among them and the expressive speech styles and, finally, to carry out the expressive speech synthesis. All things considered, the synthesis process involves a previous work in emotion recognition, which could be a complete research field, since it shows the feasibility of using the selected parameters during their discrimination and provides with the necessary knowledge for the modelling that can be used during the synthesis process.

The search for the naturalness improvement has implied a better characterization of the emotional or expressive speech, so we have researched on parameterizations that could perform this task. These are the Voice Quality (VoQ) parameters, which main feature is they are able to characterize the speech in an individual way, identifying each factor that makes it unique. The potential benefits that this kind of parameterization can provide with natural interaction are twofold: the expressive speech styles recognition and the synthesis. The VoQ parameters proposal is not trying to replace prosody, but working altogether to improve the results so far obtained.

Once the parameters selection is conducted, the VoQ modelling is raised (i. e. analysis and modification methodology), so each of them can be extracted from the voice signal and later on modified during the synthesis. Also, variations are proposed for the involved and traditionally used parameters, adjusting their definition to the expressive speech context. From here, we work on the existing relations with the expressive speech styles and, eventually we show the transformation methodology for these ones, by means of the modification of VoQ and prosody, for the ESS in a TTS system.

KEYWORDS: Voice quality, expressive speech synthesis, text-to-speech, emotion recognition, speech technologies.

Agradecimientos

Son muchas las personas que de un modo u otro han participado en esta tesis, a las que quiero dar mi sincero agradecimiento.

Primero de todo quiero agradecer a mis padres todo lo que han hecho para que yo pueda estar escribiendo hoy estas líneas. Gracias a su apoyo, el proyecto que fue un día iniciar unos estudios superiores culmina hoy con la realización de esta tesis. También quiero agradecer al resto de mi familia, hermana, abuelos, tíos y primos, los ánimos que siempre me han dado para no dejar nunca de esforzarme.

A Alba, que desde el principio me ha dado todo su apoyo, sin dejar ni un momento de creer en mí, y por quien todo esfuerzo vale la pena. Sus ánimos han hecho que el cansancio fuera ilusión y que los momentos de debilidad se convirtieran en ganas de seguir adelante.

Agradecer a mis amigos, especialmente a Íñigo Pérez, todos los buenos momentos que me han hecho pasar, por saber escuchar y por darme ánimos cuando quizás más se necesitaban.

A mis directores de tesis Ignasi Iriondo y Joan Claudi Socoró, por sus consejos, porque he aprendido de ellos al trabajar a su lado y porque no ha habido problema, tanto dentro como fuera de la tesis, en el que no me hayan intentado ayudar.

Quiero agradecer a José Antonio Morán que me animara a iniciar esta etapa, por la confianza que depositó en mí y toda la ayuda que me brindó en los primeros momentos.

A Santi Planet y Xavi Gonzalvo, porque desde el inicio de esta tesis he podido contar con ellos.

A Elisa Martínez, por permitirme iniciar esta etapa de mi vida, por buscar un hueco siempre que la necesité y por haber colaborado en parte de esta tesis.

A Francesc Alías porque siempre que ha podido ayudar en algo lo ha hecho y por haber participado activamente en mi formación profesional.

A Lluís Formiga y Jordi Adell, con los que he compartido mis últimos años de trabajo y quienes han colaborado en todo lo que se les ha necesitado.

A todos los que son y fueron mis compañeros y que de algún modo han colaborado en este trabajo: Alexandre Trilla, Àngel Calzada, Xavier Sevillano, Borja Martínez, Berta Martínez, José Antonio Montero, Rosa Maria Alsina, Germán Cobo, Diego Torres, David García, Javier Melechón y Ester Cierco.

A toda aquella persona que ha participado en las pruebas subjetivas para la evaluación de los resultados.

Finalmente, a todos los organismos e instituciones que han financiado los proyectos en los que he participado, que sin ellos no hubiera sido posible llevar a cabo esta tesis.

Índice general

Índice de figuras

Índice de tablas

Acrónimos

AH	Ruido de Aspiración —*Aspiration Noise*—
ANOVA	Análisis de Varianza —*ANalysis Of VAriance*—
AQ	Cociente de Amplitud —*Amplitude Quotient*—
AR	Autorregresivo
AT	Tensión Abductora —*Adductive Tension*—
ATR	*Advanced Telecommunications Research Institute International*
BW	*Backward Elimination*
CBR	Razonamiento Basado en Casos —*Case Based Reasoning*—
CMOS	Nota Media de Opinión sobre las Comparaciones —*Comparison Mean Opinion Score*—
CQ	Cociente de Cierre —*Closed Quotient*—
CTH	Conversión de Texto en Habla
DI	Diplofonía —*Diplophonia*—
do1000	*Drop-off of Spectral Energy above 1000 Hz*
EE	Potencia de Excitación —*Excitation Strength*—
EGG	Electroglotógrafo
EMA	Articulografía Electromagnética —*Electromagnetic Articulagraphy*—
ESS	*Expressive Speech Synthesys*
FW	*Forward Selection*

GA	Área Glótica —*Glottal Area*—
GER	Tasa de Errores Grandes —*Gross Error Rate*—
GNE	*Glottal-to-Noise Excitation Ratio*
GOG	Gradiente de Apertura Glótica —*Glottal Opening Gradient*—
GPMER	*Gross Pitch Marks Error Rate*
GTM	*Grup de Recerca en Tecnologies Mèdia*
GW	Anchura Glótica —*Glottal Width*—
HammI	*Hammarberg Index*
HMM	Modelo Oculto de Markov —*Hidden Markov Model*—
HMM-TTS	*Hidden Markov Model based Text-to-Speech*
HNM	Modelo Armónico más Ruido —*Harmonic plus Noise Model*—
HNR	*Harmonic-to-Noise Ratio*
HTK	*Hidden Markov Model Toolkit*
IC	Estado Incompleto de Cierre —*Incompleteness of Closure*—
LAICOM	Laboratorio de Análisis Instrumental de la Comunicación
LDA	Análisis Discriminante Lineal —*Linear Discriminant Analysis*—
LF	Liljencrants-Fant
LPC	*Linear Predictive Coding*
LT	Tensión Longitudinal —*Longitudinal Tension*—
LTAS	Espectro Medio a Largo Plazo —*Long-Term Average Spectrum*—
MANOVA	Análisis Multivariante de la Varianza —*Multivariate ANalysis Of VAriance*—
MBROLA	*Multi-Band Resynthesis Overlap-Add*
MC	Compresión Media —*Medial Compression*—
MDVP	*Multi-Dimensional Voice Program*
MFCC	*Mel Frequency Cepstral Coefficients*
ML	Aprendizaje Automático —*Machine Learning*—
MOS	Nota Media de Opinión —*Mean Opinion Square*—

ACRÓNIMOS

MRI	Imagen por Resonancia Magnética —*Magnetic Resonance Imaging*—
mRMR	Mínima-Redundancia Máxima-Relevancia —*Minimal-Redundancy Maximal-Relevance*—
NAQ	Cociente de Amplitud Normalizado —*Normalized Amplitude Quotient*—
NNE	*Normalized Noise Energy*
NSW	Palabra no Estándar —*Non-Standard Word*—
OQ	Cociente de Apertura —*Open Quotient*—
OQG	Gradiente del Cociente de Apertura —*Open Quotient Gradient*—
PCA	Análisis de Componentes Principales —*Principal Component Analysis*—
PDA	Algoritmo de Detección de *Pitch* —*Pitch Detection Algorithm*—
PDS	Procesamiento Digital de la Señal
pe1000	*Relative Amount of Energy above 1000 Hz*
PLN	Procesamiento del Lenguaje Natural
PMA	Algoritmo de Marcado de *Pitch* —*Pitch Marking Algorithm*—
PMFA	Algoritmo de Filtrado de Marcas de *Pitch* —*Pitch Marks Filtering Algorithm*—
PSOLA	*Pitch Synchronous OverLap and Add*
RA	Tiempo de Retorno —*Return Time*—
RAPT	*Robust Algorithm for Pitch Tracking*
RCG	Gradiente de Velocidad de Cierre —*Rate of Closure Gradient*—
RG	Frecuencia Glótica —*Glottal Frequency*—
RK	Asimetría Glótica —*Glottal Skew*—
RQ	Cociente de Retorno —*Return Quotient*—
SA	*Simulated Annealing*
SAMPA	*Speech Assessment Methods Phonetic Alphabet*
SFM	*Spectral Flatness Measure*
SHE	Síntesis del Habla Expresiva
SI	Índice de Velocidad —*Speed Index*—

SKG Gradiente de Asimetría —*Skewness Gradient*—

sPMA Marcador Simple de *Pitch* —*simple Pitch Marking Algorithm*—

SQ Cociente de Velocidad —*Speed Quotient*—

SVM Máquina de Soporte Vectorial —*Support Vector Machine*—

TD-PSOLA *Time-Domain Pitch Synchronous OverLap and Add*

TTS *Text-to-Speech*

UAB *Universitat Autònoma de Barcelona*

URL *Universitat Ramon Llull*

VoQ Cualidad de la Voz —*Voice Quality*—

CAPÍTULO 1

Introducción

La presente tesis se enmarca dentro del programa de doctorado *Las Tecnologías de la Información y las Comunicaciones y su gestión*. Se ha llevado a cabo dentro del grupo de investigación *Grup de Recerca en Tecnologies Mèdia* (GTM) de *Enginyeria i Arquitectura La Salle*, perteneciente a la *Universitat Ramon Llull* (URL), bajo la dirección del Dr. Ignasi Iriondo Sanz y del Dr. Joan Claudi Socoró Carrié.

En este primer capítulo se hace una introducción a la tesis por medio de los siguientes puntos:

- Presentación del contexto de la investigación realizada (Sección 1.1).
- Definición de los objetivos a alcanzar (Sección 1.2).
- Organización de la tesis (Sección 1.3).

1.1. Contexto de la investigación

En un mundo cada vez más tecnológico e informatizado, la interacción hombre-máquina es una barrera que en ocasiones dificulta la integración de sistemas automáticos en ámbitos donde los usuarios se verían beneficiados. Esta interacción a la que nos referimos no es más que la utilización de dispositivos y cómo estos dan una respuesta, de forma que métodos tradicionales como son el teclado, el ratón y el monitor quedan alejados de permitir a la tecnología desarrollar todo su potencial. Es por eso que se necesita de una interacción más natural entre el hombre y la máquina, donde no sólo se valore la eficacia del sistema sino que también sea importante cómo está realizando la tarea para la que se diseñó. Por ejemplo, un sistema de lectura para niños ciegos leerá un cuento de forma que se entiendan todas las palabras, pero puede ser que una parte importante del mensaje se pierda por no haber transmitido la expresividad que este requería. Otro ejemplo es el de un servicio automático de

atención al ciudadano, donde no sería suficiente un módulo de reconocimiento de la voz, ya que parte de la interacción estaría contenida en el grado de ansiedad o estado de ánimo del interlocutor, siendo necesaria una respuesta acorde a las necesidades emocionales de cada persona.

Dotar de "humanidad" o de naturalidad a la tecnología está íntimamente ligado a ser capaces de percibir y generar emociones o expresividades de forma automática, ya que en ellas reside parte de la información que utilizan las personas para entenderse y comunicarse (Cowie et al., 2001). De ahí que la problemática de una interacción natural se pueda plantear desde dos posibles puntos de vista, ambos de gran importancia y relacionados entre sí. El primero es el de ser capaces de modelar el estado de ánimo del interlocutor (reconocimiento a partir de imagen y de voz), de tal forma que el sistema sea capaz de dar la mejor respuesta posible tal y como lo haría y esperaría una persona. El segundo es el de generar una respuesta lo más natural posible, usando para ello síntesis audiovisual, donde la generación de un estímulo sonoro (síntesis del habla) y un estímulo visual (personaje virtual) transmitiría la respuesta adaptándola al interlocutor. No todas las interacciones hombre-máquina necesitan del sistema de reconocimiento y de generación de estímulos expresivos en su totalidad, ni siquiera será siempre necesario generar un mensaje visual junto a uno sonoro, pero el hecho de enfocar el mismo problema desde diferentes perspectivas permitirá descubrir soluciones comunes. Un ejemplo sería el caso de conocer, durante el proceso de síntesis del habla, aquellos parámetros de la voz que mejor caracterizan a una emoción, información que solamente podrá será extraída a partir de estudios de reconocimiento de emociones.

Vista la problemática de dotar de naturalidad a un sistema automático cualquiera, y a partir de una de las principales líneas de investigación que existen dentro del grupo de investigación GTM, la de Conversión de Texto en Habla (CTH), se decidió trabajar en la Síntesis del Habla Expresiva (SHE), es decir, en ser capaces de generar habla con contenido expresivo, dotando así de mayor naturalidad al sistema de CTH. Para ello se planteó la necesidad de aplicar nuevas parametrizaciones que caracterizaran a la voz, y sobre ello analizar qué parámetros modelarían estilos de habla expresivos y cómo podrían ser posteriormente modificados durante el proceso de síntesis. Con este objetivo se inició un proceso de investigación de los parámetros de Cualidad de la Voz —*Voice Quality*— (VoQ), parámetros que llevaban décadas siendo utilizados por su capacidad en la caracterización del habla de una persona, especialmente en el análisis clínico para la detección de patologías de la voz (Hirano, 1981), siendo usados por esta misma razón en las tecnologías del habla (Keller, 2005). A partir de aquí se comprobó su utilidad en reconocimiento de estilos de habla expresivos (línea de investigación emergente en el grupo) y en la SHE, complementando así a los primeros experimentos donde únicamente se utilizaban parámetros de prosodia (Iriondo et al., 2007c).

1.2. Objetivos de la tesis

Como se ha mostrado en la Sección 1.1, esta tesis se enmarca en el contexto de las tecnologías del habla, más concretamente en la CTH, también conocida como síntesis del habla. El interés se ha centrado en "humanizar" el sistema de CTH, haciendo que el mensaje oral que se genera sea lo más natural posible, es decir que sea lo más parecido a lo que una persona generaría en una situación normal. Para poderlo llevar a cabo tendremos que ser capaces de modelar las emociones o expresividades, inherentes a la respuesta que cualquier persona da ante un estímulo externo.

Expuesto el problema, se formula la pregunta de investigación de partida presentada a continuación:

> *"¿Qué factores rigen la capacidad de un interlocutor de transmitir oralmente emociones o expresividades, haciendo que un oyente sea capaz de interpretarlas correctamente?"*

Para dar respuesta a esta pregunta, se plantea la doble hipótesis de trabajo siguiente:

1. Relativa al análisis y capacidad discriminatoria de la expresividad en el habla. El uso de parámetros de VoQ medidos directamente sobre la señal acústica de la voz permite una mejora de la capacidad discriminatoria mediante el uso de técnicas de clasificación automática y selección de atributos.

2. Relativa a la modificación y síntesis del habla expresiva. En el ámbito de la SHE, la modificación de parámetros de VoQ combinada con modificación de prosodia ha de permitir una mejor percepción de la expresividad en el habla generada respecto al uso único de modificación prosódica.

De este modo, planteadas las hipótesis de trabajo, se define el objetivo general a conseguir: **modelar el habla para su aplicación en SHE**. Para su consecución se necesitará de una serie de objetivos específicos.

1. **Seleccionar el conjunto de parámetros que caractericen el habla.** El interés se centra especialmente en caracterizar aquellos aspectos que permiten al habla ser un medio de transmisión de emociones o expresividades. Los parámetros elegidos son los de VoQ, que se unirán a los de prosodia, empleados previamente en el grupo de investigación con esta finalidad, cuyas limitaciones son conocidas. Este proceso de búsqueda ha dado lugar al conocimiento de variantes de parámetros que podrían ser de utilidad en otras aplicaciones, con lo que esta información ha quedado documentada para servir de guía a otros investigadores.

2. **Estudiar la capacidad discriminatoria de la VoQ.** De este proceso de discriminación se extraen las relaciones entre los parámetros de VoQ y los estilos de habla expresivos, descubriendo la vinculación que los diferentes parámetros tienen con el estilo expresivo bajo estudio, los valores típicos para cada uno de

ellos, los factores de transformación necesarios para pasar de uno a otro y, en definitiva, los modelos de VoQ. Estos modelos serán aplicables durante la SHE, ya que se dispondrá de las reglas de transformación necesarias para realizar la conversión entre estilos de habla expresivos en un sistema de CTH. Dada la naturaleza de este segundo objetivo, las relaciones obtenidas entre la VoQ y los estilos de habla expresivos permiten abrir líneas de trabajo en reconocimiento de emociones, ampliando así las posibilidades de su uso.

3. **Transformar estilos de habla expresivos en la SHE.** Se propone la metodología de transformación de estilos de habla expresivos para la SHE en un sistema de CTH, mediante la modificación de los parámetros de VoQ que indique el modelo. Como parte del objetivo de llevar a cabo la SHE, aquellos parámetros seleccionados, cuya definición no se ajustaba a las necesidades del habla expresiva, fueron redefinidos, con lo que se presentan las propuestas metodológicas de análisis y de modificación correspondientes para su modelado.

Se han mostrado los objetivos principales a conseguir y sobre los que centra esta tesis. No obstante, dadas las posibilidades de los parámetros de VoQ y su modelado, se han creado líneas de colaboración con otros investigadores pudiendo de este modo explotar su utilidad. Estas aplicaciones son el reconocimiento de emociones, la validación automática de corpus orales expresivos y la evaluación objetiva de sistemas de CTH.

1.3. Organización de la tesis

En esta sección se presenta la organización de la tesis, indicando para cada uno de los seis capítulos en los que está dividida los temas que se tratarán.

El capítulo 1 es en el que nos encontramos en este momento, donde se ha introducido al lector al contexto de la investigación, se han mostrado los objetivos que se desean conseguir y se explica cómo está organizada la tesis.

La tesis comienza con el capítulo 2, en el que se establecen los fundamentos del trabajo de investigación realizado. Las temáticas aquí tratadas se inician con la producción de la voz, muestran como su análisis clínico aporta herramientas y experiencia en el proceso de identificación de tipos de habla y se definen la prosodia y la VoQ, que junto con sus parámetros caracterizarán el habla.

El capítulo 3 trata del estado de la cuestión de los temas sobre los que gira esta tesis: la VoQ y la SHE. Se presentan los ámbitos donde la VoQ ha sido aplicada y su evaluación, las teorías y los trabajos desarrollados sobre las emociones y las relaciones existentes con el habla, los sistemas de CTH junto con las estrategias de síntesis existentes y el caso concreto del grupo de investigación GTM y, finalmente, la generación y la evaluación de la SHE.

En el capítulo 4 se presenta el material de voz empleado, es decir, la información relativa a los corpus oral para el análisis y la síntesis del habla expresiva. Este capítulo contiene una parte teórica donde se introduce cómo debe de ser un corpus oral para

SHE y se establecen las bases para su desarrollo. Por otro lado, se presenta el trabajo realizado para la creación del corpus oral expresivo del GTM, que ha sido el material básico para la realización de esta tesis. Por último, también se describen dos corpus que fueron empleados en experimentos destinados a dar mayor cobertura y ampliar los resultados obtenidos.

El capítulo 5 detalla el trabajo de investigación llevado a cabo durante el transcurso de esta tesis: el modelado de la VoQ. Se muestra cada uno de los objetivos a conseguir, las metodologías seguidas, la justificación de los resultados mediante su evaluación y las colaboraciones con otros investigadores que permitieron dar mayor repercusión a la investigación realizada. Así pues, se presentan las fases del modelado de la VoQ, los parámetros que finalmente fueron seleccionados junto con su metodología de análisis y de modificación, las propuestas metodológicas para la adaptación de los parámetros al objetivo de trabajar con habla expresiva, la capacidad de la VoQ para discriminar estilos de habla expresivos, la metodología de transformación de estilos de habla expresivos para SHE y, por último, la aplicación de la VoQ en otras aplicaciones en las que ha demostrado su utilidad.

Para terminar, el capítulo 6 presenta las conclusiones y las líneas de futuro. Primero se hace una visión conjunta, recopilando el trabajo desarrollado a lo largo de la tesis y mostrando en detalle las conclusiones obtenidas para los tres objetivos principales: la parametrización de la VoQ, la discriminación de estilos de habla expresivos usando VoQ y la síntesis del habla expresiva a partir de la metodología de transformación propuesta. Asimismo, para cada uno de estos objetivos se definen un conjunto de líneas de futuro para la mejora de los resultados obtenidos, indicando en cada caso cuáles serían los pasos a seguir.

CAPÍTULO 2

Fundamentos

En este capítulo se establecen los fundamentos de la presente tesis. Se define la terminología a ser utilizada, se aclaran posibles interpretaciones que puedan llevar a equívoco y se relacionan los conceptos entre sí. La información presentada se divide en los siguientes puntos:

- La producción del habla (Sección 2.1): que muestra los mecanismos por los cuales una persona genera un mensaje oral y su modelado.
- El examen clínico de la voz (Sección 2.2): permite conocer el comportamiento de la voz desde un enfoque totalmente objetivo, mostrando el efecto que provocan en la producción del habla las alteraciones de los diferentes parámetros que la caracterizan.
- La prosodia (Sección 2.3) y la Cualidad de la Voz —*Voice Quality*— (VoQ) (Sección 2.4): se definen los conceptos de prosodia y de VoQ, discutiendo los parámetros que los describen y las diferentes opciones de medida que se pueden encontrar. Debido a la naturaleza de la VoQ, se la relaciona con la producción del habla y el examen clínico de la voz, observando las características que presenta para su caracterización.

2.1. Producción del habla

El sistema de producción del habla, aunque estrictamente no forma parte del sistema sensorial humano, es de una importancia indudable. La comunicación humana surge instintivamente en el momento en el que aparece la necesidad de comunicación entre individuos, tanto por razones de supervivencia como en respuesta al deseo de transmisión de impresiones, sentimientos y emociones. Esta comunicación primitiva hacía uso de la mímica, gritos o interjecciones, dando pie a la constitución de un

lenguaje biológico, dando paso con el tiempo al nacimiento del lenguaje hablado y a las primeras manifestaciones pictóricas, rasgo diferencial de los hombres frente a los animales.

La comunicación oral, en el sentido más amplio de la palabra, es la expresión de nuestros pensamientos por medio de la palabra hablada y con fines comunicativos, presentando una serie de ventajas prácticas frente a la comunicación escrita (Macías, 2006):

- Facilidad: es el mecanismo "natural" de la comunicación humana.
- Aprendizaje: es el mecanismo más precoz de comunicación.
- Sencillez: el conocimiento de la propia lengua es universal, incluso en los casos de analfabetismo.
- Capacidad expresiva: haciendo uso de elementos como la entonación, el ser humano es capaz de transmitir la carga expresiva deseada.

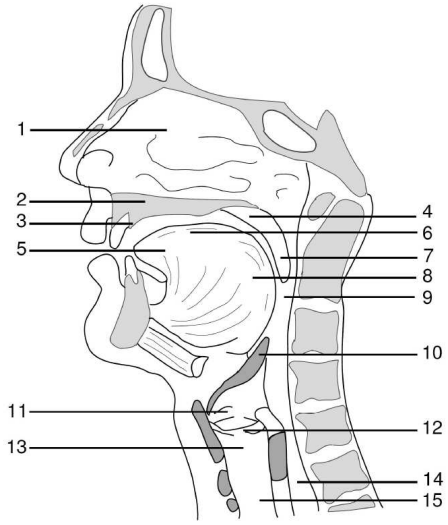
Dada la importancia que tiene la comunicación oral en la relación entre individuos, el estudio de los mecanismos de producción del habla ha sido un tema de interés por parte de investigadores, ya sea tal y como se presenta en trabajos relacionados con el análisis clínico de la voz Hirano (1981) o con las tecnologías del habla (Llisterri et al., 1999).

2.1.1. Fisiología de la producción del habla

Desde el punto de vista fisiológico, el mecanismo de producción del habla está íntimamente ligado a la función respiratoria, ya que ambas funciones usan recursos comunes. El habla, como señal acústica que es, se produce a partir de las ondas de presión que salen de la boca y de las fosas nasales de un interlocutor. El proceso comienza con la generación del flujo de aire en los pulmones, la modificación de este por parte de los pliegues vocales y su posterior perturbación por algunas constricciones y configuraciones de los órganos superiores. Así pues, en el proceso fonador intervienen distintos órganos a lo largo del llamado tracto vocal, restringiéndolos a la zona comprendida entre los pliegues vocales y las aberturas finales: los labios y las fosas nasales.

El conjunto de órganos que toman parte en el proceso de producción del habla (Figura 2.1) pueden ser divididos en tres grupos bien delimitados (Macías, 2006):

- Cavidades subglóticas: se encuentran situadas bajo la glotis, componiéndose de pulmones, bronquios y tráquea.
- Cavidad laríngea: formada por cartílagos y músculos.
- Cavidades supraglóticas: situadas sobre la glotis, se componen de las cavidades faríngea, nasal y oral.



1. Cavidad nasal
2. Paladar duro
3. Alvéolo
4. Paladar blando (velo)
5. Ápice de la lengua
6. Dorso de la lengua
7. Úvula (campanilla)
8. Raíz de la lengua
9. Faringe
10. Epiglotis
11. Falsos pliegues vocales
12. Pliegues vocales
13. Laringe
14. Esófago
15. Tráquea

Figura 2.1: *Órganos de la producción del habla (traducido de Karjalainen (2008))*

2.1.1.1. Cavidades subglóticas

Las cavidades subglóticas están formadas por los órganos propios de la respiración: pulmones, bronquios y tráquea, que son la fuente de energía durante todo el proceso de producción del habla.

Durante la inspiración la glotis está relativamente abierta, los pulmones toman aire, se baja el diafragma y la cavidad torácica se agranda. Los pulmones se expanden y se crea un paso de aire entre la boca y estos.

Durante la espiración, los pulmones se contraen empujando el aire hacia la boca. En este momento la glotis está casi cerrada y el flujo de aire es relativamente pequeño. Este flujo de aire generado en los pulmones pasa a través de la laringe y de las cavidades supraglóticas, antes de salir por la boca en forma de variaciones de presión que constituyen la señal de presión acústica.

La presión de aire en la tráquea es virtualmente la misma que la presión de aire subglótica, y es casi la misma que la que hay en los pulmones. Por el contrario, la presión de aire en las cavidades supraglóticas es prácticamente cero. La válvula glótica puede controlar el flujo de aire entre ambas cavidades subglóticas y supraglóticas, modificando de este modo la presión subglótica y así la intensidad de los sonidos producidos.

La presión subglótica es uno de los factores de mayor importancia en la producción del habla, afectando a la frecuencia fundamental (*pitch*) y a la sonía (*loudness*) (Iribar, 2008b).

2.1.1.2. Cavidad laríngea

La cavidad laríngea es la responsable de modificar el flujo de aire generado por los pulmones y convertirlo en una señal susceptible de excitar adecuadamente las posibles configuraciones de las cavidades supraglóticas. La cavidad laríngea es la responsable de la transformación del flujo de aire generado en los pulmones en sonidos audibles. A este proceso se le llama **fonación** (Sección 2.1.3).

La laringe, formada por cuatro cartílagos (tiroides, cricoides, aritenoides y epiglotis) unidos por ligamentos y membranas, conecta los pulmones con el tracto vocal a través de la tráquea. Dentro de la laringe se encuentran los pliegues vocales, el principal órgano de la cavidad laríngea, que son dos pares de repliegues compuestos de ligamentos y músculos localizados en la base de la laringe. Estos pliegues vocales pueden estar abiertos o cerrados, modificando el flujo de aire que circula hacia las cavidades supraglóticas. A la abertura que queda entre los pliegues vocales se la llama glotis. La cavidad laríngea está terminada por la epiglotis, un cartílago en forma de cuchara, que permite cerrar la apertura de la laringe en el acto de la deglución.

La distinción fundamental entre los sonidos se basa en su característica de sonoridad, debida a la acción vibratoria de los pliegues vocales. En los sonidos sonoros, incluyendo las vocales, se observa un patrón regular tanto en su estructura temporal como en su estructura frecuencial, patrón del que carecen los sonidos sordos. El mecanismo de la vibración se produce cuando estando juntos los pliegues vocales, la presión subglótica se incrementa lo suficiente para forzar su separación. Por el hecho de separarse, el aire pasa a través de ellos disminuyendo la presión subglótica, momento en el que la fuerza de los músculos hace que los pliegues vocales vuelvan a juntarse. En este momento, el flujo de aire disminuye y la presión subglótica aumenta de nuevo, con lo que se vuelve a reproducir el ciclo y la vibración provocada produce pulsos casi periódicos de aire que excitan el sistema por encima de la laringe. A esta frecuencia de vibración se la denomina frecuencia fundamental (*pitch*), y sus valores típicos, mínimo y máximo, oscilan entre los 60 Hz (hombre) y los 300 Hz (mujer o niño) (Macías, 2006). La señal generada en los pliegues vocales puede variar en frecuencia e intensidad según varíe la masa, la longitud y la tensión de los mismos.

2.1.1.3. Cavidades supraglóticas

Los pulmones proveen del flujo de aire que los pliegues vocales modulan a la hora de producir sonidos audibles. Las cavidades supraglóticas actúan filtrando este flujo de aire para producir las diferentes características que constituyen un lenguaje (Iribar, 2008a), siendo por tanto una importante componente de la producción del habla.

Las cavidades supraglóticas están constituidas por la faringe, la cavidad oral y la cavidad nasal. Su misión fundamental, de cara a la fonación, es perturbar adecuadamente el flujo de aire procedente de la laringe, para dar lugar finalmente a la señal acústica generada a la salida de la nariz y de la boca.

La faringe es un tubo musculoso que ayuda a respirar, y está situado en el cuello y revestido de membrana mucosa. Conecta la nariz y la boca con la tráquea y el

esófago respectivamente. Por la faringe pasan tanto el aire como los alimentos, por lo que forma parte tanto del aparato digestivo como del aparato respiratorio. Se divide en tres partes: laringofaringe, bucofaringe y nasofaringe, las dos últimas separadas por el velo del paladar. El volumen de la laringofaringe puede ser modificado por los movimientos de la laringe, la lengua y la epiglotis, mientras que el volumen de la bucofaringe se modifica por el movimiento de la lengua. La nasofaringe y las restantes cavidades nasales forman, desde el punto de vista de su acción sobre el flujo de aire procedente de la faringe, un resonador que puede o no conectarse al resonador oral mediante la acción del velo del paladar. Según si el resonador nasal está o no conectado, el sonido será nasal u oral respectivamente.

En cuanto a la cavidad oral, se pueden señalar las siguientes partes: los labios en el extremo, los dientes, la zona alveolar (entre los dientes y el paladar duro) y el paladar (que a su vez se puede diferenciar entre paladar duro y blando o velo). La raíz de la lengua forma la pared frontal de la laringofaringe, y sus movimientos le permiten modificar la sección de la cavidad oral (movimiento vertical), adelantar o retrasar su posición frente a la de reposo (movimiento horizontal), así como poner en contacto su ápice con alguna zona del paladar. El movimiento de los labios también interviene en la articulación, pudiendo ser de apertura o cierre y de protuberancia, alargando en este último caso la cavidad oral.

De los movimientos de los órganos en las cavidades supraglóticas surgen los distintos modos de articulación de los posibles sonidos emitidos por un interlocutor. En la mayor parte de los casos es un órgano el que se mueve (activo) y otro contra el que se efectúa la articulación (pasivo), teniendo una serie de posibles articulaciones según la pareja de órganos activo/pasivo. Por tanto, el control de los órganos que forman las cavidades supraglóticas permiten modificar los sonidos producidos por los pliegues vocales. Esta modificación se obtiene mediante mecanismos de filtrado y articulación. En primer lugar, el filtrado modifica el espectro del sonido, teniendo lugar este efecto en la faringe, cavidad nasal y vocal, que constituyen resonadores acústicos. Estos resonadores enfatizan determinadas bandas de frecuencia del espectro generado por los pliegues vocales llamadas **formantes** (Iribar, 2008a). En segundo lugar, durante el proceso de articulación algunos órganos de las cavidades supraglóticas introducen un obstáculo en el paso del flujo del aire: los labios, los dientes, el paladar y la lengua; y de acuerdo con el punto de articulación de estos órganos se obtienen los diferentes fonemas.

2.1.2. Modelo general de producción del habla

Antes de pasar a presentar el modelo de producción del habla, veamos primero aquellos elementos físicos que toman parte:

- Fuente de energía: provista por la presión de aire generada por los pulmones.
- Órgano vibratorio: los pliegues vocales, localizados en la laringe.
- Caja de resonancia: formado por las cavidades supraglóticas (faringe, cavidad oral y cavidad nasal), en los que se encuentran los órganos articulatorios.

- Elementos que en último lugar radían el sonido: las fosas nasales y los labios.

Fant (1960) introdujo el modelo de fuente y filtro para la producción del habla. Esta teoría modela el mecanismo de producción del habla como una combinación de una fuente de excitación y un filtro. La fuente y el filtro son considerados independientes el uno del otro, a pesar de que esta suposición no es totalmente cierta ya que el flujo glótico se ve influenciado por la configuración del tracto vocal. Sin embargo, la validez de esta teoría puede ser considerada suficiente para la mayoría de casos, siendo habitual esta suposición en el procesamiento del habla.

El modelado de tracto vocal se manifiesta como un filtro variable en el tiempo, cuyos parámetros varían en función de la acción consciente que se realiza al pronunciar una palabra. Este filtro variable en el tiempo tiene dos posibles señales de entrada, que dependerán de si la señal es sonora o sorda. Para señales sonoras la excitación será un tren de impulsos de frecuencia controlada, mientras que para las señales sordas la excitación será ruido aleatorio. Con la combinación de estas señales queda modelado el funcionamiento de la glotis.

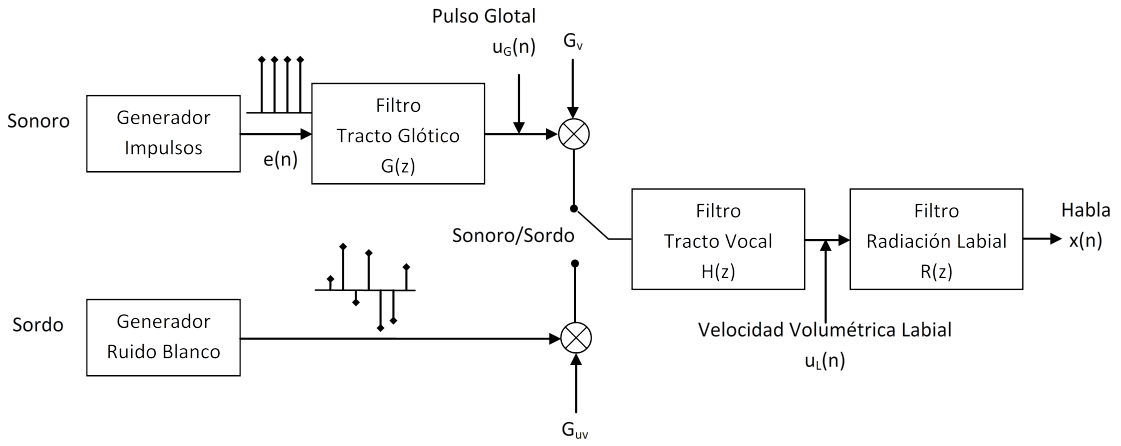


Figura 2.2: *Modelo de fuente para la producción del habla*

Los diferentes elementos que toman parte de la producción del habla son modelados como sigue. Los pulmones y las cavidades subglóticas se modelan como la fuente de excitación del sistema, la cavidad laríngea y los pliegues vocales se representan por medio del filtro de tracto glótico y, por último, las cavidades supraglóticas son modeladas por el filtro de tracto vocal. Al final de este sistema lineal se encuentran las fosas nasales y los labios, representados por el filtro de radiación labial. Este proceso se muestra en la Figura 2.2, donde se ven envueltas las señales que se describen a continuación:

- $e(n)$: Modelo de señal de excitación glótica.
- $G(n)$: Respuesta impulsional del filtro de tracto glótico.
- $u_G(n)$: Pulso glótico o señal de velocidad volumétrica glótica.

- $H(z)$: Función de transferencia del filtro de tracto vocal.
- $u_L(n)$: Señal de velocidad volumétrica labial.
- $R(z)$: Función de transferencia del filtro de radiación labial.
- $x(n)$: Señal de presión acústica.

Este modelo es utilizado para definir diferentes tipos de VoQ (Sección 2.1.3) a partir del estudio del pulso glótico (Sección 2.4.2.2). Esta señal glótica se estima por el filtrado inverso del habla producida, compensando el efecto de la radiación labial y del tracto vocal.

2.1.2.1. Modelo de tracto glótico

La entrada del filtro de tracto glótico será un tren de impulsos de frecuencia igual a la frecuencia fundamental de la voz (F_0), con una respuesta impulsional del filtro que corresponde con el pulso glótico. En el caso de sonidos sordos, este filtro no se utilizará, usando como excitación del filtro de tracto vocal una señal de ruido blanco (Figura 2.2).

El filtro de tracto glótico es el responsable de los diferentes tipos de fonación y así de las diferentes VoQ. Existen diferentes estudios que buscan modelar paramétricamente la forma del pulso glótico. Los modelos más conocidos son los modelos Rosenberg y Liljencrants-Fant (LF):

Modelo Rosenberg

Rosenberg (1971) utiliza el filtrado inverso para extraer la forma de onda glótica de la señal de habla. Basado en sus resultados experimentales desarrolló un filtro para ser usado en aplicaciones de síntesis del habla, a partir de la obtención del pulso glótico y su derivada. Para el caso del modelo Rosenberg, se pueden definir los tres intervalos para la derivada del pulso glótico definidos en el modelo LF.

Modelo LF

Originalmente propuesto por Fant et al. (1985), el modelo Liljencrants-Fant (LF) es una representación del flujo glótico y su derivada. En el modelo de fuente del mecanismo de producción del habla, el flujo glótico sirve como excitación para el filtro de tracto vocal. Ha sido un modelo utilizado en estudios analíticos.

El modelo de derivada del pulso glótico (Figura 2.3), a su vez está definido por tres intervalos de tiempo:

- $[0; t_e)$: es la fase abierta para el modelo LF. Se considera que durante este periodo la glotis está abierta. La fase abierta puede asociarse a la fase de apertura y de cierre, marcándose la división entre ellas con el instante de máximo flujo glótico t_p.

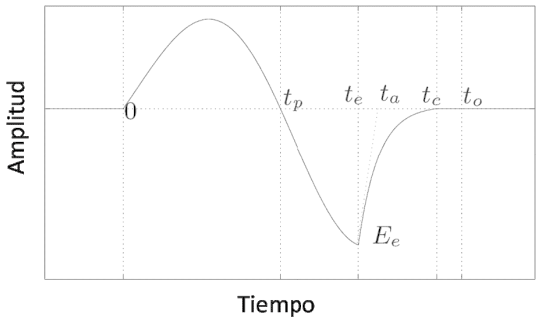


Figura 2.3: *Modelo LF: derivada del pulso glótico*

- $[t_e; t_c)$: es la fase de retorno para el modelo LF. Se trata de la duración requerida por la glotis para alcanzar el cierre completo.

- $[t_c; t_0)$: es la fase cerrada para el modelo LF. Durante este periodo se considera que la glotis está cerrada.

Tal y como se muestra en la Sección 2.4.2.2, el modelo LF se utiliza en la descripción de la VoQ a partir de la relación obtenida entre modelo y parámetros.

2.1.2.2. Modelo de tracto vocal

El tracto vocal es la cavidad formada por la laringe, faringe y la cavidad oral. La onda de presión generada por los pulmones se ve modificada primero por los pliegues vocales y posteriormente por el tracto vocal. El tracto vocal es un filtro acústico ajustable que modifica el espectro de la señal de excitación (Rubin y Vatikiotis-Bateson, 1998). Cada uno de los distintos sonidos tiene unas características espectrales producidas por una configuración del tracto vocal diferente.

El tracto vocal es básicamente un tubo curvo de diámetro cambiante (Bäckström, 2004) que puede ser modelado simplemente como un tubo cerrado en el punto donde se aplica la excitación, es decir los pliegues vocales, y abierto en los labios (Macías, 2006). La función de transferencia de un tubo sin ramificaciones, con una excitación en uno de sus extremos y midiendo su respuesta en el otro, tiene únicamente polos. Las frecuencias de resonancia, o formantes, son aquellas frecuencias donde la función de transferencia se hace infinito en condiciones ideales, apareciendo en múltiplos impares de la frecuencia fundamental de resonancia (Ecuación 2.2). La menor frecuencia a la que un tubo de estas características resuena es aquella que tiene una longitud de onda[1] cuatro veces igual a la longitud del tubo (L) (Macías, 2006). La frecuencia fundamental de resonancia (f_r) corresponde a la frecuencia más

[1] La longitud de onda λ es inversamente proporcional a la frecuencia f, siendo ésta la frecuencia del movimiento armónico simple de cada una de las partículas del medio.

baja a la que el aire que hay dentro de tal tubo vibra (Ecuación 2.1), y es igual a la velocidad del sonido en el aire[2] (c_0) dividido por la longitud de onda.

$$f_r = \frac{c_0}{4L} \tag{2.1}$$

$$f_n = f_r \cdot (2n+1), n = 0, 1, 2... \tag{2.2}$$

La posición de los formantes varía dependiendo del sonido producido y, por tanto de la posición de los órganos articulatorios que modifican la forma y el área del conducto por donde pasa el aire.

La sección del tracto vocal varía a lo largo del camino que sigue el aire desde la glotis hasta los labios, con lo que un mejor enfoque de la realidad es representar el tracto vocal como la concatenación de N tubos sin pérdidas con diferente área e igual longitud, donde cada tubo mantiene su diámetro constante. Este modelo no implica pérdidas, puesto que puede elegirse una subdivisión de tamaño suficientemente pequeño como para representar, con buena aproximación, las variaciones del tracto vocal.

Por tanto, a la hora de utilizar el modelo de tubos, se deberán de tener en cuenta las siguientes consideraciones (Bäckström, 2004):

- El tracto vocal consiste de N secciones de tubo de igual longitud y diferente diámetro.
- La longitud de cada sección es suficientemente pequeña como para que la propagación del sonido pueda ser tratada como una onda plana[3].
- Las secciones son rígidas y las pérdidas internas, debidas entre otras causas a vibraciones de las paredes, pueden ser ignoradas.
- El modelo es lineal y está desacoplado de la glotis.
- Las interacciones con el tracto nasal son ignoradas.

Las frecuencias de resonancia resultantes son la suma de las frecuencias de resonancia de cada una de las secciones. Los formantes están localizados en diferentes posiciones dependiendo del sonido producido. Sin tener en cuenta el retraso introducido ni las pérdidas, la función de transferencia resultante del tracto vocal es un filtro todo-polos (Macías, 2006), con N polos determinados por los coeficientes a_k, $k = 1, 2 \ldots N$ (Ecuación 2.3):

[2]La velocidad del sonido depende del medio por el que se propaga y la temperatura. Para el caso del aire la velocidad de propagación a 0 °C es de 331 m/s, con un incremento de $0,6$ m/s por cada °C que aumente la temperatura.

[3]Una onda plana es una onda de frecuencia constante cuyos frentes de onda (superficies con fase constante) son planos paralelos de amplitud constante normales al vector velocidad de fase. Es decir, son aquellas ondas que se propagan en una sola dirección a lo largo del espacio.

$$H(z) = \frac{1}{1 + \sum_{k=1}^{N} a_k \cdot z^{-k}} \tag{2.3}$$

2.1.2.3. Modelo de radiación labial

La onda de presión de la señal de voz $x(n)$ está relacionada con la onda de velocidad volumétrica $u_L(n)$ presente en los labios a través de una impedancia de radiación $R(z)$ (Wong et al., 1979), que se considera invariante de acuerdo al sonido producido.

Para frecuencias por debajo de los 4000 Hz, la señal de presión sonora, a una distancia l de los labios, es proporcional a la derivada temporal de la velocidad volumétrica en los labios con un tiempo de retraso de l/c_0. Excluyendo la constante de proporcionalidad y el tiempo de retraso, a bajas frecuencias se puede aproximar la impedancia de radiación $R(z)$ por un filtro pasa-altas (Ecuación 2.4).

$$R(z) = 1 - \alpha \cdot z^{-1} \tag{2.4}$$

2.1.3. Fonación

La fonación es el trabajo muscular realizado para emitir sonidos inteligibles, es decir, para que exista la comunicación oral. El objetivo último de la fonación es la articulación de palabras, a través del proceso por el cual se modifica la corriente de aire, procedente de los pulmones y de la laringe, en las cavidades supraglóticas como consecuencia de los cambios de volumen y de la forma de estas (Sección 2.1.1). El sistema fonatorio se vincula con otros sistemas, como el respiratorio, siendo su interacción parte activa en la función fonatoria, que se regula por el sistema nervioso central y periférico.

La cavidad laríngea es la responsable de la transformación del flujo de aire generado en los pulmones en sonidos audibles, es decir la **fonación**, y constituye el conjunto de los parámetros de VoQ. Las personas pueden usar diferentes tipos de fonación en distintas situaciones, como por ejemplo susurrar cuando está contando un secreto.

Como se explica en la Sección 2.1.1.2, dentro de la laringe se encuentran los pliegues vocales, principal órgano de la cavidad laríngea. Los pliegues vocales son un par de pliegues compuestos de ligamentos y músculos que están localizados en la base de la laringe, pueden estar abiertos o cerrados, modificando de este modo el flujo de aire que circula. Al espacio que queda entre los pliegues vocales se le llama glotis.

Durante la respiración normal, los pliegues vocales se mantienen separados, permitiendo el libre paso de aire sin crear ningún sonido. Si se cierran lo suficiente, obstruyendo el paso de aire, se crea una turbulencia en la glotis produciéndose así los diferentes sonidos. Estos sonidos, pueden ser fundamentalmente agrupados en

dos grupos según sean sordos o sonoros. Cuando los pliegues vocales están abiertos y, por tanto, relajados, el aire no produce vibración alguna y, de este modo, se crea un sonido sordo. Por el contrario, si estos vibran, se creará un sonido sonoro. Cuando los pliegues vocales vibran aparece el concepto de frecuencia fundamental (F_0), que se determina por la frecuencia de la vibración glótica.

Además de la distinción entre sonidos sordos y sonoros, se puede hacer una distinción en función de parámetros laríngeos, que definen las diferentes fonologías. Según Laver (1980) los principales parámetros son: Tensión Abductora —*Adductive Tension*— (AT), Compresión Media —*Medial Compression*— (MC) y Tensión Longitudinal —*Longitudinal Tension*— (LT). En la Figura 2.4 se muestra la interpretación geométrica de cada uno de ellos.

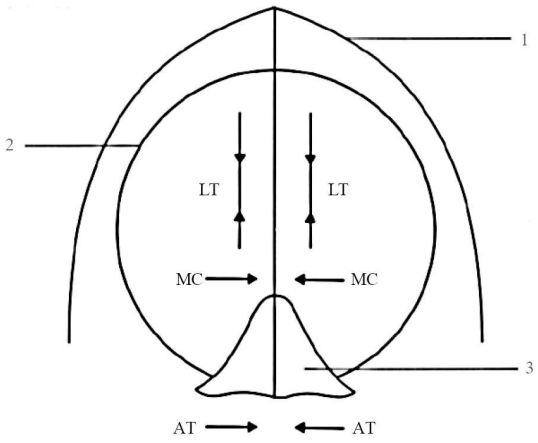


Figura 2.4: *Relación geométrica entre parámetros laríngeos (Laver, 1980). Tensión Abductora (AT), Compresión Media (MC) y Tensión Longitudinal (LT). 1. Cartílago Tiroides, 2. Cartílago Cricoides, 3. Cartílago Aritenoides*

Keller (2005) recoge en su trabajo los esquemas propuestos por Laver (1980, 1991), que se basa en las clasificaciones perceptivas de las posiciones articulatorias a partir de modificaciones de VoQ. De aquí, a partir de los tres parámetros de la configuración laríngea presentados (AT, MC y LT) se definen seis tipos de fonación o de VoQ:

- **Modal (*modal*)**: es el modo neutro de fonación. Este es el tipo ideal y normal de fonación, en el cual la vibración de los pliegues vocales es periódica.
- **Falsete (*falsetto*)**: en este tipo de voz, los pliegues vocales son fuertemente estirados, volviéndose muy finos. El resultado de las vibraciones puede llegar a doblar la frecuencia que un interlocutor produciría usando la voz modal.
- **Susurrante (*whispery*)**: caracterizada por una voz susurrante, es diferente de otros tipos de fonación por tener una constricción de la glotis resultando en un flujo de aire turbulento y un característico siseo. Se produce cuando la tensión abductora es muy baja y los pliegues vocales no vibran.

- **Chirriante (*creaky*)**: se caracteriza por una voz chirriante debido a que los pliegues vocales se comprimen, formando una gran vibración irregular. Solamente la parte delantera de los pliegues vocales vibra, produciendo una baja frecuencia de vibración. Se producen pulsos en un margen de frecuencias de 20 a 90 Hz.
- **Áspera (*harsh*)**: caracterizada por una voz áspera debido a la aparición de aperiodicidad en la frecuencia fundamental. Este tipo de voz incluye el uso de los pliegues ventriculares (los falsos pliegues vocales), que presionan la superficie superior de los pliegues vocales y vuelve su vibración ineficaz.
- **Aérea (*breathy*)**: se trata del caso donde la voz es como un murmullo, de forma que se produce en los pliegues vocales un flujo de aire que produce un ruido audible. Esto se da cuando existe un cierre glótico incompleto en el ciclo de vibración.

Por otro lado, las relaciones entre los parámetros laríngeos y cada uno de estos seis tipos de fonación quedan resumidas en la Tabla 2.1 (Keller, 2005; García, 2007).

Tipo de fonación	AT	MC	LT
Modal	Moderada	Moderada	Moderada
Falsetto	Alta	Importante	Moderadamente alta
Whispery	Baja	De moderada a alta	Moderadamente alta
Creaky	Fuerte	Fuerte	Poca
Harsh	Extrema	Extrema	Alta
Breathy	Mínima	Débil	Bastante baja

Tabla 2.1: *Relación entre parámetros laríngeos y tipos de fonación*

Para terminar, en cuanto a las fonaciones *breathy* y *whispery* comparten las características de fricción, la mayor diferencia entre ambos tipos de fonación es la cantidad de tensión en la voz, que se hace presente debido a toda la musculatura del tracto vocal (Keller, 2005). Esta configuración de tensión da lugar a que el tipo de fonación pueda por ejemplo ser relajado (*lax)*. Según Biemans (2000), debido a que el término *whispery* es más ampliamente aplicable que el término *breathy*, será este el que se tendrá en consideración.

2.2. Examen clínico de la voz

Esta sección está dedicada a la revisión del examen de la voz con el objetivo de estudiar patologías asociadas a la función fonadora. El interés radica en que a partir de las bases establecidas para el examen clínico, aparecen las definiciones de los parámetros que han sido empleados en las tecnologías del habla.

La producción de la voz implica un proceso de control complejo y preciso, por parte del sistema nervioso central, sobre una serie de eventos dados sobre los órganos fonadores periféricos. El uso efectivo de los métodos para el examen clínico de la voz,

aplicable a cada sujeto o paciente, exige de un cuidadoso entendimiento del proceso de producción de la voz y del significado y la naturaleza de cada prueba clínica.

2.2.1. Proceso periférico de la producción de la voz

En el acto comunicativo de la creación del mensaje oral toda la actividad del sistema nervioso central se ve reflejada, en última instancia, en una actividad muscular de los órganos fonadores, los cuales producen una serie de sonidos conocidos como voz. En la mayor parte de los casos, los parámetros ligados al proceso periférico[4] de la producción de la voz son evaluados o medidos durante el examen clínico de la misma. En la Tabla 2.2 se resumen los procesos periféricos implicados en la producción y percepción de la voz siguiendo el trabajo de Hirano (1981).

El punto clave de la producción de la voz es la vibración de los pliegues vocales, que transforman la energía aerodinámica en energía acústica. Desde este punto de vista, los parámetros implicados en el proceso de fonación pueden dividirse en tres grupos mayoritarios:

- Los que **regulan** el patrón vibratorio de los pliegues vocales.
- Los que **especifican el patrón vibratorio** de los pliegues vocales.
- Los que **especifican la naturaleza** del sonido generado.

Los parámetros que regulan el patrón vibratorio de los pliegues vocales pueden ser divididos, a su vez, en dos grupos: **fisiológico** y **físico**. Los factores fisiológicos están relacionados con la actividad de los músculos respiratorios, fonadores y articulatorios. Los factores físicos incluyen la fuerza espiratoria, la condición de los pliegues vocales y el estado del tracto vocal. La fuerza espiratoria es la fuente de energía de la fonación y se regula principalmente mediante los músculos respiratorios, el estado del sistema broncopulmonar y la caja torácica. La condición de los pliegues vocales, que constituyen los vibradores, se describe con respecto a la posición, forma, tamaño, elasticidad y viscosidad de los mismos; y está influenciada por la actividad de los músculos laríngeos, condiciones patológicas de los pliegues vocales y de estructuras adyacentes. Finalmente, el estado del tracto vocal, siendo este el canal que existe entre la glotis y los labios, afecta hasta cierto punto al patrón de vibración de los pliegues vocales y se ve regulado sobre todo por los músculos articulatorios. Estos factores físicos primarios determinan a su vez ciertas características secundarias, que incluyen la caída de presión en la glotis, velocidad de volumen o velocidad de flujo de aire promedia y la impedancia glótica o resistencia glótica media. Estas características secundarias están referidas a parámetros aerodinámicos.

Por otro lado, el patrón vibratorio de los pliegues vocales puede ser descrito con respecto a varios parámetros, incluyendo la frecuencia o periodo fundamental, regularidad o periodicidad en vibraciones sucesivas, simetría entre el par de mucosas

[4]Un proceso periférico, asociado al sistema nervioso periférico, es aquel que coordina, regula e integra los órganos, por medio de respuestas inconscientes.

	Reguladores del patrón vibratorio		Determinantes del patrón de vibración	Determinantes del sonido generado	
Nivel	**Fisiológico**	**Físico**	**Físico**	**Acústico**	**Psicoacústico**
Parámetros	Control neuromuscular	(Primario)	Periodo fundamental	Frecuencia fundamental	Tono
	- Músculos respiratorios	Fuerza espiratoria	Simetría Periodicidad	Amplitud (intensidad)	Sonia
	- Músculos laríngeos	Pliegue vocal - Posición - Forma y tamaño - Elasticidad - Viscosidad	Uniformidad Cierre glótico Amplitud Onda mucosa Velocidad de excursión	Forma de onda Espectro acústico Fluctuación	Cualidad Fluctuación
	- Músculos articulatorios	Estado del tracto vocal (Secundario) Caída de presión en la glotis Velocidad de volumen Impedancia glótica	Forma de onda del área glótica		

Tabla 2.2: *Parámetros utilizados en el proceso periférico de la producción y percepción de la voz (traducido de Hirano (1981))*

que forman los pliegues vocales, uniformidad u homogeneidad en el movimiento de los diferentes puntos dentro de cada pliegue vocal, cierre glótico durante la vibración, amplitud de la vibración, velocidad de excursión, la onda que viaja por la mucosa, área de contacto entre los dos pliegues vocales y forma de onda glótica entre otras.

Por último, la naturaleza del sonido generado se determina en mayor medida por el patrón vibratorio de los pliegues vocales, pudiendo ser especificado tanto en términos acústicos como psicoacústicos. Los parámetros acústicos son la frecuencia fundamental, intensidad, forma de onda o espectro acústico y sus variaciones temporales. Por otra parte, los parámetros psicoacústicos, naturalmente dependientes de los acústicos, son el tono (*pitch*), la sonía, la cualidad de la voz y sus variaciones temporales.

2.2.2. Métodos para el examen de la vibración de los pliegues vocales

El hecho de conocer la metodología de examen de la vibración de los pliegues vocales es fundamental para plantearse el análisis de la voz en cualquiera que sea la disciplina de trabajo, como por ejemplo examen de voces patológicas, estudio del género del interlocutor o de la expresividad de la voz. En esta sección se presentan los fundamentos, en base al trabajo realizado por Hirano (1981), de los métodos más habituales para el examen de la vibración de los pliegues vocales utilizados durante el análisis clínico de la voz.

2.2.2.1. Estroboscopio

Se trata de la técnica que ha resultado más práctica para el examen del patrón de vibración de los pliegues vocales. Su funcionamiento se basa en la emisión de luz mediante flashes intermitentes sincrónicos con los periodos de vibración. Cuando los flashes son emitidos a la misma frecuencia que la de vibración de los pliegues vocales se consigue observar una imagen nítida de los mismos. Esta técnica no presenta resultados finos para cada periodo de vibración, pero demuestra un patrón de vibración promediado sobre varios periodos sucesivos, hecho diferencial con la técnica de fotografía de velocidad ultra alta (Sección 2.2.2.2).

2.2.2.2. Fotografía de velocidad ultra alta

La vibración de los pliegues vocales se fotografía usando una velocidad de captura entre 20 y 30 veces la frecuencia fundamental de fonación, pudiendo observar los eventos sucedidos de forma ultra lenta. Esta técnica requiere de un dispositivo de captura más complejo y mayor capacidad de cálculo para llevar a cabo el procesado de la información capturada, por esta razón no fue un método utilizado clínicamente, sin embargo es extremadamente útil en ámbitos de investigación y enseñanza.

2.2.2.3. Glotografía foto-eléctrica

La variación del área glótica puede ser grabada mediente el uso de un dispositivo foto-eléctrico que convierte la intensidad de la luz en voltaje eléctrico. La glotis se ilumina por encima o por debajo y la intensidad de la luz que la atraviesa se mide mediante un sensor de luz colocado en el lado opuesto. Esta técnica presenta diferentes problemas:

1. La distribución de densidad de luz dentro de los pliegues vocales podría no ser constante.
2. Diferencias en el área de los pliegues vocales, en el plano anterior y posterior, podrían dar pie a su iluminación irregular.
3. Las reflexiones de luz en las superficies mucosas podrían ser variables.
4. No se tienen en cuenta movimientos verticales de los pliegues vocales hacia y desde la fuente de luz.
5. La ubicación del dispositivo de supervisión provoca diferentes formas de onda.

2.2.2.4. Electroglotografía

La electroglotografía hace uso de las variaciones de impedancia eléctrica entre dos electrodos colocados sobre el cuello. Su funcionamiento se basa en aplicar un voltaje débil y de alta frecuencia (entre $0,5$ y 10 MHz) a uno de los electrodos, mientras que en el otro se captura la corriente eléctrica que pasa a través de la laringe. La impedancia eléctrica transversal se ve modificada según la apertura y cierre de la glotis, dando como resultado una variación de la corriente eléctrica en fase con las fases de vibración de los pliegues vocales. Una de las ventajas añadidas que presenta esta técnica es que permite determinar de forma precisa el periodo de vibración de los pliegues vocales, es decir, la frecuencia fundamental de la forma de onda (F_0). Como ejemplo de inconvenientes, se tiene que la medida se ve considerablemente afectada por artefactos, incluyendo variaciones de impedancia entre los electrodos y la piel o a desplazamientos verticales de la laringe relativos a los electrodos.

2.2.2.5. Glotografía ultrasónica

Los ultrasonidos (1-10 MHz) pueden pasar a través de varios tipos de medios, incluyendo tejidos humanos. Su utilidad se basa en el hecho que cuando este tipo de onda se encuentra un cambio de medio, es decir, existe un cambio de impedancia acústica[5], esta se refleja. La relación entre la energía reflejada depende de la diferencia de impedancia acústica y del ángulo de incidencia, de forma que haciendo uso de esta característica la apertura y cierre de la glotis puede ser detectada. El principal problema que presenta esta técnica es la dificultad en la interpretación de la información capturada, ya que la compleja forma de la superficie de los pliegues vocales,

[5] Producto de la densidad del medio y la velocidad de propagación del sonido en él.

el movimiento de la zona de reflexión y los movimientos ascendentes y descendentes de la laringe, relativos a los transductores, pueden ser la mayor fuente de artefactos.

2.2.3. Parámetros para evaluar la vibración de los pliegues vocales

A continuación se introducen los parámetros más utilizados para la evaluación de la vibración de los pliegues vocales. Como se puede comprobar, algunos de los parámetros se muestran en la Sección 2.4.2.2, con la diferencia de que en esta sección se definen con relación a las características físicas de los pliegues vocales.

Excursión horizontal del borde de los pliegues vocales. El término "borde" indica la parte más centrada del pliegue vocal, la cual no es una parte fija. En cada ciclo de vibración existen variaciones de la parte del pliegue vocal, pudiéndose mover este borde tanto vertical como longitudinalmente, siendo complicada su cuantificación.

Anchura Glótica —*Glottal Width*— (GW). El término anchura glótica hace referencia a la distancia existente entre los bordes de los pliegues vocales dado un plano frontal.

Área Glótica —*Glottal Area*— (GA). El área rodeada por los bordes de los pliegues vocales se conoce como área glótica. Durante la vibración normal, la onda del área glótica se parece a la de la anchura glótica localizada en el medio de la parte membranosa. Además, la onda del área glótica es similar a la forma de onda de la velocidad de volumen, a pesar de que existen ciertas diferencias entre ambas.

Frecuencia fundamental o periodo fundamental de la vibración. El tiempo requerido por un ciclo de la vibración se conoce como periodo fundamental. Por otra parte, la frecuencia fundamental (F_0) y el periodo fundamental (T_0) están íntimamente ligados por ser uno el inverso del otro.

Fase de apertura, fase de cierre, fase abierta y fase cerrada. Un ciclo de vibración se divide en dos fases principales: fase abierta y cerrada. La fase abierta, a su vez, se divide en fases de apertura y de cierre.

Cociente de Apertura —*Open Quotient*— (OQ), Cociente de Velocidad —*Speed Quotient*— (SQ) e Índice de Velocidad —*Speed Index*— (SI). Estos parámetros se calculan a partir de relacionar los de fase abierta, de apertura y de cierre tal y como se muestra en las Ecuaciones 2.5 a 2.7.

$$OQ = \frac{\tau(\text{fase abierta})}{\tau(\text{ciclo})} \tag{2.5}$$

donde τ(fase abierta) y τ(ciclo) son las duraciones de la fase abierta y la del ciclo completo o periodo fundamental respectivamente. El valor de OQ igual a 1 se da cuando no hay un cierre glótico completo.

Por otro lado, SQ se define como:

$$SQ = \frac{\tau(\text{fase de apertura})}{\tau(\text{fase de cierre})} = \frac{\text{velocidad media de cierre}}{\text{velocidad media de apertura}} \tag{2.6}$$

donde τ(fase de apertura) y τ(fase de cierre) son las duraciones de la fase de apertura y de la fase de cierre respectivamente.

Por último, se pasa a definir SI:

$$SI = \frac{\tau(\text{fase de apertura}) - \tau(\text{fase de cierre})}{\tau(\text{fase de apertura}) + \tau(\text{fase de cierre})} = \frac{SQ - 1}{SQ + 1} \tag{2.7}$$

Amplitud. La magnitud del mayor desplazamiento es lo que se llama amplitud máxima o simplemente amplitud.

Regularidad o periodicidad de vibraciones sucesivas. Variaciones ciclo-a-ciclo del periodo fundamental, de la amplitud y/o de la onda que está siendo examinada.

Simetría bilateral de los pliegues vocales. Movimientos simétricos de los pliegues vocales indican que sus propiedades mecánicas son las mismas. Diferencias en las propiedades mecánicas de los dos pliegues vocales dan como resultado vibraciones asimétricas.

Homogeneidad. La estructura de un pliegue vocal normal es prácticamente homogénea a lo largo de su eje longitudinal. Por lo tanto, diferentes puntos a lo largo de este eje no presentan diferencias de fase sustanciales. En cuanto a la amplitud, habitualmente es mayor en el centro de la parte membranosa.

Onda Mucosa. En el pliegue vocal normal se observa como durante la vibración las ondas viajan por la mucosa desde la superficie inferior a la superior, excepto para el caso de la fonación de tipo *falsetto* (Sección 2.1.3). A esta onda se la llama onda mucosa u onda progresiva en la mucosa.

Labio superior e inferior. Para ciertas fases, durante el ciclo de vibración se observan dos formas similares a labios próximas al borde del pliegue vocal, conocidas como labios superiores e inferiores. Normalmente se observan mejor después del máximo de apertura de los pliegues vocales. Tanto los labios superiores como inferiores no son partes definidas del pliegue vocal, pudiendo variar su localización dentro de cada ciclo de vibración.

Área de contacto de los pliegues vocales. Durante la fase cerrada, el área de contacto bilateral de los pliegues vocales cambia con el tiempo. Esta información acerca del área de contacto puede ser recuperada mediante el uso de electroglotografía (Sección 2.2.2.4) y glotografía ultrasónica (Sección 2.2.2.5).

2.2.4. Análisis acústico de la señal de voz

El análisis acústico de la señal de voz es uno de los métodos más atractivos para la evaluación de la función fonadora o de patología laríngea, gracias a no ser invasiva, es decir no se necesita acceder físicamente al paciente, y por proveer de datos objetivos y cuantitativos. De muchos de los parámetros acústicos, derivados de diferentes métodos, se ha reportado su utilidad en la diferenciación entre voces patológicas y normales.

En esta sección se lleva a cabo una revisión de estudios realizados para resumir el concepto de análisis acústico, presentando los tres aspectos siguientes:

- Fuente de señal para el análisis acústico.
- Muestra para el análisis acústico.
- Parámetros acústicos para la evaluación de la función vocal.

2.2.4.1. Fuente de señal para el análisis acústico

Señal de entrada-salida. La señal más sencilla de obtener para el análisis acústico es la onda de presión sonora que sale directamente por la boca, siendo únicamente necesario un micrófono para captar la señal de voz. En este caso, además de la información de la señal glótica, se unen a ella los efectos del tracto vocal y de la radiación de los labios.

Vibración pre-traqueal o pre-laríngea. Con el objetivo de evitar los efectos del tracto vocal, a veces se usa un micrófono pre-traqueal. Es una técnica simple en la que se desconoce cómo se ven reducidos los efectos de las estructuras supraglóticas, y la información derivada de las señales de contacto pre-traqueal se ve limitada por los efectos pasa-bajos de las estructuras que intervienen.

Señal glótica obtenida por filtrado inverso basado en un modelo físico. La idea del filtrado inverso es la de obtener la señal glótica a partir de eliminar la contribución estimada del tracto vocal y de la radiación de los labios de la señal de voz. La técnica está basada en modelar teóricamente la producción vocal como un sistema físico lineal (Fant, 1960; Flanagan, 1972). El problema de esta técnica radica en la determinación precisa de los formantes y de los anchos de banda para cada señal de voz.

Residuo resultante del filtrado inverso basado en predicción lineal. Esta técnica de filtrado está basada en un modelo matemático teórico llamado "modelo de predicción lineal de producción de voz" (Atal y Hanauer, 1971; Makhoul, 1975; Markel y Gray, 1976). El filtro inverso en este caso es equivalente a la combinación de las características inversas de radiación de los labios, del tracto vocal y del espectro de la señal glótica que contribuyen a la señal de voz. La señal residuo obtenida es una estimación de una fuente periódica, teóricamente un tren de pulsos. Como se trata de una señal hipotética, no está directamente relacionada con ninguna señal físicamente observable. Debido a que cuando se empezaron a desarrollar estas metodologías no existía la capacidad de cálculo, ni las herramientas, tal y como se tiene hoy en día, esta metodología resultaba más automática y sencilla que obtener la señal glótica usando la técnica de filtrado inverso a partir de modelos físicos.

Señal glótica obtenida por un tubo sin reflexiones (tubo de Sondhi). Originalmente descrito por Sondhi (1975), es una técnica que emplea un tubo largo sin reflexiones, que se considera que actúa como una terminación pseudo-infinita del tracto vocal. Cuando una persona produce una vocal neutra en el tubo, un micrófono ubicado en su interior recoge la señal glótica debido a que la terminación sin apenas reflexiones reduce significativamente las características resonantes del tracto vocal. Cada sujeto bajo estudio requiere de un tubo distinto que encaje con su tracto vocal. Bajo condiciones ideales debería ser una forma sencilla y rápida de obtener la señal glótica, aunque a pesar de todo, grabaciones realizadas no dieron resultados satisfactorios.

2.2.4.2. Muestra para el análisis acústico

Parte estable de una vocal sostenida. Debido a la capacidad de estimar los efectos del tracto vocal y de la radiación de los labios, el análisis acústico sobre la parte estable de una vocal sostenida ha sido ampliamente utilizado, convirtiéndose en un estándar.

Partes de transición de la fonación. Muchas condiciones patológicas son más aparentes durante las fases de transición de la fonación, incluyendo la aparición y la terminación de la fonación y por lo tanto del habla. No hay un procedimiento estándar para seleccionar tales muestras, y su adecuada selección depende en gran medida de la experiencia del médico.

2.2.4.3. Parámetros acústicos para la evaluación de la función vocal

Descripción general. Los parámetros acústicos de mayor relevancia para la evaluación de la función vocal son aquellos que demuestran uno o más de los siguientes aspectos de la señal de voz:

1. **Frecuencia o periodo fundamental**
   - *a*) Media de la fonación dada
   - *b*) Posible margen de un sujeto dado (margen de la frecuencia de fonación)
   - *c*) Fluctuación de una fonación dada
     1) Magnitud
     2) Periodicidad
2. **Intensidad, presión sonora o amplitud de la forma de onda acústica**
   - *a*) Media de la fonación dada
   - *b*) Posible margen de un sujeto dado (margen de la intensidad de fonación)
   - *c*) Fluctuación de una fonación dada
     1) Magnitud
     2) Periodicidad
3. **Cantidad o riqueza de armónicos espectrales**
   - *a*) Media de la fonación dada
   - *b*) Fluctuación de una fonación dada
     1) Magnitud
     2) Periodicidad
4. **Cantidad de ruido (ruido por la turbulencia del aire)**
   - *a*) Media de la fonación dada
   - *b*) Fluctuación de una fonación dada
     1) Magnitud
     2) Periodicidad

Parámetros relacionados con la frecuencia/periodo fundamental. A partir de los valores sucesivos de la frecuencia/periodo fundamental (Figura 2.5) se pueden extraer diferentes medidas estadísticas.

- **Estadísticas estándar**
  - Media (μ_T)

$$\mu_T = \frac{1}{N} \cdot \sum_{i=1}^{N} T_i \tag{2.8}$$

  - Desviación estándar (σ_T)

$$\sigma_T = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} (T_i - \mu_T)^2} \tag{2.9}$$

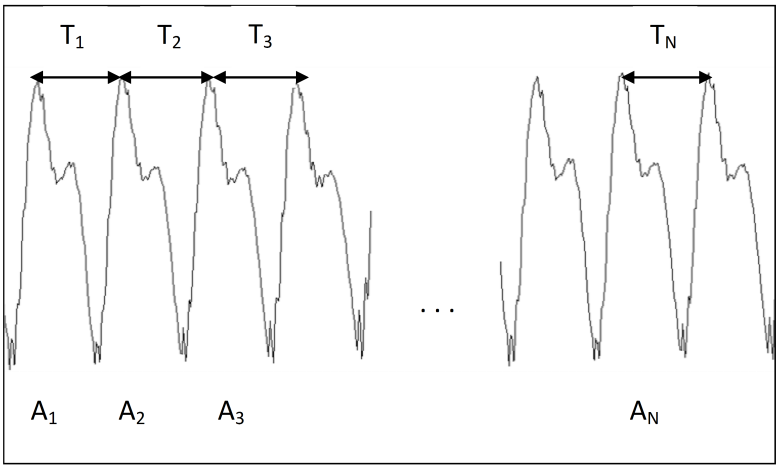


Figura 2.5: *Series de periodos fundamentales y amplitudes*

- Coeficiente de variación (v)

$$v = \frac{\sigma_T}{\mu_T} \tag{2.10}$$

- **Propuestas específicas de parámetros**

  - Perturbación de *pitch* (Lieberman, 1963)
    La perturbación de *pitch* (ΔT) se define como el valor obtenido de la diferencia de dos periodos de *pitch* consecutivos.

$$\Delta T_i = T_{i-1} - T_i \tag{2.11}$$

  - Factor de perturbación de *pitch* (Lieberman, 1963)
    El factor de perturbación de *pitch* (PPF) se define como la frecuencia relativa de perturbación de *pitch* de $0,5$ mseg o superior que ocurre en una muestra fonatoria dada.

$$PPF = \frac{\text{frecuencia de } \Delta T \text{ de } 0,5 \text{ mseg o superior}}{\text{número total de valores de } \Delta T \text{ en una muestra dada}} \tag{2.12}$$

  - Perturbación media relativa (Koike, 1973)
    A partir de la de la observación de la fonación, Koike (1973) se percató de que normalmente esta presenta cambios lentos y relativamente suaves en su periodo fundamental. Con la intención de medir perturbaciones rápidas a partir de la línea de tendencia suavizada normalizó el grado de perturbación dividiéndolo por la media del periodo fundamental, definiendo de este modo la perturbación media relativa (RAP) como:

$$RAP = \frac{\frac{1}{N-2} \cdot \sum_{i=2}^{N-1} \left| \frac{T_{i-1} + T_i + T_{i+1}}{3} - T_i \right|}{\mu_T} \tag{2.13}$$

- **Correlograma**
  Representación gráfica de los coeficientes de correlación para las series temporales de un valor dado. Lieberman (1963) fue el primero en reportar que el factor de perturbación de *pitch* era mayor para voces patológicas que normales. Desde entonces, parámetros relacionados con las variaciones ciclo a ciclo del periodo fundamental fueron investigados por médicos y científicos relacionados con la voz.

Parámetros relacionados con la intensidad vocal. Para el caso de la intensidad vocal se pueden extraer las mismas estadísticas que para la frecuencia fundamental. Hay que tener en cuenta que la máxima amplitud de una forma de onda no siempre es proporcional a la presión sonora, ya que otros factores, como el OQ entre otros, pueden afectar. Fue Koike (1969) quien por primera vez prestó atención a las variaciones de amplitud ciclo a ciclo en voces patológicas y mostró las diferencias con voces normales. A partir de aquí, otros autores adoptaron parámetros acústicos relacionados con las variaciones de amplitud en sus investigaciones sobre el análisis de voces patológicas.

Armónicos espectrales y ruido. A continuación se presentan los enfoques realizados por diferentes autores para la medida de los armónicos espectrales y ruido, junto con la descripción de cada una de las propuestas.

- **Espectrografía sonora**
  Utilizada para el análisis de voces roncas, se trabajó con una clasificación de cuatro tipos (Yanagihara, 1967a,b). Además, se relacionó estos tipos con el patrón de vibración de los pliegues vocales y la velocidad de flujo de aire promedia.

  Tipo I: Las componentes armónicas se mezclan con la componente de ruido principalmente en la región de los formantes vocálicos.

  Tipo II: Las componentes de ruido en los segundos formantes de /e/ e /i/ predominan sobre las componentes armónicas, y aparecen leves componentes de ruido en la región de alta frecuencia por encima de 3000 Hz en las vocales /e/ e /i/.

  Tipo III: Los segundos formantes de /e/ e /i/ quedan totalmente reemplazados por componentes de ruido, y las componentes de ruido adicionales por encima de 3000 Hz además intensifican su energía y expanden su margen.

  Tipo IV: Los segundos formantes de /a/, /e/ e /i/ se ven reemplazados por componentes de ruido, e incluso los primeros formantes de todas las vocales a menudo pierden sus componentes periódicas por añadirse componentes de ruido. Además, se hacen presentes con mayor intensidad las componentes de ruido de alta frecuencia.

- **Espectro Medio a Largo Plazo —*Long-Term Average Spectrum*— (LTAS) (Frokjaer-Jensen y Prytz, 1976)**
  A partir del uso de un analizador de banda estrecha de 400 canales en tiempo real, se mostraban las distribuciones espectrales de la amplitud del habla promediada a lo largo de 45 segundos de medida.

- **Parámetro α (Frokjaer-Jensen y Prytz, 1976)**
  Este parámetro se define como la relación de amplitudes por encima y por debajo de 1000 Hz.

$$\alpha = \frac{\text{nivel de amplitud por encima de } 1000 \text{ Hz}}{\text{nivel de amplitud por debajo de } 1000 \text{ Hz}} \tag{2.14}$$

2.2.4.4. Análisis multidimensional

Ya que un único parámetro no es suficiente para demostrar el espectro completo de la función vocal o de las patologías laríngeas, se necesita por tanto de un análisis multidimensional utilizando múltiples parámetros acústicos.

Aplicación de señales residuo. Aplicación de señales residuo (Davis, 1976) para el cálculo de los siguientes seis parámetros para la diferenciación de voces patológicas respecto de las normales:

1. Cociente de perturbación de *pitch* (PPQ).

2. Cociente de perturbación de amplitud (APQ), análogo a PPQ.

3. Amplitud de *pitch* (PA), definido como el pico del valor de amplitud en el periodo de *pitch* de la secuencia de autocorrelación de la señal residuo para un sonido vocálico.

4. Coeficiente de exceso (EX), definido como la relación del cuarto momento de la señal con el cuadrado del segundo.

5. Planicie espectral del espectro del filtro inverso (SFF). Equivalente a la pendiente espectral de la señal glótica representada en dB/oct, y de valor negativo. Cuanto mayor es el nivel de ruido, la SFF disminuye.

6. Planicie espectral del espectro de la señal residuo (SFR). Representa la pendiente espectral de la señal residuo medida en dB/oct. Davis (1976) sugirió que SFR podría ser considerada como una medida del enmascaramiento de las amplitudes de los armónicos por el ruido. A un incremento del nivel de ruido se corresponde un aumento de SFR.

Aplicación de la señal glótica. A partir de la señal glótica, obtenida utilizando la técnica del filtrado inverso, se derivaron los siguientes 14 parámetros (Hirano, 1975; Hirano et al., 1976; Hiki et al., 1976; Hirano et al., 1977a,b; Kakita et al., 1977a,b):

1. Amplitud de la segunda cresta de la función de autocorrelación (Po).
2. Anchura de la segunda cresta de la función de autocorrelación (Wl).
3. Amplitud de la primera depresión de la función de autocorrelación (MP).
4. Desviación estándar del periodo fundamental (PDEV).
5. Desviación estándar del pico de amplitud (ADEV).
6. Periodo de fluctuación del correlograma del periodo fundamental (PPER).
7. Periodo de fluctuación del correlograma del pico de amplitud (APER).
8. Coeficiente de correlación entre periodos fundamentales adyacentes (Pl).
9. Coeficiente de correlación entre periodos fundamentales separados por 10 intervalos (P10).
10. Velocidad de variación del correlograma del periodo fundamental (PINC).
11. Coeficiente de correlación entre picos de amplitud adyacentes (Al).
12. Coeficiente de correlación entre picos de amplitud separados por 10 intervalos (A10).
13. Velocidad de variación del correlograma del pico de amplitud (AINC).
14. Velocidad de variación de la envolvente espectral (SPINC).

Estos parámetros acústicos se relacionaron con otros factores ligados con la función fonadora, incluyendo el patrón de vibración y las propiedades físicas de los pliegues vocales, medidas aerodinámicas y parámetros psico-acústicos de la voz.

Aplicación del sonograma. La posibilidad de aplicar el sonograma para el análisis de voces patológicas fue investigado por Imaizumi et al. (1980). El sonograma era un método más fácilmente accesible que los ordenadores en muchas clínicas debido a la limitación de capacidad de computación existente en la época. De este modo, se midieron los siguientes 9 parámetros acústicos sobre el sonograma de vocales sostenidas:

1. Magnitud de las fluctuaciones de frecuencia fundamental, siendo representado como la relación del valor pico a pico (ΔF_0) y la media de la frecuencia fundamental ($\overline{F}_0$).

2. Velocidad de la fluctuación de la frecuencia fundamental, medida como el número de picos positivos dentro de un segundo.

3. Magnitud de la fluctuación de la amplitud global, medida sobre la amplitud mostrada en la pantalla, determinando los valores pico a pico.

4. Velocidad de la fluctuación de la amplitud global, definida como el número de picos positivos dentro de un segundo mostrados en la pantalla de amplitudes.

5. Sonoridad de los armónicos de alta frecuencia, adoptando una relación entre el nivel medio del margen de frecuencias entre $3,5$ y $4,5$ KHz con el inferior a 1 KHz.

6. Nivel de ruido relativo, obtenido a partir de la medida de dos envolventes en una sección de la pantalla. La primera a partir de la unión de los picos de los armónicos y la segunda a partir de conectar las depresiones. Se medía el nivel de ruido relativo en dos márgenes de frecuencias, el del primer y el del segundo formante.

7. Tiempos de subida y de bajada, siendo parámetros relacionados con la amplitud. Se definió el tiempo de subida como el tiempo necesario para pasar, en la amplitud global, desde el valor del $10\,\%$ al $90\,\%$ del valor estable. Por otro lado, el tiempo de bajada se definió como el tiempo necesario para pasar del $90\,\%$ al $10\,\%$ del valor estable.

2.3. Prosodia

2.3.1. Definición de prosodia

La prosodia, tal y como define Gil (2007), está determinada por aspectos suprasegmentales y revestida de un gran valor comunicativo. Se trata de una dimensión expresiva que contribuye determinadamente al proceso de comunicación, y a través de la cual se transmiten contenidos significativos que ni el léxico ni la sintaxis por sí solos podrían proporcionar. Dichos aspectos suprasegmentales se definen como variables fonéticas o fonológicas que sólo pueden definirse en relación a dominios superiores al segmento (fonemas y sonidos), como la sílaba o la palabra.

Los rasgos prosódicos están establecidos como unidades independientes de la lengua, y están basados en determinadas cualidades perceptivas de los sonidos: su tono, su sonía y su duración. Estas propiedades perceptivas se corresponden principalmente con dimensiones físicas de la onda sonora: la frecuencia fundamental (F_0), la intensidad y el tiempo. El tiempo corresponde a la duración de la onda sonora, la intensidad está relacionada con su amplitud y la F_0 se trata del inverso del periodo de vibración de los pliegues vocales.

La prosodia tiene principalmente una función lingüística, como por ejemplo distinguir entre una afirmación o una pregunta. Además, el habla presenta variaciones

en sus rasgos prosódicos y en el timbre[6] que no son relevantes desde el punto de vista estrictamente lingüístico. En este caso, distinguimos entre la función paralingüística, que complementa el mensaje con una intención determinada o que refleja una actitud o estado emocional del interlocutor, y la función extralingüística que aporta información sobre las características del interlocutor, como son entre otras su edad, sexo o estatus socioeconómico (Iriondo, 2008).

2.3.2. Parámetros de prosodia

El timbre, cualidad acústica propia de cada sonido, los diferencia entre sí. Además, los sonidos pueden diferir entre sí, desde el punto de vista acústico/perceptivo, por su tono, su sonía y su duración. A continuación se muestra la explicación que Gil (2007) da sobre estos atributos.

Tono. El tono o la tonía, *pitch* en inglés, se podría definir como la impresión perceptiva que nos produce la frecuencia fundamental (F_0) de la onda sonora. Es, por tanto, una cualidad subjetiva dependiente de una propiedad física. Es un hecho comprobado que la capacidad discriminatoria del oído humano va variando según vamos descendiendo por la escala de frecuencias en la que se mueve. Debido a que la relación entre los cambios frecuenciales y los tonales es no lineal, la unidad que se emplea para medir la tonía es el *mel*, a partir de la aplicación de una escala no lineal de naturaleza subjetiva (Stevens et al., 1937).

Sonía. La sonía, *loudness* en inglés, es el atributo fundamental del sonido. A menudo suscita a equívoco el concepto sonía e intensidad, en cuanto que este último se suele usar tanto para aludir a una característica física de las ondas, vinculada a su amplitud, como para hacer referencia a la impresión subjetiva que de ella se desprende; esto es estrictamente la sonía. La intensidad, en acústica, es una propiedad física inherente a la onda sonora, mensurable y definible como la potencia acústica. En cambio, la sonía es la impresión de fuerza o energía que apreciamos en los sonidos o en las secuencias de sonidos, es decir, es en última instancia el correlato perceptivo del aumento de energía en el flujo de aire procedente de los pulmones. A pesar de que el principal factor del que depende la sonía es la intensidad de la onda sonora, en su determinación influyen también otros elementos como la F_0, las características espectrales y la duración del sonido del que se trate en cada caso.

Duración. La longitud de un sonido y su correlato perceptivo, la duración, es la cantidad de tiempo empleado para su producción (p. ej. milésimas de segundo). Como pasa con el tono y la sonía, los atributos fundamentales del sonido y las sensaciones de ellos derivadas mantienen entre sí una compleja red de interrelaciones. En este sentido, es interesante señalar que mientras el tono no parece influir en la discriminación de la duración, cualquier cambio que se imprima a la sonía afecta a la

[6] Timbre es la cualidad acústica propia de cada sonido y dependiente de la configuración general de su espectro (Gil, 2007).

percepción del tiempo, es decir, para una mayor intensidad mayor será la capacidad de discriminación.

2.4. Cualidad de la voz

2.4.1. Definición de cualidad de la voz

La VoQ es un término confuso debido a la multitud de conceptos sobre el que se aplica. En cuanto a su definición, tal y como expone Biemans (2000) en su trabajo, han sido los fonetistas Abercrombie (1967) y Laver (1980) los que han desarrollado las teorías más influyentes, definiendo la VoQ como se presenta a continuación:

Abercrombie (1967)

> "El término cualidad de la voz se refiere a aquellas características que aparecen prácticamente todo el tiempo que una persona está hablando: es una cualidad cuasipermanente presente en todo sonido que sale de la boca."

Laver (1980)

> "La cualidad de la voz se concibe en un sentido amplio como la característica auditiva de la voz individual de un interlocutor, y no en el sentido más reducido de la cualidad derivada únicamente de la actividad laríngea. Tanto las características laríngeas y supralaríngeas serán vistas como contribuciones a la cualidad de la voz."

La VoQ es el centro de varios temas relacionados con el tratamiento del habla. Como ejemplos de su utilidad, se tiene por un lado la detección de patologías de voz, donde su uso es importante ya que permite la caracterización y detección de posibles problemas, ámbito donde se empezó a aplicar este concepto. Por otro lado, su estudio ha ido ganando una importancia considerable en las tecnologías del habla, como es el caso del reconocimiento de voz, donde diferencias en la voz, particularmente divergencias de la normal, permiten conocer las degradaciones de rendimiento. Asimismo, en síntesis del habla sería deseable modelar la VoQ, puesto que serían multitud de tipos de voz los que teóricamente se podrían llegar a generar. Gracias a las posibilidades que da la VoQ en la caracterización individual de la voz, se pueden encontrar otras aplicaciones como por ejemplo es el ámbito de la seguridad, con sistemas de control de acceso por voz.

La VoQ y la prosodia se complementan en el control de los aspectos que caracterizan a un estilo de habla particular (Keller, 2005). No obstante, existen autores que consideran que la VoQ no es más que la cuarta dimensión de la prosodia, junto a la duración, la sonía y el tono (Campbell y Mokhtari, 2003).

Según Biemans (2000), la descripción de la VoQ se lleva a cabo en base a dos aspectos: marco temporal (Sección 2.4.1.1) y características del locutor (Sección 2.4.1.2).

2.4.1.1. Marco temporal

Las características de la voz pueden ser a corto, medio o largo plazo. Estos tres dominios temporales tienen asociadas respectivamente las funciones: lingüística, paralingüística y extralingüística.

- **Corto plazo.** Las características a corto plazo de la información lingüística transmiten su significado a través de unidades fonológicas y gramaticales en grandes estructuras, es decir consonantes, vocales, palabras y así sucesivamente con construcciones mayores. Tiene una función comunicativa e informativa. Comunicativa porque se usa conscientemente por el interlocutor para hacer consciente al oyente de algo, mientras que es informativa debido a que es utilizada por el oyente para inferir información acerca del interlocutor en relación con su intención.

- **Medio plazo.** Las características a medio plazo tienen una función paralingüística. Transmiten el estado emocional del interlocutor, como por ejemplo enfado, pudiendo ser expresada por una voz *harsh* (Sección 2.1.3), enérgica y aguda, estando ligadas al tono de voz. Tal y como ocurre con las características lingüísticas, las paralingüísticas tienen una función comunicativa e informativa, pero al contrario de las primeras no tienen una estructura secuencial, es decir, la elección del tono de voz en un instante específico de la conversación no tiene por qué estar directamente relacionado con el tono de voz en cualquier otro.

- **Largo plazo.** Las características a largo plazo son indicadoras del comportamiento extralingüístico del habla, estando formadas por elementos que están presentes en la voz del interlocutor de una forma más o menos permanente, por ejemplo el *pitch* medio que un interlocutor tiende a usar. Biemans (2000) puntualiza que todas las características de voz a largo plazo combinadas, caracterizando a la voz de un interlocutor, se las conoce como VoQ, afirmación que liga con las definiciones presentadas por Abercrombie (1967) y Laver (1980), donde se habla de características cuasipermanentes y auditivas de la voz en un sentido amplio respectivamente. La información extralingüística es informativa pero no comunicativa, siendo esta información inferida por el oyente sin tener en cuenta las intenciones del interlocutor. En la base de la información extralingüística un oyente atribuye características de personalidad al interlocutor y valora su edad, género y rasgos regionales. Estas pueden caracterizar al individuo por si mismo, o al grupo social/regional al que pertenece. Las características de la voz a largo plazo son a nivel suprasegmental.

A pesar de definirse la VoQ en términos de largo plazo, son múltiples los trabajos que han trabajado a lo largo del tiempo con ella en el análisis y la síntesis de estados emocionales: Scherer (1989); Banse y Scherer (1996); Montero et al. (1998, 1999); Cowie et al. (2001); Campbell y Mokhtari (2003); Drioli et al. (2003); Gobl y Ní Chasaide (2003); Cabral y Oliveira (2006); Türk y Schröder (2008), de modo que su definición se amplia con respecto a las características lingüísticas (corto plazo) por

poder extraer la información de los estilos de habla expresivo transmitidos a partir del análisis de las unidades del mensaje, como por ejemplo las vocales (Drioli et al., 2003), y a las paralingüísticas (medio plazo) por verse modificados los parámetros de VoQ en función del estilo expresivo transmitido.

2.4.1.2. Características del locutor

Continuando con la visión que da Biemans (2000) en su trabajo, junto con el marco temporal, se hace distinción de dos tipos de características del interlocutor a largo plazo: características de la voz inducidas anatómicamente y configuraciones vocales. La primera de ellas no puede ser cambiada, mientras que la segunda está influenciada por los interlocutores.

- **Características anatómicas.** Las características de la voz inducidas anatómicamente aparecen a partir de diferencias anatómicas entre interlocutores. Estas diferencias pueden explicar diferencias de VoQ entre interlocutores individuales o en grupo (p. ej. la F_0 media entre hombres y mujeres es diferente).
- **Configuración extralingüística.** Puede ser definida como la manera en la que un interlocutor individual habla. Más específicamente, se constituye por el aparato vocal, manteniendo una configuración dada sobre largos tramos de segmentos, p. ej. usar voz nasal, lo que hace que las características de VoQ se compartan a lo largo de segmentos de habla. Una consideración a tener en cuenta es que el margen de actuación de una configuración va desde el corto plazo, pasando por el medio plazo, hasta el largo plazo. A corto plazo, un ejemplo lingüístico es la coarticulación. Por ejemplo, para el caso de configuración a medio plazo, se usa una voz de tipo *whispery* (Sección 2.1.3) para dar al mensaje confidencialidad. Tal y como se dijo en la Sección 2.4.1.1, no sólo aspectos extralingüísticos caracterizan a la VoQ, ya que como puede apreciarse, las características a corto y a medio plazo también se ven involucradas, aplicándose a la transmisión del estado emocional del interlocutor.

2.4.2. Parámetros de cualidad de la voz

La parametrización de la VoQ se puede realizar según los siguientes enfoques:

1. Parámetros glóticos:
   a) Señal glótica capturada a partir de transductores.
   b) Señal glótica extraída a partir de la señal acústica y del modelado del tracto vocal.
   c) Parámetros propios de la señal glótica extraídos directamente de la señal acústica.
2. Parámetros acústicos.

Dependiendo de las necesidades, o de su aplicación, será más conveniente utilizar los glóticos o los acústicos. En primer lugar, los glóticos (Fant et al., 1985), basados en la señal glótica y no acústica (casos 1a y 1b), tienen el hándicap de requerir de algún tipo de transductor como es el Electroglotógrafo (EGG), necesitando acceder al cuerpo del interlocutor de forma más o menos invasiva, o bien de modelar el tracto vocal con lo cual pueden haber errores debidos al propio método. En segundo lugar, podemos utilizar los acústicos (Drioli et al., 2003) (caso 2) de manera que su obtención se hace de forma no invasiva, evitando además errores intrínsecos del modelado. En este caso, un corpus de voz cualquiera puede servir para hacer el análisis. Finalmente, existe un enfoque intermedio que trabaja con parámetros típicos de la señal glótica a partir de parámetros acústicos (caso 1c), tal y como presenta Lugger y Yang (2006b) basándose en las observaciones de Stevens y Hanson (1994).

2.4.2.1. Parámetros glóticos: captura de la señal glótica

Respecto a la captura de la señal glótica, sean cualesquiera las técnicas utilizadas para tal fin, todas ellas son en un mayor o menor grado invasivas (incluyendo en el término invasivo cualquier metodología que implique acceder al cuerpo del interlocutor). Esto quiere decir que el interlocutor debe de entrar en contacto con algún dispositivo que haga la medición de esta señal, cosa que hace que no sea útil en aplicaciones donde esto no será posible, como por ejemplo en aplicaciones del estilo *Call Center*, donde se pretenda realizar un análisis del estado anímico del interlocutor para adoptar las medidas oportunas. Otro ejemplo es que en el caso de realizarse la grabación de un corpus emocionado, el hecho de tener dispositivos conectados al cuerpo puede afectar a la naturalidad de la voz grabada.

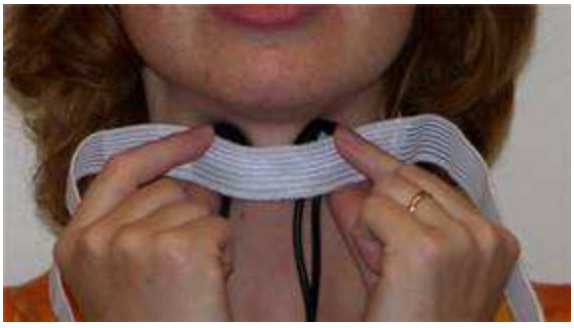


Figura 2.6: *Colocación del electroglotógrafo*

En cuanto a las técnicas existentes, se pueden encontrar diferentes aproximaciones dependiendo de la calidad esperada de la señal glótica. En Gobl y Ní Chasaide (2003) se muestra la opción presentada por Kitzing y Löfqvist (1975) y Cranen y Boves (1985). Se trata de una técnica que utiliza transductores de presión en miniatura entre los pliegue vocales, con el objetivo de obtener señales glóticas de alta calidad. Esta técnica, además de ser altamente invasiva, ya que incluso necesitaría de anestesia local, tiene el problema añadido tanto de la estabilidad del transductor así como

de las interferencias que éste podría causar en la producción vocal. Debido a estos inconvenientes no se ha obtenido gran volumen de información a partir del uso de esta técnica. Por otro lado, se tiene el uso del Electroglotógrafo (EGG), también llamado laringógrafo, método usado en la investigación del comportamiento laríngeo, utilizando un par de electrodos colocados sobre la piel a ambos lados de la laringe (Figura 2.6). Esta metodología, desarrollada por Fabre (1957) e influenciada por las contribuciones de Frokjaer-Jensen y Thorvaldsen (1968) y Fourcin y Abberton (1971), aunque considerándose rigurosamente no invasiva, todavía necesita del acceso al cuerpo del interlocutor para la captura de la señal glótica.

2.4.2.2. Parámetros glóticos: señal acústica y modelado del tracto vocal

Antes de definir los principales parámetros que describen la VoQ, usando la señal glótica extraída a partir de la señal acústica y del modelado del tracto vocal, presentemos brevemente su medida. La producción del habla puede ser modelada como la convolución de la señal fuente y la respuesta del tracto vocal, con lo que esta señal glótica se consigue a partir del filtrado inverso de la señal acústica, tratándose por tanto de una alternativa no invasiva, obteniendo de este modo un modelo del tracto vocal (Gobl y Ní Chasaide, 2003). El filtrado inverso de la señal de habla separa la fuente y el filtro por la cancelación del tracto vocal, siendo la señal resultante una estimación de la fuente.

Según Gobl y Ní Chasaide (2003), se deben de tener en cuenta diferentes consideraciones en este proceso. Se han desarrollado multitud de algoritmos para el filtrado inverso automático, la mayoría basados en alguna forma de análisis de predicción lineal, como por ejemplo: Wong et al. (1979), Ljungqvist y Fujisaki (1985), Veenevan (1985), Chan y Brookes (1989), Talkin y Rowley (1990), Lee y Childers (1991), Alku (1992), Strik et al. (1992), Alku y Vilkman (1994), Ding et al. (1994), Kasuya et al. (1999), McKenna y Isard (1999), Fröhlich et al. (2001), Alku et al. (2004) o Naylor et al. (2007). Las técnicas automáticas tienden a actuar ineficazmente cuando no hay una auténtica fase de cierre para el ciclo glótico y donde la estimación automática de los picos de los formantes no es fiable, hecho que se da para tipos de fonación no modales (Sección 2.1.3).

Otro problema es cómo de eficaz resulta la medida de parámetros a partir de la señal glótica. No existe un único conjunto de parámetros claramente definido, lo que complica las comparaciones. Además, los valores estimados de los parámetros resultantes del filtrado inverso implican típicamente algún nivel de compromiso, ya que los tiempos críticos y las amplitudes de los eventos de los pulsos glóticos no siempre quedan claramente definidos. De este modo, cómo obtener medidas óptimas a partir del filtrado inverso a menudo no es evidente. En algunas técnicas, como las presentadas por Ljungqvist y Fujisaki (1985), Milenkovic (1986), Kasuya et al. (1999) o Fröhlich et al. (2001), los parámetros de la fuente y del filtro se estiman simultáneamente, pero a menudo los parámetros se miden sobre la estimación de la señal fuente. Esto se puede hacer directamente sobre la forma de onda, por lo tanto usando solamente información en el dominio temporal, pero es más habitual la técnica

de ajuste de un modelo de fuente paramétrico para capturar las características de los pulsos glóticos obtenidos del filtrado inverso.

La técnica de la adaptación del modelo tiene la ventaja de permitir la optimización de los parámetros tanto en el dominio temporal como en el frecuencial, y también de proporcionar datos apropiados para la síntesis del habla. Sin embargo, la parametrización llevada a cabo de este modo dependerá del modelo usado. Se han propuesto numerosos modelos de fuente, como los de Rosenberg (1971), Rothenberg et al. (1975), Fant (1979a,b, 1982), Ananthapadmanabha (1984), Hedelin (1984), Fant et al. (1985), Fujisaki y Ljungqvist (1986), Price (1989), Klatt y Klatt (1990), Schoentgen (1993), Qi y Bi (1994) o Veldhuis (1998). A pesar de todas las propuestas, el más utilizado en estudios analíticos ha sido el modelo LF (Fant et al., 1985), mostrado en la Figura 2.7.

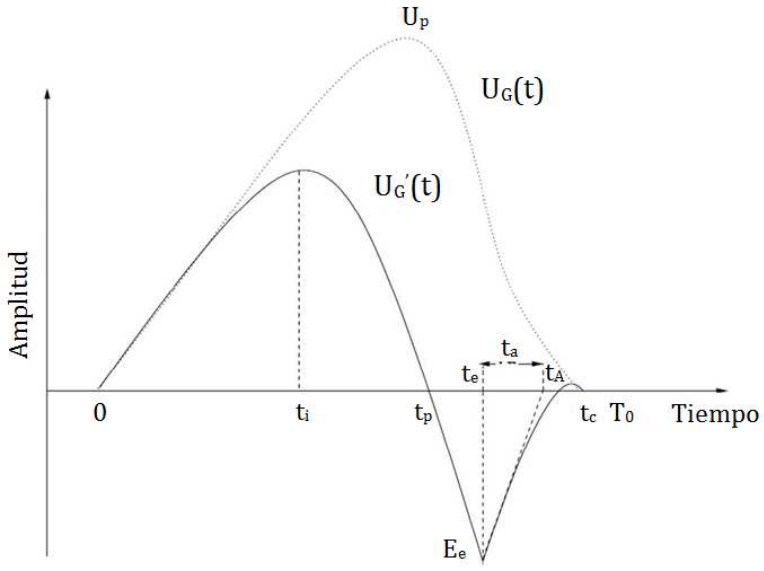


Figura 2.7: *Modelo LF: pulso glótico y su derivada*

Donde:

- $U_G(t)$: Pulso glótico
- $U'_G(t)$: Derivada del pulso glótico
- T_0: Periodo fundamental e instante de apertura glótica
- t_i: Instante del máximo de la derivada del pulso glótico
- t_p: Instante del máximo del pulso glótico
- t_e: Instante del máximo valor negativo de la derivada del pulso glótico
- t_A: Instante donde la tangente a la fase de retorno pasa por 0
- t_a: Intervalo de tiempo entre t_e y t_A

- t_c: Instante de cierre glótico
- E_e: Valor de la derivada del pulso glótico en el instante t_e
- U_p: Valor del pulso glótico en el instante t_p

Respecto a los parámetros glóticos, en la bibliografía se encuentran diferentes propuestas para su cálculo, tanto a partir de medir directamente la señal glótica como de ser extraída mediante su modelado. Jiang et al. (1998) presenta una serie de trabajos que han tratado la parametrización de la señal glótica: Hornbeck (1975), Gerratt et al. (1986, 1988), Baken (1987), Hanson et al. (1988), Zemlin (1988), Trapp et al. (1989), Murty et al. (1991a), Murty et al. (1991b), Murty et al. (1991c) y Hanson et al. (1995). Junto a estos autores existen otros como Gobl y Ní Chasaide (2003) o Keller (2005) donde las medidas se aplican sobre el modelado de dicha señal glótica. A continuación se presentan los parámetros más comúnmente utilizados, con relación a la señal glótica, junto con su medida basada en el modelo LF (Figura 2.7):

Potencia de Excitación —*Excitation Strength*— (EE). Amplitud máxima negativa de la derivada del flujo glótico (E_e), que está positivamente correlacionada con la energía armónica global.

Cociente de Apertura —*Open Quotient*— (OQ). Se define como el tiempo entre que el borde del pliegue vocal está abierto hasta que el pliegue vocal se cierra, indicando la relación de trabajo del pulso glótico. Tal y como se presentó en la Sección 2.2.3 (Ecuación 2.5), se calcula como el cociente entre la duración de fase abierta y el periodo glótico, señalando la relación entre el tiempo en el que la glotis está abierta y el tiempo de un periodo glótico completo. Variaciones en este parámetro cambian el espectro de la excitación y la relación entre la amplitud del primer y segundo armónico.

$$OQ = \frac{t_e}{T_0} \tag{2.15}$$

Cociente de Cierre —*Closed Quotient*— (CQ). Se define como el porcentaje del periodo glótico en el que los pliegues vocales impiden el paso de aire. Se calcula como el cociente entre la duración de la fase cerrada y el periodo glótico.

$$CQ = \frac{T_0 - t_c}{T_0} \tag{2.16}$$

Tiempo de Retorno —*Return Time*— (RA). El tiempo de retorno mide cómo de abrupto es el cierre glótico, es decir, el tiempo que tarda el pliegue vocal en cerrarse. Los cierres abruptos se asocian con la aparición de amplitudes en las altas frecuencias.

$$RA = \frac{t_a}{T_0} \tag{2.17}$$

Cociente de Retorno —*Return Quotient*— (RQ). Cociente entre la duración de la fase de retorno y el periodo glótico. En el dominio de la frecuencia este parámetro afecta a la pendiente espectral.

$$RQ = \frac{t_c - t_e}{T_0} \tag{2.18}$$

Asimetría Glótica —*Glottal Skew*— (RK). Mide la asimetría glótica y se calcula como el cociente entre la fase de apertura y cierre del flujo glótico. En general, los pulsos glóticos tienden a desviarse hacia la derecha. Un incremento de la asimetría da como resultado un aumento de las bajas frecuencias.

$$RK = \frac{t_e - t_p}{t_p} \tag{2.19}$$

Cociente de Velocidad —*Speed Quotient*— (SQ). Está relacionado con la asimetría del pulso glótico. Es el cociente entre la duración de la fase de apertura y de cierre, y afecta a las amplitudes de los primeros armónicos. Se trata del inverso de RK.

$$SQ = \frac{t_p}{t_e - t_p} \tag{2.20}$$

Cociente de Amplitud —*Amplitude Quotient*— (AQ). Es el cociente entre el máximo del pulso glótico y el pico mínimo de su derivada.

$$AQ = \frac{U_p}{E_e} \tag{2.21}$$

Cociente de Amplitud Normalizado —*Normalized Amplitude Quotient*— (NAQ). Es el cociente de amplitud normalizado al periodo fundamental. Es un parámetro estrechamente relacionado con CQ, pero de medida más fiable (Alku et al., 2002).

$$NAQ = \frac{AQ}{T_0} \tag{2.22}$$

Frecuencia Glótica —*Glottal Frequency*— (RG). Estima el grado de estímulo, encontrado en algunas voces, en las áreas del primer y segundo armónico.

$$RG = \frac{1}{2 \cdot t_p \cdot F_0} \tag{2.23}$$

Además de estos parámetros, existen otros derivados del modelado del pulso glótico (Gobl y Ní Chasaide, 2003):

Ruido de Aspiración —*Aspiration Noise*— (AH). Se trata de un parámetro que a menudo no se incluye explícitamente en los modelos de generación de voz. A partir de un modelo de excitación mixta, donde la señal de voz se genera a partir de una componente periódica y de una no periódica, extraer y cuantificar esta componente de ruido es un reto complejo. Sin embargo, con fines de síntesis, se puede generar un ruido pseudoaleatorio filtrado convenientemente para compensar la ausencia de una adecuada estimación empírica del ruido de aspiración en la glotis.

Diplofonía —*Diplophonia*— (DI). Es un parámetro que refleja la alteración que sufre cada segundo el pulso glótico por su desplazamiento hacia el pulso que lo precede, afectando al mismo tiempo a su amplitud. Tanto el desplazamiento como la cantidad de reducción de la amplitud se determina por el valor de este parámetro. Por lo tanto, el periodo fundamental respecto al pulso precedente se reduce, dando como resultado un incremento equivalente del periodo fundamental con respecto al siguiente pulso y proporcionando a la voz un carácter bitonal.

2.4.2.3. Parámetros propios de la señal glótica a partir de la señal acústica

Hasta el momento se ha mostrado como el modelado del pulso glótico, a partir del filtrado inverso de la señal de voz, es el método más común para la extracción de los parámetros de VoQ. Otros métodos han sido propuestos para la estimación de estos parámetros de VoQ directamente desde la señal acústica, con la ventaja de no necesitar hardware extra ni tener que acceder al cuerpo del interlocutor para obtener la información deseada (Lugger y Yang, 2006b). El método propuesto por Lugger y Yang (2006b) se basa en las observaciones presentadas en Stevens y Hanson (1994), donde las propiedades glóticas *cociente de apertura*, *apertura glótica*, *asimetría del pulso glótico* y *velocidad de cierre glótico* afectan al espectro de la excitación de la señal de voz en un margen de frecuencias, reflejando de este modo la VoQ del interlocutor.

La propuesta que se realiza es la de estimar estos estados glóticos, a partir de la señal acústica y de la relación de las amplitudes de los correspondientes armónicos más elevados, respecto de la frecuencia fundamental. Además, se observó que el ancho de banda del primer formante está correlacionado con el estado incompleto de cierre glótico. La metodología presentada por Lugger y Yang (2006a) calcula gradientes espectrales en lugar de relaciones de amplitudes puras, debido a que los gradientes caracterizan de un modo más efectivo la forma del espectro de la señal glótica. Complementariamente, realiza una compensación del tracto vocal antes de la estimación de los gradientes, basada en el trabajo de Wokurek y Pützer (2003), ya que los parámetros de VoQ dependerán solamente de la excitación y no del proceso articulatorio. A partir de un análisis LPC (*Linear Predictive Coding*) (Apéndice F.2) se determinan las frecuencias y los anchos de banda de los cuatro primeros formantes, convirtiendo todas las frecuencias a la escala Bark (Zwicker, 1961) y obteniendo de este modo las siguientes características de la voz:

- F_0: frecuencia fundamental o *pitch* medio

- F_1, F_2, F_3, F_4: frecuencias de los formantes
- B_1, B_2, B_3, B_4: ancho de banda de los formantes
- H_1, H_2: amplitud en F_0 y $2F_0$
- $F_{1p}, \ldots, F_{3p}$: frecuencia de los picos del espectro cercanos a los formantes
- $A_{1p}, \ldots, A_{3p}$: valores de amplitud en $F_{1p}, \ldots, F_{3p}$

A partir de estas características de la voz se calcula:

1. La compensación de la influencia del tracto vocal (Ecuaciones 2.24 a 2.26), usando la estimación de la contribución de los cuatro formantes al espectro en la frecuencia f de Fant (1960). Esta compensación se realiza para las amplitudes de los dos primeros armónicos y de las frecuencias cerca de los tres formantes, dando lugar a $\tilde{H}_1$ y $\tilde{H}_2$, y para las amplitudes $\tilde{A}_{1p}$, $\tilde{A}_{2p}$ y $\tilde{A}_{3p}$.

$$V(f; F_i, B_i) = 20 \cdot \log \frac{F_i^2 + \left(\frac{B_i}{2}\right)^2}{\sqrt{\left((f-F_i)^2 + \left(\frac{B_i}{2}\right)^2\right) \cdot \left((f+F_i)^2 + \left(\frac{B_i}{2}\right)^2\right)}} \tag{2.24}$$

$$\tilde{H}_k = H_k - \sum_{i=1}^{4} V(kF_0; F_i, B_i) \qquad (k = 1, 2) \tag{2.25}$$

$$\tilde{A}_{kp} = A_{kp} - \sum_{\substack{i=1 \\ i \neq k}}^{4} V(F_{kp}; F_i, B_i) \qquad (k = 1, 2, 3) \tag{2.26}$$

2. La estimación de los parámetros de VoQ (Ecuaciones 2.27 a 2.31).

Gradiente del Cociente de Apertura —*Open Quotient Gradient*— (OQG).

$$OQG = \frac{\tilde{H}_1 - \tilde{H}_2}{F_0} \tag{2.27}$$

Gradiente de Apertura Glótica —*Glottal Opening Gradient*— (GOG).

$$GOG = \frac{\tilde{H}_1 - \tilde{A}_{1p}}{F_{1p} - F_0} \tag{2.28}$$

Gradiente de Asimetría —*Skewness Gradient*— (SKG).

$$SKG = \frac{\tilde{H}_1 - \tilde{A}_{2p}}{F_{2p} - F_0} \tag{2.29}$$

Gradiente de Velocidad de Cierre —*Rate of Closure Gradient*— (RCG).

$$RCG = \frac{\tilde{H}_1 - \tilde{A}_{3p}}{F_{3p} - F_0} \tag{2.30}$$

Estado Incompleto de Cierre —*Incompleteness of Closure*— (IC).

$$IC = \frac{B_1}{F_1} \tag{2.31}$$

Para terminar, se puede ver una relación entre los parámetros calculados sobre la señal glótica, ya sea directamente medida sobre la señal del EGG o bien a partir del filtrado inverso de la señal acústica asociada, desde tres puntos de vista principalmente: **apertura**, **cierre** y **asimetría**. De este modo se tienen cubiertos los mismos aspectos en ambas metodologías de medida, obteniéndose en cada momento diferentes parámetros considerados representantes de la VoQ.

2.4.2.4. Parámetros acústicos

La última de las alternativas de medida mostrada en la bibliografía para la parametrización de la VoQ, ya sea en el ámbito clínico de análisis de la disfonía[7] (Núñez et al., 2004) o en la caracterización emocional de la voz (ámbito de interés de esta tesis) (Banse y Scherer, 1996; Alter et al., 2003; Drioli et al., 2003), se trata de su parametrización mediante parámetros extraídos directamente desde la señal acústica sin realizar ninguna transformación hacia parámetros glóticos (Sección 2.4.2.3).

Esta metodología tiene como beneficio principal no necesitar acceder al cuerpo del interlocutor, con lo que el único equipamiento requerido sería un micrófono, y la independencia de los resultados con el modelado de la señal glótica, que puede inducir a errores inherentes al propio modelo. A continuación se presentan los parámetros que se encuentran en la bibliografía:

***Jitter*.** Mide las variaciones a corto plazo del periodo fundamental, debidas a fluctuaciones en los tiempos de apertura y de cierre de los pliegues vocales, de un periodo al siguiente, describiendo un ruido que aparece en forma de modulación en frecuencia. Existen diversas medidas sobre este mismo parámetro dependiendo de si se normalizan, promedian o se consideran tal cual los valores entre varios periodos consecutivos. Tal y como se muestra en la Sección 2.2.4.3, la medida de este parámetro se basa en el trabajo de Lieberman (1963) sobre la perturbación de *pitch*. La Ecuación 2.32 muestra la medida del *jitter* (expresado en $\%$), utilizado tanto por herramientas como Praat (Boersma, 2001) o *Multi-Dimensional Voice Program* (MDVP)[8], donde N

[7] Tal y como define la Real Academia de la lengua (http://www.rae.es/rae.html), la disfonía es el trastorno cualitativo o cuantitativo de la fonación por causas orgánicas o funcionales.

[8] http://www.kayelemetrics.com/Product %20Info/CSL %20Options/5105/5105.htm

es el número de periodos de la señal de voz, i representa a cada uno de ellos y T_{0i} es su valor.

$$jitter = \frac{\frac{1}{N-1} \cdot \sum_{i=2}^{N} |T_{0i} - T_{0i-1}|}{\frac{1}{N} \cdot \sum_{i=1}^{N} T_{0i}} \cdot 100 \qquad [\,\%] \tag{2.32}$$

***Shimmer*.** Mide las variaciones a corto plazo de la amplitud de la forma de onda entre periodos, describiendo un ruido que aparece como una modulación en amplitud. Del mismo modo que para el *jitter*, presenta distintas alternativas de ser calculado manteniendo siempre la misma filosofía. Se basa en las explicaciones que se encuentran en la Sección 2.2.4.3 en relación con los parámetros relacionados con la intensidad vocal de Koike (1969). En la Ecuación 2.33 se presenta la medida del *shimmer* (expresado en $\%$), empleado por herramientas como Praat o MDVP, donde N es el número de periodos de la señal de voz y U_i es el valor de la amplitud pico a pico en cada periodo i.

$$shimmer = \frac{\frac{1}{N-1} \cdot \sum_{i=2}^{N} |U_i - U_{i-1}|}{\frac{1}{N} \cdot \sum_{i=1}^{N} U_i} \cdot 100 \qquad [\,\%] \tag{2.33}$$

***Harmonic-to-Noise Ratio* (HNR).** Describe ruido aditivo y se define como la relación entre la energía de la parte armónica y la energía del resto de la señal. Tiene como inconveniente que puede verse afectado por el *jitter* y el *shimmer*. En el trabajo de Severin et al. (2005) se presenta este parámetro y se describen y prueban tres algoritmos de estimación para su medida.

***Normalized Noise Energy* (NNE).** Parámetro que describe ruido aditivo, cuya definición difiere de la presentada para el HNR levemente para diferentes autores. NNE es la relación entre la energía de ruido y la energía total de la señal. Existen diferentes implementaciones para el cálculo de este parámetro tal y como sucedía para el HNR. Michaelis et al. (1997) presenta la opción de obtener la energía de ruido directamente entre los armónicos de su espectro, asumiendo dentro de un armónico la energía de ruido como el valor medio de las mínimas adyacentes en el espectro.

***Glottal-to-Noise Excitation Ratio* (GNE).** Del mismo modo que el HNR, la medida GNE, propuesta por Michaelis et al. (1997), se trata de un parámetro que describe ruido aditivo. Su cálculo está basado en la correlación de la envolvente de la transformada de Hilbert (Faúndez, 2001) entre diferentes canales de frecuencia, dando como resultado un valor comprendido entre 0 y 1. Se trata de una buena alternativa al uso de HNR por hacerlo prácticamente independiente del *jitter* y *shimmer*.

***Spectral Flatness Measure* (SFM).** Parámetro que calcula la relación entre la medias geométrica y aritmética de la distribución espectral de energía. Es una medida útil para dar una estimación de la tonicidad de una señal (Johnston, 1988). En la Ecuación 2.34 se presenta su cálculo (expresado en dB), donde N son el número de muestras espectrales tomadas, representadas por i, para frecuencias situadas entre $f = 0$ y la mitad de la frecuencia de muestreo ($f = f_s/2$) y E_i es la energía para cada una de estas frecuencias.

$$SFM = 10 \cdot \log \frac{\sqrt[N]{\prod_{i=1}^{N} E_i}}{\frac{1}{N} \cdot \sum_{i=1}^{N} E_i} \quad [\text{dB}] \tag{2.34}$$

***Drop-off of Spectral Energy above 1000 Hz* (do1000).** Se calcula como la aproximación por mínimos cuadrados de la pendiente espectral por encima de 1000 Hz (Banse y Scherer, 1996). En la Ecuación 2.35 se muestra el cálculo de este parámetro, utilizando la aproximación por mínimos cuadrados de la pendiente presentada por Abdi (2003). N es el número de muestras espectrales tomadas, representadas por i, para frecuencias situadas por encima de $f = 1000$ Hz hasta la mitad de la frecuencia de muestreo ($f \in (1000 \dots f_s/2]$), E_i es la energía para cada frecuencia, $\overline{E}$ es la media de la energía calculada para todas las frecuencias, $f1000_i$ son las frecuencias por encima de 1000 Hz y $\overline{f1000}$ el valor medio de estas frecuencias superiores a 1000 Hz.

$$do1000 = \frac{\sum_{i=1}^{N} (E_i - \overline{E}) \cdot (f1000_i - \overline{f1000})}{\sum_{i=1}^{N} (f1000_i - \overline{f1000})^2} \tag{2.35}$$

***Hammarberg Index* (HammI).** Drioli et al. (2003), basándose en el trabajo de Hammarberg et al. (1980), define este parámetro como la diferencia entre los máximos de energía en las bandas de frecuencia de $f \in [0 \dots 2000]$ Hz y la banda de $f \in (2000 \dots 5000]$ Hz (expresados en dB). En la Ecuación 2.36 se muestra su cálculo, donde $E_{0\dots2000}$ es la energía dentro del margen de frecuencias $f \in [0 \dots 2000]$ Hz, y $E_{2000\dots5000}$ se trata de la energía para las frecuencias $f \in (2000 \dots 5000]$ Hz.

$$HammI = 10 \cdot \log \frac{\text{máx}(E_{0\dots2000})}{\text{máx}(E_{2000\dots5000})} \quad [\text{dB}] \tag{2.36}$$

***Relative Amount of Energy above 1000 Hz* (pe1000).** Se mide la cantidad de energía relativa en el margen de las frecuencias superiores a 1000 Hz respecto a las inferiores (Scherer, 1989). La Ecuación 2.37 presenta su cálculo (expresado en dB), donde E_h representa la energía para cada una de las frecuencias por encima de 1000

Hz ($f \in (1000 \ldots f_s/2]$ Hz, donde f_s es la frecuencia de muestreo), y E_l es la energía en la banda de frecuencias de $f \in [0 \ldots 1000]$ Hz.

$$pe1000 = 10 \cdot \log \frac{\sum_{h=1}^{H} E_h}{\sum_{l=1}^{L} E_l} \quad [\text{dB}] \tag{2.37}$$

CAPÍTULO 3

Estado de la cuestión

Este capítulo presenta al lector el estado de la cuestión sobre las temáticas tratadas en esta tesis. La información se agrupa en los siguientes 4 puntos:

- El uso y la evaluación de la Cualidad de la Voz —*Voice Quality*— (VoQ) (Sección 3.1): se presentan los ámbitos de aplicación donde la VoQ ha demostrado ser de interés, así como las metodologías destinadas a su evaluación.
- La relación entre el habla y las emociones (Sección 3.2): antes de plantearse el reconocimiento o la síntesis de estilos de habla expresivos o emociones, hay que entenderlas, saber cómo pueden ser representadas y cómo tratar el habla para que su transmisión y percepción sea posible. Para ello, se relacionan los parámetros del habla (prosodia y VoQ) con la expresión de las emociones, mostrando los principales estudios existentes en la bibliografía.
- La Conversión de Texto en Habla (CTH) (Sección 3.3): donde se explica cómo se lleva a cabo la generación o síntesis del habla a partir de texto, presentando las estrategias que han ido apareciendo a lo largo de los años e introduciendo el caso concreto del sistema CTH del grupo de investigación GTM en el que se enmarca esta tesis.
- La Síntesis del Habla Expresiva (SHE) (Sección 3.4): una vez introducidos los conceptos anteriores se puede pasar a la descripción de la SHE, mostrando las principales aportaciones para su generación y su evaluación.

3.1. Uso y evaluación de la cualidad de la voz

En esta sección se presentan los ámbitos de aplicación de la VoQ y su evaluación. En primer lugar, en los ámbitos de aplicación se hace una revisión de los principales

usos que esta ha tenido y las posibilidades que presenta. En segundo lugar, se muestran las propuestas de cómo llevar a cabo su evaluación, relacionándolas a su vez con la aplicación en la que esté siendo empleada.

3.1.1. Ámbitos de aplicación de la cualidad de la voz

En esta sección se muestran los ámbitos de aplicación donde la VoQ se encuentra presente. Su utilización nació de la evaluación de las características de la voz, concretamente de la señal producida por el aparato fonador, de forma que es en el análisis clínico donde se encuentran sus inicios. En el caso del análisis clínico se trataba de evaluar el habla, cuantificar los resultados obtenidos a partir de la realización de mediciones y, en base a los parámetros que más tarde han hecho suyos otras aplicaciones, detectar patologías. Gracias a los avances técnicos producidos a finales del siglo XX se acentúa el estudio de la voz, a partir de técnicas que permiten su análisis y parametrización (Sección 2.2).

Junto a la aplicación en análisis clínico, como es por ejemplo el trabajo de Núñez et al. (2004) donde se evalúa perceptualmente la disfonía, existen diferentes tipos de uso donde se emplea la VoQ:

- Estudios sociológicos.
- Caracterización del contenido expresivo o emocional:
  - Clasificación de estilos de habla expresivos.
  - Síntesis del Habla Expresiva (SHE).
- Evaluación objetiva del habla.
- Reconocimiento del habla.

Primero, como ejemplo de estudios sociológicos se tiene el trabajo desarrollado por Biemans (2000), en el que se extrae información de la variación de VoQ en función del género, presentando relaciones entre parámetros y tipos de fonación. En segundo lugar, en cuanto a la caracterización del contenido expresivo o emocional, existen diferentes trabajos que relacionan la parametrización de la VoQ con diferentes tipos de fonación y a su vez con estilos de habla expresivos (Gobl y Ní Chasaide, 2003). Además, otros trabajos proponen utilizar la VoQ como información que permita la discriminación automática de estilos de habla expresivos y aplicarla en SHE (Drioli et al., 2003). Por otro lado, existen trabajos como el de Stylianou (1999), que utiliza la VoQ para la evaluación objetiva del habla en diseño de corpus. Finalmente, tal y como expone Keller (2005), debido a las diferencias sistemáticas existentes en distintos tipos de voz, el potencial diferenciador de la VoQ puede ser explotado por sistemas de reconocimiento del habla. Por ejemplo, tales divergencias son una parte de un patrón de identificación biológica individual, que puede ser usado como una componente de un sistema de identificación de locutor para el control de acceso.

En la mayoría de aplicaciones, la prosodia también juega un papel importante, por ello no se usa habitualmente la parametrización de VoQ de forma exclusiva. Por

ejemplo, se tiene el caso de la evaluación objetiva de la calidad del habla expresiva generada por un sistema de CTH donde se utiliza información prosódica, tanto si el sistema únicamente modela a esta (Iriondo et al., 2007c) como si la prosodia y la VoQ se ven implicadas en la generación del habla (Tesser et al., 2005).

3.1.2. Evaluación de la cualidad de la voz

Para la evaluación de la VoQ se realiza una medida desde dos enfoques: medida perceptiva y medida acústica. Por un lado, la medida perceptiva se trata de analizar subjetivamente el habla, presentando una serie de locuciones a los oyentes para que estos las juzguen. Por otro lado, la medida acústica se trata de una evaluación objetiva del habla, de forma que es a partir de su parametrización y análisis que se extrae la información relevante.

3.1.2.1. Medida perceptiva

Tal y como se describe en Gerratt y Kreiman (2003), la VoQ es la percepción auditiva de elementos acústicos de fonación que caracteriza a un interlocutor individual. Así pues, es una interacción entre la señal acústica del habla y la percepción que tiene un oyente de la misma. La medida perceptiva de la VoQ se basa en juzgar subjetivamente ejemplos del habla bajo estudio por uno o varios oyentes. Este tipo de medida es especialmente utilizado en análisis clínico aunque se aplica a otros ámbitos de las tecnologías del lenguaje, como la síntesis del habla, donde la realización de pruebas perceptivas es habitual utilizando métodos como Nota Media de Opinión —*Mean Opinion Square*— (MOS) o Nota Media de Opinión sobre las Comparaciones —*Comparison Mean Opinion Score*— (CMOS) (ITU-P.800, 1996).

Existen diferentes aproximaciones al problema de la evaluación perceptiva de la VoQ. Por una parte se tiene por ejemplo, tal y como menciona Gerratt y Kreiman (2003), el trabajo de Fairbanks (1960) donde se recomendaba que las voces fueran evaluadas usando una escala de 5 puntos para las cualidades *harshness* (áspera), *hoarseness* (ronca) y *breathiness* (aérea). Por otro lado existe el protocolo GRBAS (Hirano, 1981), propuesto por el "Comité para el Análisis de la Función Fonatoria" de la "Sociedad Japonesa de Logopedia y Foniatría" para la evaluación de la *hoarness*, o ronquera, mediante 5 escalas, utilizando para ello una graduación de cuatro puntos para cada una de ellas, "0" para la no-*hoarse* o normalidad, "1" ligera, "2" moderada y "3" extrema:

G ***Grade*** **(grado).** Representa el grado de *hoarseness* o anomalía vocal, mientras que las otras cuatro escalas representan diferentes aspectos de estas anomalías vocales.

R ***Rough*** **(aspereza).** Representa la impresión psicoacústica de pulsos glóticos irregulares. Corresponde con las fluctuaciones irregulares de la frecuencia fundamental y/o de la amplitud de la señal glótica fuente (excitación).

B ***Breath*** **(aérea).** Representa una impresión psicoacústica de la cantidad de aire que se escapa a través de la glotis.

A ***Asthenic*** **(astenia).** Denota una intensidad débil del sonido generado por la fuente glótica y/o por la falta de los armónicos más altos.

S ***Strained*** **(forzada).** Representa la impresión psicoacústica de excesivo esfuerzo, de tensión asociada con la fonación espontánea. Está relacionada con una frecuencia fundamental anormalmente elevada, ruido en el margen de altas frecuencias y/o sonoridad de los armónicos de alta frecuencia.

Debido a la subjetividad del protocolo GRBAS, el examinador debe poseer un oído entrenado. Con este fin, el "Comité para el Análisis de la Función Fonatoria" de la "Sociedad Japonesa de Logopedia y Foniatría" realizó una grabación con muestras de voz típicas representadas por este protocolo.

El protocolo GRBAS ha sido revisado y ampliado a GIRBAS (Dejonckere et al., 1998), añadiendo una nueva escala: **I** ***Instability*** **(inestabilidad)**. Junto a estos protocolos, muchos otros han sido propuestos a lo largo de los años (Gerratt y Kreiman, 2003; Aronson y Bless, 2009). El "Sistema de Perfil de Voz de Wilson" (Wilson, 1977) emplea escalas con 7 puntuaciones para: tono laríngeo, tensión laríngea, abuso vocal, sonía, tono, inflexiones vocales, cambios repentinos de tono (*pitch breaks*), diplofonía (percepción de dos tonos en la voz), resonancia, emisión nasal, margen y eficiencia vocal global. Adicionalmente pueden aparecer subperfiles para poblaciones o afecciones específicas. En Suecia se desarrolló el "Enfoque de Evaluación de Voz de Estocolmo", utilizando 13 parámetros (Hammarberg y Gauffin, 1995): tensión, relajación, *breathiness* (aérea), *creakiness* (chirriante), *roughness* (áspera), *grating* (similar a *harshness*), *falsetto* (falsete), inestabilidad de tono, cambio de voz, afonía, diplofonía, tono y sonía. Otro enfoque más elaborado fue el presentado por Gelfer (1988), donde se proponía el uso de 17 parámetros. Para terminar, en el Reino Unido apareció otro desarrollo, el "Análisis de Perfiles Vocales", protocolo que contiene más de 30 parámetros, incluyendo características del tracto vocal y prosódicas (Laver, 1980, 2000). A modo de resumen, en la Tabla 3.1 se muestran las distintas propuestas presentadas, junto a una descripción de las mismas.

La utilidad de tales protocolos está limitada debido a las dificultades existentes a la hora de establecer el correcto y adecuado conjunto de escalas necesario. Los investigadores nunca han estado de acuerdo con la estandarización de las escalas para evaluar la VoQ, habiendo evidencias que sugieren que las diferencias entre oyentes, en estrategias perceptivas, son suficientemente altas como para que los esfuerzos de estandarización no sean efectivos (Kreiman y Gerratt, 1996); y además, aparentemente los oyentes no se ponen de acuerdo en su clasificación. Los resultados señalan que de media, más de un $60\,\%$ de la discrepancia que se da en la clasificación de la VoQ, es debida a factores ajenos a las diferencias de VoQ que se están evaluando. Como ejemplo, Kreiman y Gerratt (1998) presentan los factores siguientes:

- La atención del oyente varía.

- Dificultad en el aislamiento de dimensiones perceptivas individuales dentro de estímulos acústicos complejos.
- Diferencias en las experiencias previas de los oyentes.

Autor	Descripción
Fairbanks (1960)	Evaluación de voces en escalas de 5 puntos para *harshness*, *hoarseness* y *breathiness*
Wilson (1977)	Protocolo "Sistema de Perfil de Voz de Wilson", emplea escalas con 7 puntuaciones para: tono laríngeo, tensión laríngea, abuso vocal, sonía, tono, inflexiones vocales, cambios repentinos de tono, diplofonía, resonancia, emisión nasal, margen y eficiencia vocal global.
Hirano (1981)	Protocolo GRBAS, donde los oyentes evalúan las voces sobre 5 escalas: ***G**rade*, ***R**ough*, ***B**reath*, ***A**sthenic*, ***S**trained*; puntuando de 0 a 3 cada una de ellas.
Gelfer (1988)	Propone un enfoque elaborado, usando 17 parámetros.
Hammarberg y Gauffin (1995)	Protocolo "Enfoque de Evaluación de Voz de Estocolmo", utilizando 13 parámetros: tensión, relajación, *breathiness*, *creakiness*, *roughness*, *grating*, *falsetto*, inestabilidad de tono, cambio de voz, afonía, diplofonía, tono y sonía.
Dejonckere et al. (1998)	Protocolo GIRBAS, aparecido a partir de la revisión del protocolo GRBAS, añadiendo una nueva escala de ***I**nstability* (inestabilidad).
Laver (1980, 2000)	Protocolo "Análisis de Perfiles Vocales", contiene más de 30 parámetros incluyendo características del tracto vocal y prosódicas.

Tabla 3.1: *Resumen de protocolos para la medida perceptiva de la cualidad de la voz*

Es en respuesta a las dificultades de criterio entre oyentes, que aparece la idea de reemplazar estas medidas por medidas objetivas de la función fisiológica, flujo de aire, o de la señal acústica. Un ejemplo del trabajo realizado en este aspecto es el desarrollado por Sáenz-Lechón et al. (2006), donde se presenta una alternativa para la evaluación automática de la VoQ de acuerdo a la escala GRBAS.

Según Gerratt y Kreiman (2003), el enfoque para la medida objetiva refleja como los oyentes son inherentemente incapaces de consensuar su percepción ante estímulos auditivos complejos, aunque aparecen dificultades tanto teóricas como prácticas

en la utilización de esta metodología. Teóricamente, no se puede conocer la importancia porcentual de aspectos particulares de la señal acústica sin disponer de medidas válidas de la respuesta perceptiva, debido a que la VoQ es, por definición, la respuesta perceptiva a un estímulo acústico particular. De este modo, medidas acústicas orientadas a la cuantificación de la cualidad vocal pueden solamente derivar su validez como medidas de la VoQ a partir de su asociación causal con la percepción auditiva. De forma práctica, no existen correlaciones sistemáticas claras entre medidas perceptivas e instrumentales de la voz, sugiriendo que tales medidas instrumentales no son índices estables de percepción. Finalmente, correlación no implica causalidad, ya que simplemente por conocer la relación de una variable acústica con la perceptiva no necesariamente ha de aclarar su contribución a la percepción de la VoQ. Es más, se da el caso de que aún siendo una variable acústica importante en la decisión tomada por el oyente sobre la VoQ, la naturaleza de dicha contribución no se muestra por un coeficiente de correlación. Además, dada la gran variabilidad existente en las estrategias perceptivas y en los hábitos que el oyente individual demuestra en su uso de escalas de clasificación tradicionales, la correlación global entre variables acústicas y perceptivas, promediadas a través de muestras de oyentes y de voces, falla a la hora de proveer elementos para el proceso perceptivo.

Como solución al problema anterior, en Gerratt y Kreiman (2001) se propuso una alternativa a esta problemática. Este trabajo propone la medida de la cualidad vocal a partir de preguntar a los oyentes sobre voz natural y sintética. Los oyentes, a partir de variar los parámetros de síntesis del habla, crean estímulos de una similitud auditiva aceptable a los naturales. Cuando un oyente elige la mejor correspondencia para un estímulo de prueba, la configuración paramétrica de síntesis representa la percepción de la VoQ del oyente. De este modo, comparando directamente la voz sintética con la natural, los oyentes no necesitan estándares para ninguna VoQ en particular, permitiendo a su vez centrar la atención en dimensiones acústicas individuales, reduciendo así la complejidad perceptiva de la tarea de evaluación y la variabilidad de la respuesta. Los resultados preliminares de este proceso mostraron una elevada concordancia entre las evaluaciones sobre VoQ de distintos oyentes, presumiblemente motivado por el hecho que este método de análisis-síntesis permite el control de la mayoría de fuentes de discrepancia en los criterios de decisión sobre la cualidad vocal, mientras que evitan el uso de escalas de dudosa validez.

Dedicar una mayor investigación en la búsqueda de un conjunto de parámetros acústicos, que permitiera caracterizar los distintos tipos de voces según sus características de VoQ, permitiría tanto a investigadores como a personal clínico, en el caso concreto de tratarse de patologías de la voz, la sustitución de “etiquetas” por los parámetros acústicos ligados a su percepción auditiva, cuyos niveles especifican a la VoQ de interés de forma objetiva y completa.

3.1.2.2. Medida acústica

La medida acústica tiene como objetivo el análisis de la voz mediante su estudio y parametrización. Keller (2005) presenta una serie de consideraciones a tener en

cuenta a la hora de hacer este tipo de medida, de forma que las medidas acústicas de la VoQ deben satisfacer una serie de requisitos:

- Diferencias perceptibles de la voz deberían de poder quedar reflejadas en variaciones predecibles en la forma de onda de la señal o en una o varias de sus derivadas.
- Las medidas deberían de reflejar estados o conjunto de estados del tracto vocal típico de un cierto individuo, y debería de ser separable de estados compartidos por un gran número de interlocutores, que son relevantes para la producción de segmentos fonéticos o de características prosódicas en una comunidad de interlocutores, también conocido como características lingüísticas.
- Debido a que la percepción de la VoQ refleja configuraciones supralaríngeas, laríngeas y sublaríngeas, o respiratorias, del tracto vocal, las medidas de la forma de onda del habla deberían capturar todos esos tipos de información y separarlos si es posible.

Satisfacer todos los requisitos expuestos a partir de una única medida no es una tarea sencilla, ya que medidas que suelen satisfacer uno de los requisitos tienden a fallar con el resto y, además, la evaluación de la VoQ probablemente requiere en última instancia la aplicación en paralelo de diferentes medidas. Las medidas de mayor relevancia utilizadas son:

- Espectro Medio a Largo Plazo —*Long-Term Average Spectrum*— (LTAS).
- Medidas acústicas relacionadas con la prosodia.
- Modelado de fuente.

Espectro Medio a Largo Plazo (LTAS)

El Espectro Medio a Largo Plazo —*Long-Term Average Spectrum*— (LTAS) es el espectro promediado de la amplitud o intensidad a lo largo de la cadena hablada en una gama de frecuencias seleccionada. Se obtiene haciendo la media de un gran número de análisis espectrográficos, p. ej. cada ms, de fragmentos de una muestra de habla y, opcionalmente, se resumen como un conjunto de bandas espectrales. Las diferencias medias entre los perfiles espectrales en el mismo fragmento de habla presumiblemente refleja las configuraciones a largo plazo, e idealmente esperadas, para capturar las diferencias de VoQ.

Típicamente, el LTAS se estabiliza después de 40 segundos, identificando de forma fidedigna las características que se mantienen prácticamente constantes en el tiempo tales como es el formante de un cantante. Sin embargo esta propuesta se ve afectada principalmente por cuatro limitaciones (Keller, 2005):

1. Largo plazo promedia características espectrales, relevantes para la información segmental, con aquellas más ligadas a la voz. Esto dificulta la comparación de diferentes fragmentos de habla, incluso fragmentos que léxicamente son iguales aunque pronunciados de forma distinta. En consecuencia aparece la tendencia de reemplazar el LTAS simple por una medida más localizada, p. ej. aplicar el LTAS únicamente sobre vocales núcleo (Klasmeyer, 2000) o reemplazarlo por promediados basados en un gran número de espectros obtenidos en el centro de las vocales núcleo (Keller, 2003).

2. El promediado no tiene en cuenta las dinámicas temporales que a menudo contribuyen a la definición de una VoQ dada, como por ejemplo las perturbaciones de frecuencia fundamental entre periodos consecutivos de la señal de voz cuando se evalúa el *jitter*.

3. Existen divergencias entre perfiles obtenidos a partir de señales con una amplitud elevada y baja.

4. La baja capacidad de obtención de diferencias entre interlocutores o grupos de interlocutores.

Como consecuencia, mientras que LTAS innegablemente ilustra ciertas diferencias en la VoQ, estas son aparentemente insuficientes para modificar el efecto de la VoQ o la particularidad de un interlocutor en un sistema de síntesis del habla. Por tanto, se debe asumir que existen características en la forma de onda acústica que contribuyen de manera importante en la percepción de una VoQ.

Medidas acústicas relacionadas con la prosodia

Vistas las limitaciones que presenta el LTAS, la VoQ no puede ser definida únicamente en términos de características acústicas estáticas. Tal y como indica Laver en su esquema (1980; 1991), elementos estructurados temporalmente se tienen en consideración con los indicadores acústicos estáticos como también parámetros que tradicionalmente han sido asociados a la prosodia:

- Frecuencia fundamental (F_0) para la percepción de tono (*pitch*).
- Intensidad para la percepción de la sonía (*loudness*).
- Duración, típicamente de vocales o sílabas.

La asignación de valores tanto de prosodia como de VoQ depende en parte de como se define esta última y del contexto, siendo la decisión de qué es VoQ un proceso complejo por cubrir gran variedad de conceptos. Por ejemplo, una combinación de F_0, intensidad y duración, junto con la suficiente información de contexto, puede hacer que un oyente deduzca la información lingüística y social relevante a partir de varias componentes acústicas. El oyente puede interpretar que una palabra es parte de una pregunta, que el interlocutor es un hombre, que los parámetros prosódicos son demasiado elevados para el interlocutor concreto, la lengua utilizada, el contexto

y que el valor de los parámetros de VoQ como el *jitter* es demasiado alto. Con todos estos indicadores se consigue extraer que se está hablando en inglés, que la palabra forma parte de una pregunta, que el interlocutor es un hombre y que posiblemente tenga un cierto grado de ansiedad.

Tal y como explica en su trabajo Keller (2005), en aplicaciones de síntesis de habla natural, la prosodia debería ser combinada con la manipulación adecuada de la VoQ. Existen estudios como los de Gobl et al. (2002) y Yanushevskaya et al. (2005) donde la frecuencia fundamental y los parámetros de VoQ fueron manipulados separada y conjuntamente para sintetizar diferentes tipos de voces, mostrando como las manipulaciones de VoQ contribuyeron en gran medida a la comunicación de la expresividad.

Por tanto, para la evaluación y el procesamiento automático de la VoQ individual, es importante examinar la contribución de los tres parámetros prosódicos clásicos en interacción con los indicadores de su significación lingüística y paralingüística. Estos parámetros deben ser combinados con los parámetros apropiados de VoQ para aumentar la naturalidad de la SHE.

Modelado de fuente

Aunque todas las partes del tracto articulatorio contribuyen en un grado u otro a la VoQ, la comunidad científica está de acuerdo en que ciertas condiciones afectan al flujo de aire en la glotis, es decir, la laringe o configuraciones de la fuente son responsables de aspectos principales de esta componente del habla. Las condiciones relevantes para la VoQ pueden ser de tres tipos:

- Transitoria: como sería el caso de *voice onsets*[1] y *voice offsets*[2].
- Corto plazo: por ejemplo la duración de una vocal.
- Largo plazo: afectando a todas las componentes sonoras del habla individual.

Son muchos los investigadores que han tratado de obtener la forma de onda glótica de forma fiable y automática, minimizando las molestias ocasionadas al interlocutor. No obstante, no es fácil recuperarla debido a que aparece el efecto del tracto vocal. Aun así, son múltiples las técnicas utilizadas para conseguir una aproximación de la forma de onda glótica (Sección 2.4.2.1) como el uso del EGG o cálculos sobre la forma de onda acústica del habla, es decir, los conocidos como métodos de filtrado inverso (Sección 2.4.2.2). Este último caso es de interés para aquellas aplicaciones relacionadas con el procesamiento del habla, ya que se puede aplicar directamente sobre grabaciones capturadas con un micrófono.

A partir de la señal glótica obtenida, se pueden calcular las características glóticas que parametrizan la VoQ del habla bajo evaluación. Estas características y parámetros son los presentados en la Sección 2.4.2.2.

[1] En fonética, el tiempo de *voice onset*, es una característica de la producción de las consonantes oclusivas. Se define como el tiempo transcurrido entre que una consonante oclusiva se produce y cuando empieza la excitación glótica (Lieberman y Blumstein, 1998).

[2] El tiempo de *voice offset* se define como el lapso de tiempo que pasa entre la finalización de un sonido, anterior a una oclusiva, y el inicio de la oclusión o *closure* de dicha oclusiva (Laver, 2002)

3.2. Habla y emociones

El habla es una característica humana que por sí sola es capaz de comunicar al oyente el estado emocional del interlocutor. En una conversación normal los interlocutores esperan la aparición de cierto grado de contenido emocional en la voz, con lo que es una parte esencial en el habla humana. La componente emocional, expresiva o afectiva del habla es principalmente no-léxica, a pesar de que otros elementos importantes de la comunicación sí que deben tenerse en consideración donde sea posible: el contexto, el contenido del mensaje, los gestos y la expresión facial.

3.2.1. Las emociones

3.2.1.1. Introducción

Las emociones, o en general los estilos expresivos, están interconectadas con el estado mental de una persona. En su definición más general, la emoción es un intenso estado mental que se presenta en el sistema nervioso y que evoca o bien una respuesta positiva o bien negativa. No obstante, las definiciones que se pueden encontrar en la bibliografía son múltiples, y variarán dependiendo del campo donde son definidas, por ejemplo la filosofía, la biología o la psicología. Aunque no existe un acuerdo general para su descripción, la bibliografía psicológica contiene algunas guías para parametrizar los distintos estados emocionales. En cuanto a las diferentes acepciones de esta palabra en el ámbito de la investigación que ocupa esta tesis, "habla y emoción" (del inglés *speech and emotion*), se buscan las relaciones entre ambos dominios.

En Cowie y Cornelius (2003) se presenta un estudio exhaustivo de términos y de conceptos relacionados con la emoción y el habla, proponiéndose abordar la descripción de las emociones, en la disciplina de las tecnologías del habla, sin esperar una solución completa aportada desde cualquier otra.

En primer lugar se trata el término de "emoción plena" (Scherer, 1999), al que otros autores se refieren como "emociones primarias" (Plutchik, 2001) o "emociones básicas" (Ekman, 1999). Con estos términos se denota la forma más intensa de las emociones, estando presentes todos los aspectos considerados relevantes de una emoción en concreto, tales como la evaluación de la situación, los acontecimientos previos, la respuesta conductual, los aspectos psicológicos y las señales universales distintivas.

En segundo lugar se presenta la "emoción subyacente", que denota una clase de colorido emocional presente en todos los estados mentales. A pesar de que las emociones subyacentes son más comunes en la comunicación humana que las plenas, su descripción no es trivial en absoluto.

Finalmente, se denominan "estados emocionales" a toda la variedad de estados que van desde las emociones subyacentes más débiles hasta las emociones plenas. Este abanico de posibilidades contiene todo un conjunto de estados intermedios que tienen sentido en el ámbito de la comunicación humana. Asimismo, se introduce el

concepto de estados relacionados con la emoción, los cuales presentan ciertos aspectos propios de las emociones sin ser una de ellas. Estos estados se manifiestan en la persona como una cierta actitud, tal como el humor o la excitación.

3.2.1.2. Teoría de las emociones plenas

Cabe destacar que las emociones, tal y como plantea Scherer (1986), son la interfaz utilizada por un organismo para interactuar con el medio que le rodea. Para ello, son tres las funciones principales que las emociones desarrollan:

- Valoran la situación.
- Producen cambios fisiológicos y psicológicos preparando al organismo para una acción determinada.
- Comunican la reacción a otros individuos mediante comportamiento expresivo facial, corporal y oral.

En el ámbito de las emociones en psicología, las teorías contemporáneas indican cuatro perspectivas básicas para definir, estudiar y explicar las emociones. Estas teorías van desde las primeras aproximaciones de Charles Darwin hasta las teorías de finales del siglo XX. Cornelius (2000) define estas cuatro perspectivas como: Darwiniana, Jamesiana, Cognitiva y Constructivista Social.

- **Darwiniana.** Los inicios de la perspectiva Darwiniana se remontan a 1872, cuando Charles Darwin escribió el libro *The Expression of Emotion in Man and Animals* (Darwin, 1872). La idea básica es que las emociones son fenómenos desarrollados como funciones importantes de supervivencia, seleccionadas como tal para solucionar ciertos problemas a los que la especie humana ha tenido que hacer frente.
- **Jamesiana.** Fue inspirada por los escritos de William James sobre la emoción (James, 1884). Según James, los cambios corporales siguen directamente la percepción de una excitación, y la emoción es el sentimiento experimentado al aparecer estos mismos cambios.
- **Cognitiva.** La aproximación cognitiva moderna se basa en los estudios de las emociones realizados por Arnold (1960), y es la más dominante de las cuatro debido a que ha sido minuciosamente incorporada dentro de las otras tres. El eje central de esta perspectiva es que la emoción y el pensamiento son inseparables y, más específicamente, todas las emociones son enjuiciadas mediante un proceso de evaluación, que consiste en discernir qué acontecimientos del entorno son tomados como buenos o malos por nosotros.
- **Contructivista Social.** Tal y como recoge Cornelius (2000), Averill (1980) propone que las emociones no son remanentes de nuestro pasado psico-genético, ni pueden ser explicadas en términos estrictamente psicológicos. Más bien son construcciones sociales, que solo pueden ser plenamente entendidas a partir de un análisis social.

3.2.1.3. Descripción de las emociones

Son muchas las teorías sobre la emoción, en especial las que siguen las tradiciones Darwiniana y Jamesiana, que utilizan el concepto de emociones básicas, a partir de las cuales se generan todas las demás mediante variaciones o combinaciones de estas. No existe un criterio único para definir el conjunto básico de emociones. Las 4 emociones básicas más aceptadas, considerándose directamente ligadas a procesos biológicos, son: la alegría (*happiness*), la tristeza (*sadness*), el enfado (*anger*) y el miedo (*fear*).

La mayor parte de teorías coinciden en que el número de emociones básicas es inferior a 10, a pesar de que estudios más recientes definen entre 10 y 20 (Cowie y Cornelius, 2003). Por otro lado, hay que destacar el término "*The Big Six*" (Cornelius, 2000), en el que se engloba al conjunto formado por la alegría (*happiness*), la tristeza (*sadness*), el miedo (*fear*), el asco (*disgust*), el enfado (*anger*) y la sorpresa (*surprise*). Hay que tener en cuenta que las emociones que forman parte de estos conjuntos, denominadas emociones plenas, básicas o primarias, se consideran fundamentales ya que representan los patrones relacionados con la supervivencia del individuo y porque el resto de emociones derivan de ellas.

Ekman (1999) propone el concepto de familias de emociones, ya que se considera que cada emoción no es un único estado afectivo, sino una familia de estados relacionados. En total son 15 las emociones básicas o familias propuestas. Cada familia se caracteriza por un tema, fruto de la evolución, y unas variaciones, reflejo del aprendizaje. Por ejemplo, la familia del enfado abarcaría emociones como enojo, enfado y rabia; todas con un tema común, pero con diferentes matices fruto de elementos adquiridos previamente.

Es interesante destacar que existe un número importante de términos que describen "estados relacionados con la emoción", como por ejemplo confiado, relajado o aburrido, hecho que refleja el sentido generalizado de que estos constituyen una parte significativa de la vida emocional diaria. Por tanto, podemos concluir que las teorías sobre las emociones se refieren mayoritariamente a emociones básicas, y consideran que las demás emociones son combinaciones o modificaciones de estas, pese a que no hay un consenso claro sobre cuáles deben ser las llamadas así.

Referente a la representación que las emociones pueden tener, especialmente una vez visto como a partir de la combinación de diferentes factores aparecen nuevas emociones, algunos investigadores han concluido que las emociones se pueden representar mediante una estructura circular. La proximidad de dos categorías representa emociones conceptualmente similares, mientras que las emociones contrarias están separadas 180 grados. Tal y como describe Tryon (1997), Plutchik (1958) propuso un modelo con 8 emociones básicas bipolares: alegría-aburrimiento, enfado-miedo, aceptación-asco y sorpresa-expectación. La evolución de esta teoría ha llevado al modelo circunflejo tridimensional (Plutchik, 2001), presentado en la Figura 3.1, en el que se representan 4 aspectos:

1. **Intensidad**: dimensión vertical.

2. **Similitud**: el círculo representa los grados de similitud entre emociones, de forma que las similares están próximas y las opuestas están separadas 180 grados.

3. **Emociones básicas bipolares**: representadas por cada uno de los ocho sectores.

4. **Mezcla de emociones**: representada por cada uno de los espacios en blanco.

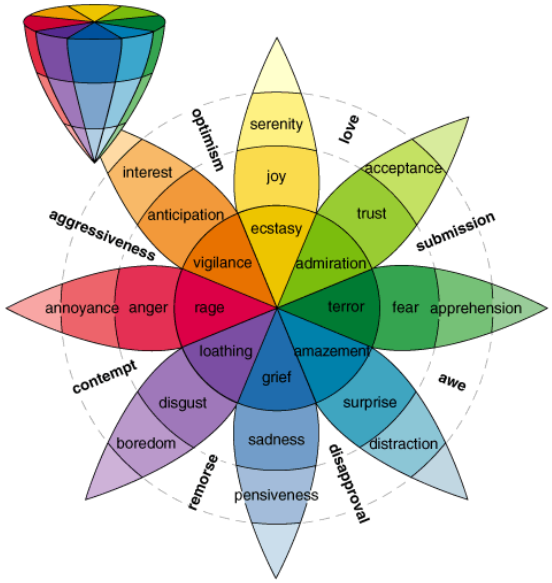


Figura 3.1: *Modelo circunflejo tridimensional de Plutchik (2001)*

Por último, el hecho fundamental para disponer de una descripción sistemática de las emociones, consiste en poder representar los estados emocionales como coordenadas en un espacio a partir de un pequeño número de dimensiones. Se han llevado a cabo numerosas investigaciones que abordan este tipo de clasificación, pudiendo encontrar los trabajos de Cowie y Cornelius (2003) y Schröder (2004) donde se profundiza en los estudios realizados con este fin. La mayor parte de los estudios intentan representar el espacio emocional en dos dimensiones, aunque algunos añaden una tercera. También la terminología asociada a cada eje presenta algunas diferencias según el estudio. A continuación se presentan las tres dimensiones más utilizadas junto con diferentes términos para referirse a ellas:

- **Evaluación / agrado / valoración**: corresponde al eje “Positivo-Negativo”, clasificando las emociones según lo placentero o desagradable de estas (p. ej. desde la alegría hasta el enfado).

- **Activación / actividad**: corresponde a la escala “Activo-Pasivo”, indicando la presencia o ausencia de energía o tensión (p. ej. desde estar furioso a estar aburrido).

- **Potencia / fuerza**: corresponde a la escala "Dominante-Sumiso", distinguiendo emociones iniciadas por el sujeto de aquellas causadas por el entorno (p. ej. desde el desprecio al temor o a la sorpresa).

Aquellas emociones con una actividad similar, como puede ser el caso de la alegría o del enfado, se confunden más entre sí que emociones con valoración o fuerza parecida.

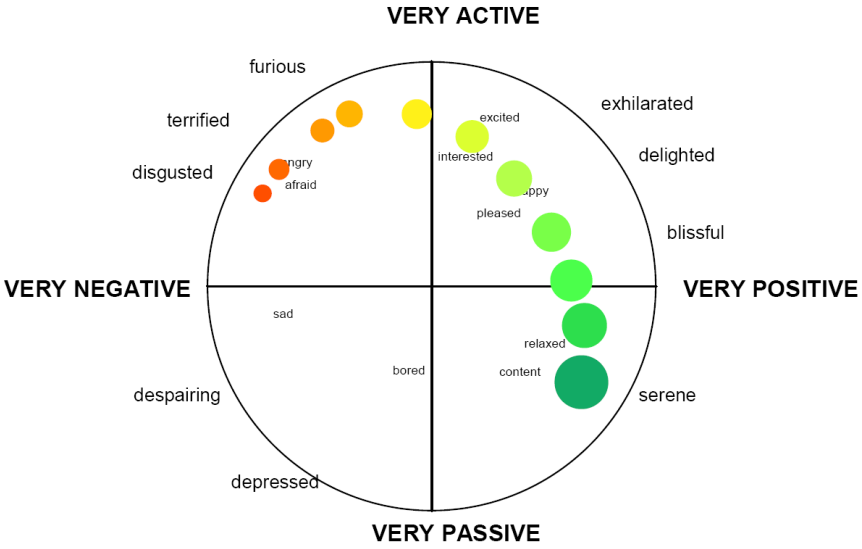


Figura 3.2: *Imagen de la herramienta Feeltrace (Cowie et al., 2000)*

Esta representación del espacio emocional es muy utilizada, pudiéndose destacar la herramienta *Feeltrace* (Cowie et al., 2000) (Figura 3.2). Esta herramienta permite el etiquetado de un estímulo sonoro o visual en dos dimensiones emocionales: evaluación en el eje horizontal y activación en el eje vertical.

3.2.2. Parámetros del habla y la transmisión de emociones

El habla es una característica humana que por sí sola es capaz de comunicar al oyente el estado emocional del interlocutor. En una conversación normal los interlocutores esperan la aparición de cierto grado de contenido emocional en la voz, con lo que es una parte esencial en el habla humana. La componente emocional, expresiva o afectiva del habla es principalmente no-léxica, a pesar de que otros elementos importantes de la comunicación sí que deben tenerse en consideración donde sea posible: el contexto, el contenido del mensaje, los gestos y la expresión facial.

Podemos hablar de efectos fisiológicos en el habla (acústicos, prosódicos y léxicos) y en el lenguaje corporal (gestos, expresión facial y movimientos corporales). Se parte de la hipótesis de que la voz sufre cambios acústicos causados directamente por alteraciones fisiológicas cuando una persona se encuentra en un determinado estado emocional (Scherer, 1986). Por ejemplo, cuando sentimos rabia o miedo se produce una activación del sistema nervioso simpático, provocando cambios en el organismo

como un incremento de la presión arterial o de la frecuencia cardíaca, temblores, sequedad de boca, entre otros. Estos cambios fisiológicos provocan cambios en el habla y en la expresión facial. Por lo tanto, la investigación en el campo de la expresividad emocional requerirá de modelos acústicos consistentes en la definición de los parámetros del habla y su cuantificación para cada uno de los estados emocionales.

Tanto la prosodia como la VoQ son parámetros utilizados en la representación del contenido emocional del habla (Cowie et al., 2001; Gobl y Ní Chasaide, 2003) y, a pesar de que la VoQ ha sido menos estudiada que la prosodia, trabajos recientes proponen ambas informaciones para mejorar el modelado acústico del habla expresiva (Cabral y Oliveira, 2005; Iriondo et al., 2007b). La aplicación de VoQ para la obtención de mejoras ha sido tanto en aplicaciones de reconocimiento automático del habla como de Conversión de Texto en Habla (CTH). Ejemplos de esta investigación los encontramos en estudios sobre reconocimiento de emociones (Cowie et al., 2001) o transformación de expresividades en la síntesis del habla (Drioli et al., 2003; Türk et al., 2005).

En esta sección se muestran las relaciones existentes en la bibliografía entre la prosodia y la VoQ con respecto a la transmisión de emociones. En primer lugar, en base al trabajo presentado por Iriondo (2008), se pasan a definir las propiedades acústicas de los sonidos del habla relacionadas con la expresividad vocal.

1. **Propiedades relacionadas con la melodía**[3]**:**

   - **Frecuencia fundamental (F_0).** Se define como el ciclo periódico de la señal de voz, siendo el resultado de la vibración de los pliegues vocales. Su medida habitual es el hercio (Hz), que da una medida de los ciclos por segundo.
   - **Curva de F_0 o melódica.** Se trata de la secuencia de valores de F_0 para una elocución, y se relaciona con la percepción de la entonación[4] del habla.
   - ***Jitter*.** Parámetro que caracteriza la perturbación de F_0 debida a fluctuaciones en los tiempos de apertura y de cierre de los pliegues vocales de un ciclo al siguiente (Sección 2.4.2.4).

2. **Propiedades relacionadas con la intensidad:**

   - **Intensidad.** Medida de la energía de la onda acústica. Habitualmente se utiliza una transformación logarítmica de la amplitud de la señal, llamada decibelio (dB), que representa mejor la percepción humana del sonido.
   - ***Shimmer*.** Parámetro que caracteriza la perturbación en la intensidad debida a fluctuaciones en la amplitud de un ciclo al siguiente (Sección 2.4.2.4).

3. **Propiedades relacionadas con los aspectos temporales del habla:**

[3] La melodía (*pitch* en inglés) es el fenómeno que se relaciona con la curva de frecuencia fundamental (F_0) o curva melódica de un grupo fónico (Garrido, 1991).

[4] La entonación es la sensación perceptiva que producen, fundamentalmente, las variaciones de tono a lo largo de un enunciado (Gil, 2007).

- **Velocidad del habla.** Se mide a partir de la duración de los segmentos del habla o como el número de unidades lingüísticas por unidad temporal (p. ej. sílabas por segundo).
- **Pausas.** El número y la duración de los silencios en la señal de voz es un parámetro del que habitualmente se realiza su medida.

4. **Propiedades relacionadas con el timbre:**

   - **Energía de alta frecuencia.** Proporción relativa de la energía por encima de una frecuencia de corte respecto a la energía total.
   - **Frecuencias de los formantes.** Se trata de regiones de frecuencia que presentan una mayor concentración de energía. También puede definirse como cada una de las resonancias del tracto vocal. Se suelen representar por la frecuencia central de la región y su ancho de banda.
   - **Precisión en la articulación.** Mide la desviación de las frecuencias de los formantes en las vocales desde las frecuencias formantes neutras (Juslin y Laukka, 2003).

Aunque *jitter* y *shimmer* estén relacionados con la frecuencia fundamental (F_0) y la intensidad respectivamente, no se asocian a propiedades prosódicas. Junto a este conjunto de propiedades que forman las perturbaciones de F_0 y de intensidad, se tienen aquellas propiedades relacionadas con el timbre que se conocen como VoQ, y que relacionan la energía en diferentes bandas frecuenciales, la frecuencia de los formantes y la desviación de estas.

Existen numerosos estudios sobre la correlación entre habla y emoción. En Murray y Arnott (1993) se presenta un resumen de los trabajos más significativos en la bibliografía sobre el habla y la emoción, en el que para cada una de las emociones se indican los efectos más comúnmente asociados respecto del estilo de habla neutro (resumido y traducido en la Tabla 3.2). Murray y Arnott (1993) distinguieron entre emociones primarias (enfado (*anger*), alegría (*happiness*), tristeza (*sadness*), miedo (*fear*) y asco (*disgust*)) y emociones secundarias (pena (*grief*), ternura (*tenderness*), ironía (*irony*) y sorpresa (*surprise*)).

La VoQ que aparece en la Tabla 3.2 hace referencia a la característica de la voz relacionada con el tipo de fonación (Sección 2.1.3) y no a los parámetros (Sección 2.4.2). El enfoque de qué tipo de emociones se tratan en este resumen es suficiente para la aplicación en la que se desarrolló el trabajo, síntesis del habla, pero no lo es en estudios que tratan de averiguar qué signos genera un ser humano cuando transmite estados emocionales, ya que este pasa por multitud de estados utilizando gran variedad de formas de expresión (Cowie et al., 2001). Por tanto, es imprecisa en cuanto a la cuantificación de los parámetros del habla que aparecen, ya que para la obtención de modelos acústicos de las emociones se necesitan enfoques con un mayor nivel de precisión (Iriondo, 2008).

Cowie et al. (2001) recogen en su investigación el trabajo realizado en reconocimiento de emociones para la interacción hombre-máquina. En la Tabla 3.3 se resumen

	Enfado	Alegría	Tristeza	Miedo	Asco
Velocidad del habla	Ligeramente más rápida	Más rápida o más lenta	Ligeramente más lenta	Mucho más rápida	Mucho más rápida
F_0 promedia	Mucho más alta	Mucho más alta	Ligeramente más baja	Mucho más alta	Mucho más baja
Margen de F_0	Muy amplio	Muy alto	Ligeramente más estrecho	Muy amplio	Ligeramente más amplio
Intensidad	Más alta	Más alta	Más baja	Normal	Más baja
VoQ	Jadeante	Estrepitosa	Resonante	Sonoridad irregular	Ruidosa
Cambios de F_0	Abruptos en sílabas tónicas	Suaves inflexiones ascendentes	Inflexiones descendentes	Normal	Amplios en inflexiones descendentes finales
Articulación	Tensa	Normal	Arrastrada	Precisa	Normal

Tabla 3.2: *Traducción del resumen de Murray y Arnott (1993) acerca de los efectos de las emociones sobre el habla*

las características relacionadas con la VoQ para cinco de las seis de las emociones llamadas "*The Big Six*" (Cornelius, 2000), presentadas en la Sección 3.2.1.3: enfado (*anger*), alegría (*happiness*), tristeza (*sadness*), miedo (*fear*) y sorpresa (*surprise*). Los parámetros acústicos considerados prosódicos: frecuencia fundamental (F_0), intensidad y duración, están ligados con la VoQ ya que su variación aportará información sobre ella (p. ej. variaciones de F_0 informarán sobre el valor del *jitter*).

Juslin y Laukka (2003) recopilaron los indicadores acústicos más utilizados para expresar emociones de forma discreta. El hecho de comparar multitud de estudios con datos cuantitativos no uniformes lo soluciona el autor agrupando los resultados obtenidos en categorías: alto/a, medio/a, bajo/a, arriba, sube, baja, rápido, lento, ancho, estrecho, empinado, equilibrado, grande, pequeño, regular o irregular. En la Tabla 3.4 se muestra la asignación de categorías, marcando en **negrita** la mayoritaria para cada par parámetro-emoción y en *cursiva* el número total de estudios en los que se ha visto envuelto el parámetro. Las emociones implicadas en el estudio son enfado (*anger*), miedo (*fear*), alegría (*happiness*), tristeza (*sadness*) y ternura (*tenderness*). El parámetro mayoritariamente analizado es el valor medio de F_0, seguido por su variabilidad, velocidad del habla y la intensidad; siendo los rasgos prosódicos, F_0, velocidad del habla e intensidad los más analizados en los estudios de expresión vocal. Se comprueba como los parámetros relacionados con la VoQ han sido menos estudiados que los prosódicos.

	Enfado	Alegría	Tristeza	Miedo	Sorpresa
F_0	Incremento de la media, mediana, margen, variabilidad	Incremento de la media, margen, variabilidad	Debajo de la F_0 media normal, margen de F_0	Incremento en la F_0 media, margen de F_0, perturbación, variabilidad del movimiento de F_0	Margen amplio, medio/-normal/alto
Intensidad	Elevada	Mayor	Menor	Normal	– –
Duración	Margenes altos/reducidos	Margen elevado, tempo lento	Ligeramente lento, caídas largas de *pitch*	Márgenes elevados/reducidos	Tempo normal/moderado
Espectro	Elevado en el punto medio del espectro promedio para zonas sin fricción	Incremento de la energía de alta frecuencia	Disminución de la energía de alta frecuencia	Aumento de la energía de alta frecuencia	– –
Manera de realización	Tensa, *breathy*, tono profundo, atronador	Tensa, *breathy*, atronadora	Relajada, resonante	Tensa	*Breathy*
Otros	Habla cortada, ritmo fundamental de apertura y cierre irregular, gestos articulatorios para la alternancia vocal/consonante	Distribución irregular de acentos, alternancia a voluntad del nivel de las sílabas acentuadas	Arrastrando, ritmo con pausas irregulares	Articulación precisa de vocal/consonante, irregularidad en la sonorización debido al trastorno del patrón respiratorio	– –

Tabla 3.3: *Relación entre la cualidad de la voz y emociones (traducida de Cowie et al. (2001))*

Parámetro	Categoría	Enfado	Miedo	Alegría	Tristeza	Ternura
	Indicadores vocales relacionados con la F_0					
F_0 media	Alta	**33**	**28**	**34**	4	1
	Media	5	8	2	1	0
	Baja	5	3	2	**40**	**4**
	TOTAL	*43*	*39*	*38*	*45*	*5*
Variabilidad de F_0	Alta	**27**	9	**33**	2	0
	Media	4	6	2	1	0
	Baja	4	**17**	1	**31**	**5**
	TOTAL	*35*	*32*	*36*	*34*	*5*
Contorno de F_0	Sube	**6**	**6**	**7**	0	1
	Baja	2	0	0	**11**	**3**
	TOTAL	*8*	*6*	*7*	*11*	*4*
Perturbación de F_0	Alta	**6**	4	**5**	1	0
	Baja	1	4	3	**5**	0
	TOTAL	*7*	*8*	*8*	*6*	*0*
	Indicadores vocales relacionados con la intensidad					
Intensidad media	Alta	**30**	**20**	**11**	1	0
	Media	1	3	6	2	0
	Baja	1	8	0	**29**	**4**
	TOTAL	*32*	*22*	*26*	*32*	*4*
Variabilidad de intensidad	Alta	**9**	**7**	**8**	2	0
	Media	1	4	3	1	0
	Baja	2	1	2	8	0
	TOTAL	*12*	*12*	*13*	*11*	*0*
Voice onsets	Rápido	1	1	**2**	1	0
	Lento	1	1	0	1	1
	TOTAL	*2*	*2*	*2*	*2*	*1*
	Indicadores vocales relacionados con aspectos temporales					
Velocidad del habla	Rápida	**28**	**24**	**22**	1	0
	Media	3	3	5	5	1
	Lenta	4	2	6	**30**	**3**
	TOTAL	*35*	*29*	*33*	*36*	*4*

Continúa en la página siguiente

Tabla 3.4 – continúa de la página anterior

Parámetro	Categoría	Enfado	Miedo	Alegría	Tristeza	Ternura
Proporción de pausas	Grande	0	2	1	**11**	**1**
	Media	0	3	2	0	0
	Pequeña	**8**	**4**	**3**	1	0
	TOTAL	*8*	*9*	*6*	*12*	*1*
Regularidad microestructural	Regular	0	0	**2**	0	1
	Irregular	**3**	**2**	0	**4**	0
	TOTAL	*3*	*2*	*2*	*4*	*1*
	Indicadores vocales relacionados con la VoQ					
Energía de alta frecuencia	Alta	**22**	**8**	**13**	0	0
	Media	0	2	3	0	0
	Baja	0	6	1	**19**	**3**
	TOTAL	*22*	*16*	*17*	*19*	*3*
Media del formante F_1	Alta	**6**	1	**5**	1	0
	Media	0	0	1	0	0
	Baja	0	**3**	0	**5**	0
	TOTAL	*6*	*4*	*6*	*6*	*0*
Ancho de banda del formante F_1	Estrecho	**4**	0	**2**	0	0
	Ancho	0	**2**	1	**3**	0
	TOTAL	*4*	*2*	*3*	*3*	*0*
Precisión de la articulación	Alta	**7**	2	**3**	0	0
	Media	0	2	2	0	0
	Baja	0	2	0	**6**	**1**
	TOTAL	*7*	*6*	*5*	*6*	*1*
Forma de onda glótica	Empinada	**6**	2	**2**	0	0
	Equilibrada	0	**4**	0	**4**	0
	TOTAL	*6*	*6*	*2*	*4*	*0*

Tabla 3.4: *Estudios realizados para cada uno de los pares parámetro-emoción, indicando en* ***negrita*** *la asignación mayoritaria y en cursiva el total de los estudios (traducida de Laukka (2004))*

Por último, se puede relacionar el tipo de fonación con el estilo de habla expresivo. Esto es interesante desde el punto de vista de que el tipo de fonación está íntimamente ligada a la VoQ, dado que la fonación constituye el conjunto de los parámetros de VoQ (Sección 2.1.3). Gobl y Ní Chasaide (2003) demuestran que cambios

en la VoQ pueden evocar diferencias en la expresividad del hablante. Las pruebas se realizaron presentando 7 tipos de fonación o VoQ al oyente para que decidiera su relación con 8 parejas de expresividades opuestas, puntuando la percepción de cada una de las expresividades en una escala de 0 a ± 3, pudiendo así conocer la relación entre diferentes tipos de VoQ (tipos de fonación) y expresividades (Figura 3.3). De los resultados también se desprende que a diferencia del mapeo uno-a-uno, una VoQ es multicolor en términos de expresividad, asociándose normalmente con un conjunto de atributos relacionados. Se muestra como los atributos asociados con los estímulos *tense*/*harsh* tienen una alta activación y/o potencia, pero incluyen expresividades con una evaluación positiva (seguro (*confident*), interesado (*interested*), alegre (*happy*)) y negativa (enfadado (*angry*), estresado (*stressed*)). El otro grupo de estímulos, los no modales (*breathy*, *whispery*, *creaky* y especialmente los estímulos sonoros *lax-creaky*), están asociados con los atributos que tienen una baja activación pero tanto una evaluación positiva (relajado (*relaxed*), contento (*content*), íntimo (*intimate*), amistoso (*friendly*)) como negativa (triste (*sad*), aburrido (*bored*)).

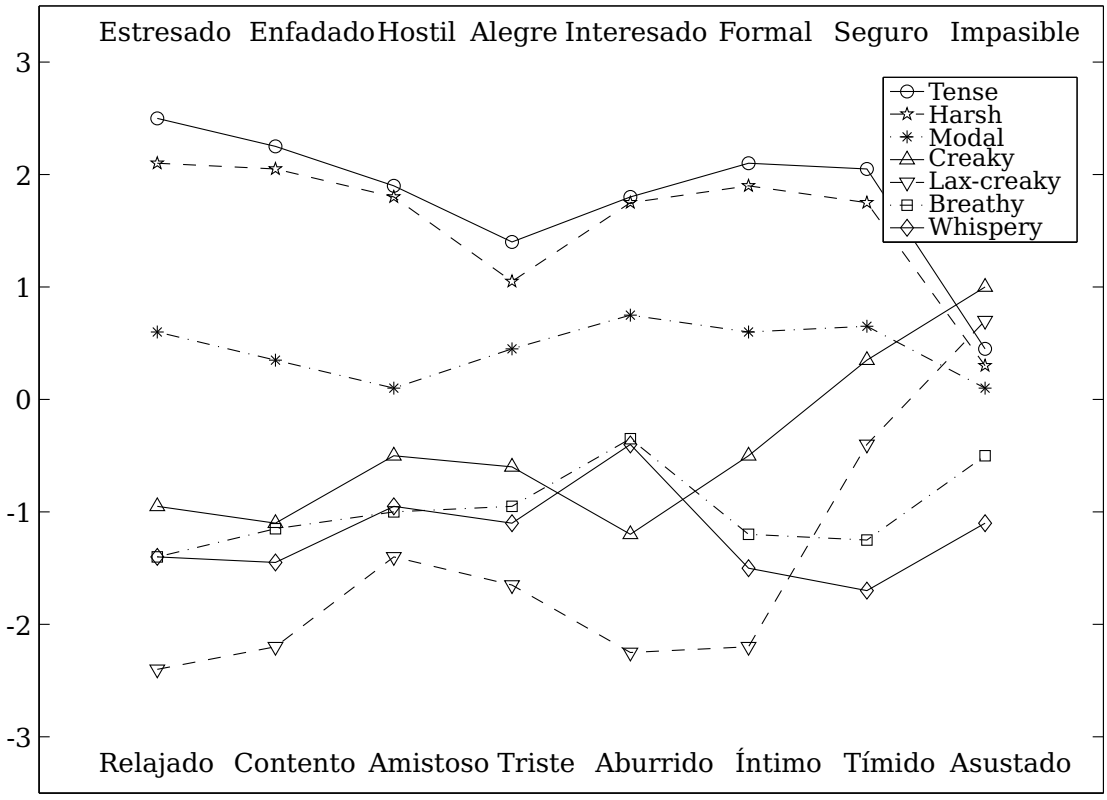


Figura 3.3: *Relación entre cualidad de la voz y expresividad (traducido de Gobl y Ní Chasaide (2003))*

3.3. Conversión de texto en habla

En los siguientes apartados se presentan los sistemas de Conversión de Texto en Habla (CTH), también llamados sistemas de síntesis del habla. Se introduce al lector en la temática y se presenta el caso del grupo de investigación GTM, en el que ha sido desarrollado este trabajo de tesis.

3.3.1. Introducción

Un sistema de Conversión de Texto en Habla (CTH), o *Text-to-Speech* (TTS) en inglés, es una herramienta, tanto de software como de hardware, que transforma un texto de entrada en la señal de voz correspondiente a ese texto maximizando la calidad de la forma de onda. El texto de entrada puede provenir de un correo electrónico, de una web o bien puede ser escrito directamente desde un teclado. Algunas de las aplicaciones de este tipo de sistemas son la ayuda a discapacitados, soporte para el aprendizaje de lenguas, aplicaciones telefónicas, aplicaciones multimedia e interfaces hombre-máquina en general.

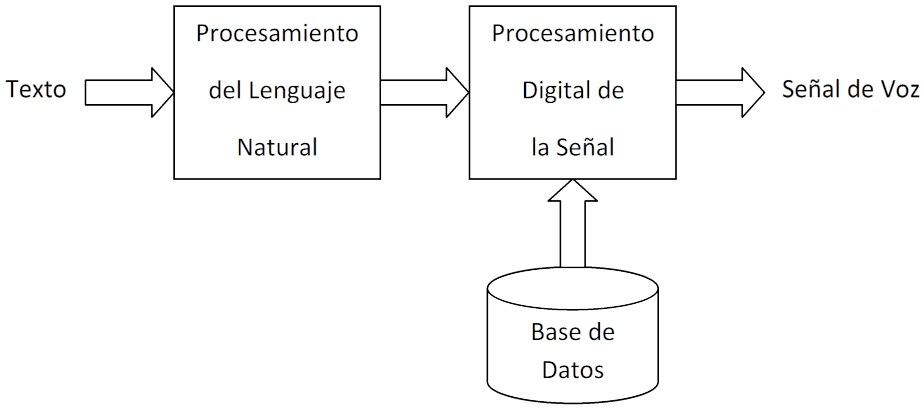


Figura 3.4: *Diagrama de bloques de un sistema de conversión de texto en habla*

Según el enfoque presentado por Dutoit y Stylianou (1997), un sistema de CTH se divide principalmente en dos bloques (Figura 3.4):

1. **Bloque de Procesamiento del Lenguaje Natural (PLN)**, se trata del bloque encargado de analizar el texto de entrada al sistema para obtener su transcripción fonética y sus características (como por ejemplo lingüísticas o prosódicas). El primer módulo del PLN es el preprocesador, encargado de normalizar el texto de entrada de forma que los siguientes bloques interpreten correctamente esta información textual. En este preprocesamiento se realizan tareas como la de pasar de números a letras o expansión de acrónimos, es decir, tratamiento en general de lo que se conoce como Palabra no Estándar —*Non-Standard Word*— (NSW) (Sproat et al., 1999) de la lengua de interés. A continuación el texto se pasa por el analizador morfosintáctico que se encarga de subdividir los textos en

grupos sintácticos, para una vez analizados aplicarles el proceso de conversión grafema-fonema. Finalmente se obtiene información prosódica (intensidad, duraciones y frecuencia fundamental) de las unidades fonéticas que corresponden al texto de entrada. A parte de esta información prosódica, puede ser añadida información de VoQ que la complemente.

2. **Bloque de Procesamiento Digital de la Señal (PDS)**, es el módulo encargado de generar las muestras de señal de voz a partir de la información obtenida del PLN (la transcripción fonética, prosódica y de VoQ en el caso que se utilice). Para la generación del mensaje oral final se lleva cabo el proceso de síntesis, del que existen diferentes estrategias (Sección 3.3.2): por formantes, articulatoria, concatenativa, basada en Modelo Oculto de Markov —*Hidden Markov Model*— (HMM), basada en Modelo Armónico más Ruido —*Harmonic plus Noise Model*— (HNM) e híbrida.

De las diferentes técnicas existentes para el bloque PDS, la más utilizada ha sido la síntesis concatenativa. En cuanto a las unidades utilizadas (por ejemplo fonemas, difonemas, trifonemas, alófonos o semisílabas), los difonemas (junto con trifonemas) son las más extendidas. Un difonema es una unidad sonora que comienza en la parte estable de un fonema y acaba en la parte estable del fonema que le sigue, de forma que las concatenaciones se realizan por las zonas más estables. En el caso de los trifonemas se tiene que el fonema central tiene unas características que lo hacen poco estable de modo que hacen que sea recomendable tenerlo grabado en ese contexto particular, ya que el uso de difonemas no garantizaría la calidad de la síntesis generada (por ejemplo el caso de consonante + vibrante + vocal). En cuanto a la síntesis concatenativa por selección de unidades, se tiene el caso donde se utilizan unidades mayores, de modo que la calidad de la síntesis mejora en naturalidad (Alías et al., 2005). El uso de uno u otro sistema depende principalmente del dominio de aplicación. Si deseamos un sistema de CTH que pueda cubrir cualquier dominio, sintetizando por tanto cualquier texto de entrada, se deberá trabajar con difonemas, mientras que si estamos en un dominio restringido, p. ej. meteorología, se podrá crear un corpus tal que las unidades que se utilicen sean mayores que el difonema y así la calidad se incremente. Por último se pueden combinar ambas estrategias, es decir se puede disponer de un corpus para una aplicación de dominio restringido y disponer de los difonemas necesarios para asegurar poder sintetizar cualquier texto de entrada (Alías et al., 2005).

Finalmente, también se puede hacer una última distinción en los sistemas de síntesis, no tanto por la metodología utilizada para la generación de la forma de onda sino por el hecho de que la voz generada transmita cierta emotividad o expresividad. Hay diferentes aproximaciones al problema, desde la creación de diferentes corpus expresivos de forma que en función de la emoción o expresividad que queramos transmitir se utilizará uno u otro (Iriondo et al., 2007a), hasta la conversión de una expresividad en otra a partir de las modificaciones de parámetros de la voz, como son la prosodia (Iriondo et al., 2007c), los de cualidad de la voz (Drioli et al., 2003) o una combinación de ambos (Cabral y Oliveira, 2006).

3.3.2. Estrategias de síntesis

En esta sección se presentan las principales estrategias de síntesis aparecidas a lo largo de los años y de las que los sistemas de CTH han hecho uso: formantes, articulatoria, concatenativa, basada en HMM, basada en HNM e híbrida.

3.3.2.1. Síntesis por formantes

La síntesis por formantes fue la primera técnica de síntesis en ser desarrollada y fue la dominante hasta los primeros años de la década de los 80. A menudo se la llamaba "síntesis por reglas" debido a que se generaba habla de la nada; mientras que en esa época se reservaba el término "síntesis" para hacer referencia a la regeneración de una forma de onda, que previamente había sido parametrizada, en aplicaciones de codificación de habla.

La síntesis por formantes adopta el siguiente enfoque al problema de la síntesis: modular, basado en modelo y acústico-fonético. Se hace uso de un modelo acústico de tubo de un modo particular, ya que los elementos de control del tubo se pueden relacionar fácilmente con propiedades acústico-fonéticas. A modo de ejemplo, en la Figura 3.5 se muestra un esquema básico de la síntesis por formantes, donde se observa como el sonido se genera en la fuente, periódica para sonidos sonoros y ruido blanco para sonidos obstruyentes[5]. Esta señal base se introduce en el modelo de tracto vocal, haciendo diferencia en el modelado de las cavidades oral y nasal, de modo que la señal pueda pasar por esta última únicamente en el caso que el sonido requiriera de ser nasalizado. Por último, la salida de estas componentes se combina y se pasa a través de una componente de radiación que simula las características de radiación de los labios y de la nariz.

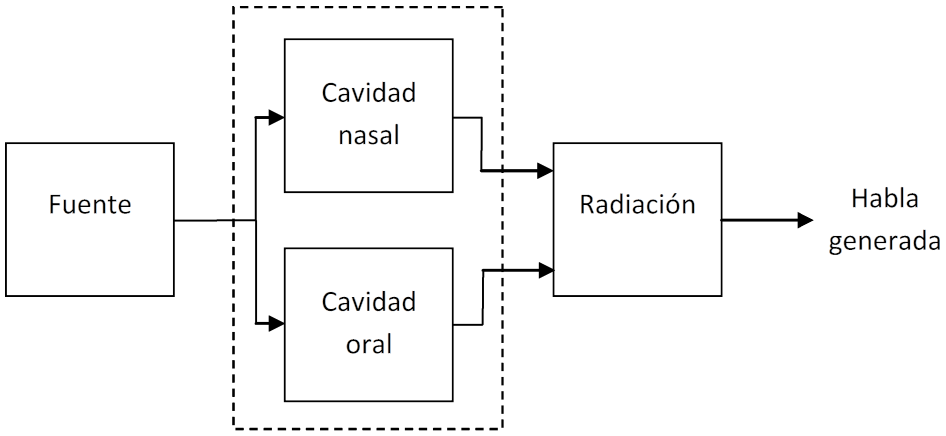


Figura 3.5: *Diagrama básico de la síntesis por formantes*

Los formantes se modelan individualmente, permitiendo así al diseñador del sistema controlar a cada uno de ellos, ya que en la época en la que estos sintetizadores

[5]Sonido que se articula obstruyendo el paso del aire por el tracto vocal, como por ejemplo sonidos oclusivos, africados o fricativos.

fueron desarrollados era mucho más sencillo leer valores reales de formantes a partir de un espectrograma que determinar configuraciones del tracto vocal. Por esta razón, un sintetizador por formantes no es un buen modelo de tracto vocal.

La síntesis por formantes tiene mucho en común con el modelo todo-polos de tracto vocal. Así como sucede en el modelo de tubo, este sintetizador es modular respecto a la fuente y filtro del tracto vocal. Los componentes de la cavidad oral están formados por entre 3 y 6 resonadores de formante individual en serie, siendo cada resonador un filtro de segundo orden.

En cuanto a implementaciones de síntesis por formantes, el sintetizador Klatt (Klatt, 1980) es uno de los desarrollos de sintetizador por formantes más sofisticado, incluyendo tanto un resonador en paralelo como en cascada. Fue configurado para trabajar a 10 KHz utilizando 6 formantes principales. Es interesante apreciar el hecho que en la mayoría de la bibliografía relacionada con la síntesis por formantes se usa una frecuencia de muestreo de 8 KHz ó 10 KHz, debido principalmente a requisitos de espacio, velocidad y salida que impedían altas velocidades. Por tanto en el caso que se necesiten elevados valores de muestreo este número puede ser fácilmente modificado. Con esto, los tres primeros formantes son utilizados por los oyentes para discriminar sonidos, mientras que los formantes más elevados se usan simplemente para dar naturalidad al habla.

Una vez presentado el sistema de síntesis pasemos a hablar de la calidad que se consigue con este tipo de sistemas. La evaluación general de la síntesis por formantes es que es inteligible, también llamado “sonar limpiamente”, aunque está lejos de ser natural. Esto se debe principalmente a que tanto el modelo de fuente como el de destino y el de transición son demasiado simplistas, dejando de lado muchas de las sutilezas que realmente están implicadas en la dinámica del habla. Mientras las formas de las trayectorias de los formantes se miden con un espectrograma, el proceso subyacente es el control motor y movimiento muscular de los articuladores. Aún cuando cada articulador puede moverse de una manera bastante simple, cuando se combinan al sistema global este es altamente complejo, complicando además el efecto del tracto vocal sobre el paso de la forma de onda fuente. Finalmente, las asunciones hechas sobre la naturaleza del modelo de tracto vocal, con la consiguiente falta de precisión, se van sumando y acaban por afectar al modelo global. A pesar de los efectos adversos por las asunciones realizadas, estas pueden ser evitadas mediante la manipulación de los valores fuera de su interpretación natural, es decir, mientras que una manipulación apropiada del sintetizador por formantes puede producir habla muy natural, esto afecta en un coste de tener que usar los parámetros de forma poco habitual, complicando su interpretación. Por tanto, existe un conflicto real entre tener un modelo fácilmente controlable y otro que produce un habla sintética de alta calidad.

Dicho esto y para terminar, la mayor crítica hacia la síntesis por formantes es que esta no suena natural. Aun así, por producir habla inteligible la hace aún hoy competitiva según sea la aplicación en la que vaya a ser utilizada.

3.3.2.2. Síntesis articulatoria

Posiblemente la forma más obvia de sintetizar habla es la de intentar simular directamente la producción del habla humana. Es a este enfoque al que se ha llamado síntesis articulatoria, siendo el más antiguo de los planteamientos, ya que la conocida "máquina parlante" de von Kempelen (1791) (Figura 3.6), descrita por Dudley y Tarnoczy (1950), puede ser vista como un sintetizador articulatorio (Taylor, 2009b).

Figura 3.6: *Reconstrucción de la máquina parlante de von Kempelen (1791) (Universität des Saarlandes)*

Esta máquina era un dispositivo mecánico con tubos, fuelles y conductos; los cuales con un pequeño entrenamiento podrían ser utilizados para producir habla reconocible. Este dispositivo solamente imita el tracto vocal usando fuentes y filtros, con la ventaja de que al ser controlado por un ser humano en tiempo real, los ajustes llevados a cabo se aprovechan de los mecanismos usados por él para el control de la producción de la voz.

En la actualidad, la síntesis articulatoria se aborda desde una perspectiva distinta como es lógico, ya que no tiene sentido que nadie sea el que esté controlando el dispositivo mecánico. Muchos sintetizadores modernos son extensiones de los modelos acústicos de tubos, pudiéndose construir modelos generales complejos a partir del conocimiento de las propiedades de propagación del sonido.

Existen principalmente dos dificultades en la síntesis articulatoria. En primer lugar, la decisión de cómo generar el control de parámetros desde las especificaciones y, en segundo lugar, hallando el punto medio entre un modelo altamente preciso que se ajuste fielmente a la fisiología humana y un modelo, más pragmático, más sencillo de diseñar y de controlar. El primer problema es similar a lo que ocurre en síntesis por formantes. Sin embargo en aquella, en muchos casos aunque no en todos, es sencillo encontrar los valores de formante del habla real, ya que simplemente se tiene que grabar el habla, calcular el espectrograma y determinar el valor de los formantes. El problema en síntesis articulatoria es considerablemente más complejo, puesto que no se pueden averiguar los parámetros articulatorios a partir de grabaciones, sino que se deben utilizar medidas más invasivas tales como la fotografía de

rayos-X, Imagen por Resonancia Magnética —*Magnetic Resonance Imaging*— (MRI) o Articulografía Electromagnética —*Electromagnetic Articulagraphy*— (EMA), con el consiguiente problema de la recopilación de la información debido a que muchas de las técnicas existentes son desarrollos recientes, que no existían en los primeros días de la síntesis articulatoria, siendo particularmente complicada su adquisición. El segundo de los problemas concierne a cómo de preciso debería de ser nuestro modelo de articulación. En el modelo de tubos siempre existe un compromiso entre cómo de bien se lleva a cabo la imitación deseada y lo simple y tratable que resultará el modelo. Los modelos más actuales incluyen modelado de pérdidas del tracto vocal, interacción fuente-filtro, radiación de los labios y características glóticas. Además, muchos de los modelos han pretendido ser tanto modelos de tracto vocal como de control, disponiendo de modelos tanto para el movimiento de los músculos como de control motor.

Ambos problemas presentados acarrean una considerable dificultad, ya que la mejor síntesis articulatoria es pobre si la comparamos con la mejor síntesis usando otras técnicas. Debido a esto se ha ido abandonando como técnica de generación de habla de alta calidad en el ámbito de la ingeniería. Sin embargo, aunque este enfoque podría no ser una buena solución de ingeniería en términos de CTH, sigue despertando interés en otras disciplinas relacionadas. Primero de todo, existe un considerable interés en el campo de la producción de habla, donde existe la discusión de que si el dominio articulatorio es el dominio natural y más correcto para la producción del habla y, por tanto, ayuda a explicar la organización sistemática de los niveles más elevados de esta. Por ejemplo la fonología articulatoria (Browman y Goldstein, 1986) está basada en la idea de “gestos articulatorios” como primitivas fonológicas más que características basadas en segmento (Taylor, 2009a). Boersma (1998) se distingue también por desarrollar un teoría de fonología junto con un sintetizador articulatorio. Un segundo campo de interés relacionado es la “fisiología articulatoria” donde la meta es crear modelos completos del movimiento articulador. En este caso, el énfasis viene dado por intentar modelar articuladores específicos o efectos con precisión, más que en construir un modelo aproximado, o enlazar este con un modelo lingüístico/fonético (Wilhelms-Tricarico, 1995; Vatikiotis-Bateson y Yehia, 1997). Finalmente, la síntesis articulatoria está implícitamente conectada con el campo de la síntesis audiovisual o síntesis de cabezas parlantes, donde la idea principal es la de construir un modelo visual completo de la cabeza mientras se habla. Estas cabezas parlantes pueden ser construidas directamente mediante el modelado de los articuladores o bien utilizando datos reales a partir de fotografías o vídeos y técnicas de *morphing* para crear la animación.

3.3.2.3. Síntesis concatenativa

La síntesis concatenativa es la opción que ha tenido más seguidores dentro de la comunidad científica. Parte de unidades de voz pregrabadas que debe de unir entre sí para generar el mensaje oral. Durante años, dentro de la síntesis concatenativa, la basada en concatenación de difonemas y trifonemas fue la más utilizada en el desarrollo de sistemas de CTH. El problema fundamental que presentan estos sistemas es

que se basan en un corpus donde cada unidad solamente dispone de una realización, es decir que cada unidad se grabó solamente una vez. Normalmente la grabación se realiza utilizando frases portadoras de las que se selecciona la unidad de interés o mediante palabras vacías (Black y Lenzo, 2001). Los problemas fundamentales que presenta esta tecnología son:

1. **Modificación prosódica**. Cuando la prosodia indicada por el bloque PLN difiere en exceso de la prosodia almacenada en el corpus. Por lo tanto, los cambios prosódicos que deben padecer las unidades provocan un descenso de la naturalidad de la señal sintética resultante y, por tanto de la calidad. Éste es el motivo por el que es necesario disponer de una gran variedad de las mismas unidades grabadas, variando los diferentes contextos donde ésta se encuentre, como por ejemplo prosódicos, lingüísticos o fonéticos.

2. **Concatenación de unidades**. Durante el proceso de síntesis llevado a cabo por el PDS se deben unir las unidades que han sido modificadas prosódicamente. Debido a que solamente se dispone de una realización por unidad y diferentes contextos donde se pueden encontrar, es decir fonemas que tengan delante y detrás (contexto izquierdo y derecho), las uniones entre unidades serán tantas como el número de unidades a sintetizar menos una, por tanto habrá un elevado número de concatenaciones sin llegar nunca a ser tan naturales como el proceso fisiológico que las genera, a pesar de la bondad del algoritmo de concatenación utilizado, provocando discontinuidades espectrales. Utilizando más realizaciones de las unidades es probable que se encuentren contextos iguales al deseado de modo que el número de concatenaciones se vea disminuido y así la calidad incrementada.

A pesar de las mejoras introducidas en el proceso de concatenación a partir de la reducción de las discontinuidades espectrales en el punto de concatenación, mediante un mejor diseño del corpus, la señal sintética de un sistema de CTH basado en difonemas tiene falta de naturalidad debido al elevado número de puntos de concatenación (Möbius, 2000). Por este motivo, cuando la tecnología permitió la creación y gestión de bases de datos mayores, es decir corpus mayores, se cambió de estrategia pasando a trabajar con corpus más grandes que consideraban unidades de duración variable. Los primeros trabajos en síntesis basada en unidades de longitud variable fueron desarrollados por el grupo de *Advanced Telecommunications Research Institute International* (ATR) (Sagisaka, 1988; Takeda et al., 1990; Sagisaka et al., 1992), mientras que en paralelo, en la misma institución, se desarrolló el trabajo que daría nombre a la nueva estrategia de síntesis: selección de unidades (Black y Campbell, 1995; Hunt y Black, 1996).

Llegados a este punto, pasemos a ver las características generales de la síntesis basada en selección de unidades:

1. Se dispone de un **corpus** de voz con un elevado número de repeticiones, para cada una de las unidades consideradas (p. ej. difonemas y trifonemas), obteniendo diversidad.

2. Se **selecciona** la secuencia de unidades del corpus más larga posible que se ajuste a las características prosódicas de la secuencia de unidades a sintetizar, información obtenida por el PLN en tiempo de ejecución. Otras alternativas a la secuencia más larga puede ser una ponderación de pesos de forma que la que genere una mejor síntesis sea la elegida.

3. Se **minimiza** el número de puntos de concatenación y la necesidad de modificación prosódica de la señal, aumentándose de este modo la naturalidad de la señal generada.

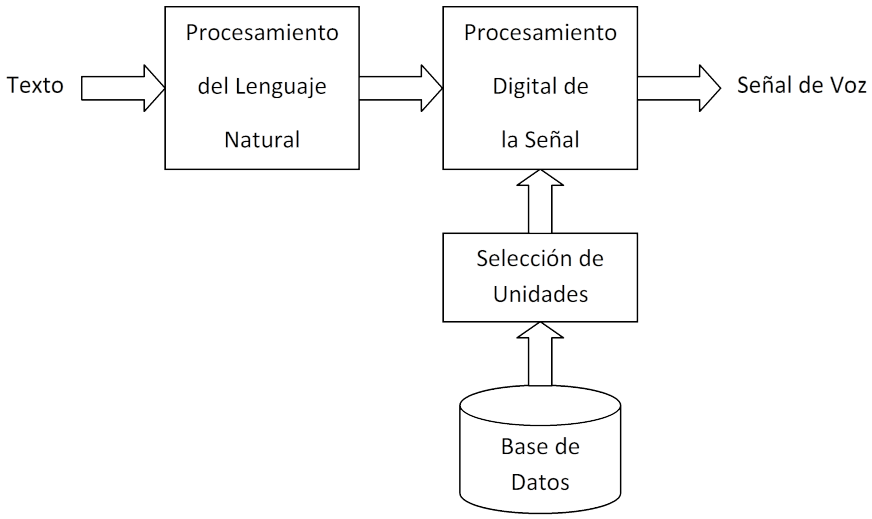


Figura 3.7: *Diagrama de bloques de un sistema de conversión de texto en habla basado en selección de unidades*

Resumiendo, en la arquitectura del sistema de CTH basado en selección de unidades, por un lado se incorpora una base de datos (corpus) más importante que la de un sistema CTH basado en difonemas, coincidiendo el esquema con el mostrado en la Figura 3.4 y, por otra parte, se incorpora un módulo encargado de seleccionar la cadena óptima de unidades en tiempo de ejecución (Figura 3.7).

3.3.2.4. Síntesis basada en Modelo Oculto de Markov (HMM)

Presentada la estrategia de síntesis más extendida, la de síntesis concatenativa basada en selección de unidades (Sección 3.3.2.3), se presenta una alternativa que está cobrando importancia en los últimos años: la síntesis basada en HMM (Tokuda et al., 1995).

Los principales intereses que existen respecto a los sistemas de CTH son por un lado aumentar la calidad y, ligado con esto, la naturalidad de la señal de voz generada en aplicaciones de propósito general. La síntesis concatenativa, especialmente en diseños de dominio restringido como el presentado por Alías et al. (2005), presenta inconvenientes cuando se intenta utilizar fuera del dominio para el que fue diseñada. Nuevas grabaciones tienen la desventaja de ser costosas, tanto en tiempo como en

dinero, ya que se requiere el diseño de nuevos textos, la grabación, el etiquetado, entre otras.

Por el contrario, *Hidden Markov Model based Text-to-Speech* (HMM-TTS) tiene como principal característica la capacidad de modelar voces para así sintetizar diferentes características del interlocutor, estilos de locución y expresividades (o emociones). Además, en el caso de aplicar transformación de voz a través de síntesis concatenativa, todavía implica grandes corpus en comparación a la basada en HMM, la cual obtiene mejores resultados con corpus menores (Yoshimura et al., 1999). Por otro lado, el uso de HMM para síntesis del habla podría ser usado en nuevos sistemas juntamente con selección de unidades, permitiendo de este modo que se unificaran ambas estrategias y se aprovecharan las ventajas de sus características particulares (Taylor, 2006), disponiendo así de un sistema de síntesis híbrido.

Un sistema típico de síntesis basada en HMM tiene tres estados por fonema, utiliza coeficientes *Mel Frequency Cepstral Coefficients* (MFCC) juntamente con los delta y aceleración (primera y segunda derivada de los coeficientes respectivamente), y utiliza modelos basados en contexto con estados enlazados determinados por agrupación de un árbol de decisión. La noción de contexto HMM se generaliza en una descripción de características, así se dispone de un modelo para cada una de estas descripciones. Esto puede resultar en algunos millones de modelos potenciales, de los cuales solamente unos pocos miles habrán sido observados en los datos de entrenamiento.

La síntesis se lleva a cabo por la generación de un HMM a nivel de frase usando los modelos HMM individuales que encajan con la especificación. Si una combinación de características no fuera observada durante el entrenamiento, el siguiente mejor modelo se selecciona por medio del árbol de decisión. Usando HMM a nivel de frase, se generan las secuencias de observaciones más probables. La clave en la síntesis basada en HMM es realizar observaciones que obedecen a las dinámicas de estado por los coeficientes delta y aceleración. En primer lugar, esto asegura que dentro de un modelo la trayectoria de un coeficiente concreto está casi siempre evolucionando, con lo que no se ven saltos de coeficientes en los límites de los estados. En segundo lugar, estas mismas restricciones dinámicas son aplicadas al estado de transición entre modelos de fonema, de ese modo se asegura la suavidad en las transiciones fonema a fonema. Esto es particularmente hábil y evita la idea de coste de unión en síntesis basada en HMM.

La lengua es otro tema importante cuando se diseña un sistema de CTH. El esquema de HMM-TTS basado en factores contextuales por agrupación (*clustering*) puede ser usado para cualquier lengua, como muestra Tokuda et al. (2002) para el inglés o da S. Maia et al. (2003) para el portugués. Los fonemas, unidades básicas de síntesis, y sus pares contextuales atributo-valor, como p. ej. el número de sílabas en la palabra o acentuación, son la principal información que cambia de una lengua a otra.

3.3.2.5. Síntesis basada en Modelo Armónico más Ruido (HNM)

A continuación se muestra el modelo Modelo Armónico más Ruido —*Harmonic plus Noise Model*— (HNM) (Laroche et al., 1993; Stylianou et al., 1995; Stylianou, 1996) y su desarrollo específico para sistemas de CTH (Stylianou, 1998). Como el propio nombre sugiere, el modelo está formado por una componente determinista o armónica ($s(n)$) y otra estocástica o de ruido ($r(n)$), que sumadas dan idealmente como resultado la señal de voz ($x(n)$) (Ecuación 3.1). Esta componente estocástica o de ruido es más sofisticada que en algunos modelos en los que se parte de que el habla real tiene patrones temporales específicos.

$$x(n) = s(n) + r(n) \tag{3.1}$$

Tal y como se presenta en el trabajo de Erro (2008),existen diferentes propuestas para el análisis y síntesis usando HNM, según si se utilizan técnicas *pitch* sincrónicas o asincrónicas, las formas de minimización del error y la reconstrucción de la señal de voz a partir de los parámetros extraídos durante el proceso de análisis (p. ej. interpolación lineal o cúbica).

En esta sección se introduce la metodología de síntesis basada en HNM presentada por Stylianou (1996). En las primeras etapas del análisis se realiza la clasificación de zonas sonoras y sordas, que establecen los parámetros de las componentes determinista y estocástica, así como las aportaciones de cada una de las componentes a la trama analizada. Primero se estima el *pitch* y a partir de aquí se lleva a cabo un análisis *pitch* sincrónico. No tiene por qué ser realizado con el objetivo de hallar el instante de cierre glótico, ya que una simple localización de los periodos de *pitch* es suficiente. Le sigue el modelado armónico (Ecuación 3.3) utilizando la frecuencia fundamental (F_0) estimada a partir del marcado de *pitch* para cada una de las tramas. El error existente entre la señal de voz original ($x(n)$) y la componente determinista ($s(n)$) será menor en aquellas tramas con una componente armónica elevada, mientras que en las más ruidosas el error será más elevado. En aquellas zonas consideradas sonoras se determina la frecuencia del armónico mayor que será el límite entre la parte armónica y la de ruido.

$$F_{ik} = F_{0k} \cdot i \tag{3.2}$$

$$s_k(n) = \sum_{i=1}^{I_k} A_{ik} \cdot \cos(2\pi \cdot F_{ik} \cdot n + \phi_{ik}) \tag{3.3}$$

Donde:

- A_{ik}: amplitud del armónico i en la trama $k-$ésima.
- F_{ik}: frecuencia del armónico i en la trama $k-$ésima.
- ϕ_{ik}: fase del armónico i en la trama $k-$ésima.
- k: índice de trama correspondiente a un periodo de análisis del HNM.

- i: índice del armónico de interés.
- I_k: número máximo de armónicos i en la trama $k-$ésima.
- n: índice de muestra de la señal temporal.

Una vez esta frecuencia de corte ha sido hallada, se procede a afinar la estimación de *pitch* usando solamente la parte más baja de la señal. Las amplitudes y las fases son encontradas minimizando el error cometido entre las formas de onda real y sintética correspondiente a la componente armónica.

Se requiere de un paso final para asegurar que no ocurren desajustes de fase entre tramas, ya que debido a que se llevó a cabo un análisis *pitch* sincrónico sin referencia del instante de cierre glótico, no estarán necesariamente alineados. Se usa una técnica en el dominio temporal para ajustar las posiciones relativas de las formas de onda dentro de las tramas para asegurar su alineamiento.

Durante el proceso de síntesis, se genera un conjunto de marcas de síntesis y una función de mapeo proveniente de las marcas de análisis, calculada tal y como se haría para *Pitch Synchronous OverLap and Add* (PSOLA) (Moulines y Charpentier, 1990), realizando la modificación temporal del mismo modo que si de PSOLA se tratara. Para la modificación del *pitch*, los armónicos de las tramas se ajustan mediante un remuestreo del espectro de la trama, aplicando estos valores a cada componente sinusoidal que generará la componente determinista. La componente estocástica se crea haciendo pasar un ruido Gaussiano de potencia constante ($b(n)$) a través del filtro $h(n, \tau)$ y multiplicando su salida por una función que le dará el correcto patrón temporal (Ecuación 3.4). Para tramas sonoras, el ruido es filtrado con un filtro pasaaltas con la frecuencia de corte situada en la componente armónica, evitando así la generación de ruido de baja frecuencia. El ruido queda modulado en el dominio temporal mediante la función $e(n)$ para conseguir emular su dinámica energética. Esta etapa es esencial para garantizar la percepción de un único sonido en lugar de dos separados. Finalmente, se realiza la síntesis mediante solapamiento y suma (en inglés *overlap and add*), con PSOLA.

$$r(n) = e(n) \cdot [h(n, \tau) * b(n)] \tag{3.4}$$

3.3.2.6. Síntesis híbrida

La síntesis híbrida es aquella en la que se combinan aspectos de la síntesis paramétrica y de la concatenativa en general (como se ejemplifica en la Figura 3.8). La síntesis por formantes era la más usada, aunque con el auge de los sistemas de síntesis basada en HMM o HNM éstos también han pasado a combinarse con la concatenativa. Con esta unión de metodologías se busca minimizar los defectos acústicos debidos a la concatenación de segmentos.

Según el trabajo de Aylett y Yamagishi (2008), uno de los puntos débiles de la síntesis concatenativa es la escasez de ejemplos para llevar a cabo la síntesis. Por tanto, para poder generar el habla de forma correcta se debe disponer de las unidades apropiadas, siendo especialmente costoso si pensamos en que el hecho de que se

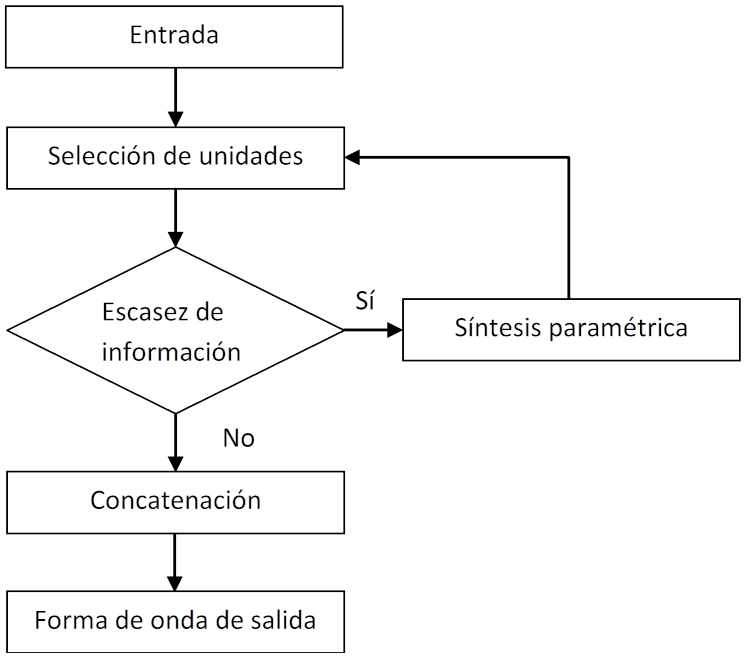


Figura 3.8: *Combinación de síntesis paramétrica y concatenativa por selección de unidades*

necesitarán los difonemas/trifonemas, en función de la lista de fonemas de la lengua de interés, para dar la máxima cobertura; y además si es necesario dar cobertura prosódica, el número de unidades distintas puede dispararse. El contexto de la mayoría de las unidades es vital para la concatenación debido a la coarticulación, con lo cual, la cosa se complica aún más si pensamos en todos los contextos por la izquierda y derecha necesarios, a pesar de que afortunadamente no todos estos contextos serán tenidos en cuenta por no darse en la lengua de interés. Es por esta razón, que en una base de datos para síntesis, puede ocurrir que haya casos no contemplados, produciendo así errores de concatenación que afectarán a la calidad de la síntesis.

Respecto al enfoque paramétrico, por un lado ofrece una solución atractiva al problema de escasez de datos. En primer lugar, como la voz se sintetiza con parámetros a partir de un modelo no existen errores de concatenación. En segundo lugar, debido a que la voz deriva de un modelo es posible usar técnicas de adaptación para aprovechar información de otros interlocutores, y mejorar así el modelo sobre una pequeña cantidad de datos. Por otro lado, las desventajas de este enfoque son que la generación del habla desde el modelo no reproduce completamente la naturalidad de la misma, pudiéndose detectar en ocasiones un zumbido debido al vocoder empleado. Sin embargo, hay un problema mayor ya que es necesario modelar todas las características del habla, incluyendo variaciones prosódicas, de VoQ y de estructura.

Un enfoque híbrido procura usar las ventajas de ambos sistemas para maximizar la calidad y la naturalidad del habla final. Por ejemplo se puede extraer la máxima cantidad de información prosódica utilizando el módulo de selección de unidades, mientras que se evita la falta de datos mediante el módulo paramétrico.

3.3.3. Conversión de texto en habla en el grupo de investigación

El grupo de investigación GTM ha sido uno de los grupos pioneros en el campo de la síntesis del habla desde la década de los años ochenta con los trabajos del Dr. Josep Martí (Martí, 1985, 1987, 1990). Se llevaron a cabo trabajos de investigación y desarrollo como síntesis articulatoria, síntesis por formantes y síntesis basada en predicción lineal (*Linear Predictive Coding* (LPC)). Estos sistemas se aplicaron principalmente a productos orientados a personas invidentes (colaborando con el grupo ONCE), llegándose a obtener equipos de síntesis del habla capaces de leer la información contenida en la pantalla o desde el teclado de un ordenador.

Hasta ese momento, la calidad de los sistemas de síntesis del habla solamente estaba condicionada por su grado de inteligibilidad, factor muy apreciado por un sistema de CTH orientado a personas discapacitadas. El siguiente reto fue incrementar la naturalidad de la señal sintética, motivo por el cual se centraron los esfuerzos del grupo en las siguientes mejoras:

- Análisis lingüístico del texto.
- Modelado y automatización de la prosodia.
- Procesamiento digital de la señal para síntesis del habla.

Posteriormente, con el objetivo de mejorar el bloque de PDS, se optó por la síntesis concatenativa basada en difonemas y trifonemas. Durante este periodo se implementó un sintetizador en catalán (Camps et al., 1992; Guaus et al., 1996) basado en la técnica PSOLA (Moulines y Charpentier, 1990) que tuvo una gran repercusión en la comunidad científica. Este sistema de CTH en catalán se convirtió en la base de los posteriores sistemas de síntesis desarrollados por el grupo.

Ya trabajando en síntesis concatenativa, durante el año 1997 se inició el desarrollo de un nuevo sistema de CTH en catalán con el objetivo de obtener alta inteligibilidad. Se fijó como objetivo mejorar los diferentes módulos que componían el sistema de CTH existente. En primer lugar se mejoró el preprocesamiento del texto de entrada, desarrollando a continuación un lenguaje de reglas compiladas para llevar a cabo la transcripción fonética (conversión grafema-fonema) y determinar automáticamente la prosodia de las unidades. En cuanto al módulo de síntesis (PDS) se realizaron mejoras en el proceso de concatenación de las unidades, y se grabó, etiquetó y segmentó un nuevo corpus para catalán cubriendo 1207 unidades, 895 de ellas difonemas y 312 trifonemas, con una realización por unidad. Se obtuvo así un sistema de CTH con buena inteligibilidad y calidad aceptable.

En paralelo a la mejora del sistema de CTH en catalán se comenzó a trabajar en el contexto de la síntesis del habla expresiva, es decir, la transmisión del estado emocional del interlocutor a partir de voz sintetizada. A pesar de que la mayoría de los sistemas actuales de síntesis se caracterizan por una gran inteligibilidad y una buena naturalidad, no existen resultados concluyentes por lo que hace referencia al habla expresiva. A partir de la colaboración con el Departamento de Comunicación Audiovisual y Publicidad (CAP) de la *Universitat Autònoma de Barcelona* (UAB) se

inició una línea de investigación en el campo del modelado acústico de la expresión emocional (Rodríguez et al., 1999) utilizado con un sistema de CTH (Iriondo et al., 2000). De este trabajo se concluyó que ciertas emociones requieren de un módulo de PDS que permita grandes variaciones prosódicas.

El uso de *Time-Domain Pitch Synchronous OverLap and Add* (TD-PSOLA) durante la síntesis no tenía suficiente versatilidad para obtener unos resultados suficientemente satisfactorios. Por esta razón se estudiaron alternativas para la concatenación de voz aparecidas en los últimos años: desde *Multi-Band Resynthesis Overlap-Add* (MBROLA) (Dutoit y Leich, 1996; Dutoit y Stylianou, 1997) hasta el modelo HNM de Stylianou (1996). Para no introducir un gran cambio en el sistema de CTH, aunque permitiendo las nuevas ideas aparecidas en el campo de la concatenación de tramas de voz, se planteó el diseño de un modelo híbrido entre PSOLA y HNM que permitiera variaciones prosódicas de mejor calidad (Iriondo et al., 2002) de forma que se incorporara al sistema de CTH del grupo. Para profundizar en la evolución de la síntesis del habla dentro del grupo de investigación de Ingeniería y Arquitectura La Salle se recomienda el trabajo realizado por Alías y Iriondo (2002).

Para terminar, la línea de trabajo ha evolucionado hacia la mejora de la síntesis, tanto en inteligibilidad como naturalidad, expresividad y modularidad del sistema. Se ha trabajado, dentro del bloque de PLN, en la transcripción automática de acrónimos (Apéndice E), y se está trabajando en la introducción de nuevas técnicas de síntesis como HMM y HNM, y estudiando nuevas parametrizaciones basadas no sólo en modelar la prosodia, sino también la VoQ con el objetivo final de aumentar la calidad, naturalidad y expresividad de la voz generada.

3.4. Síntesis del habla expresiva

La Síntesis del Habla Expresiva (SHE) o *Expressive Speech Synthesys* (ESS) en inglés, es un área de investigación multidisciplinar dirigida a uno de los problemas más complejos existentes en el procesamiento del habla y del lenguaje. El reto planteado por la SHE ha sido el tema de varios proyectos de investigación colaborativos entre universidades y laboratorios alrededor del mundo. La SHE se ha beneficiado de los avances que el procesamiento del habla y del lenguaje ha tenido en los últimos años, así como también de la disponibilidad de grandes bases de datos de habla conversacional. Estos avances han estimulado la investigación de la expresividad en el habla y de la información paralingüística que se transmite, incluyendo la emoción, el estado del hablante y las relaciones entre interlocutores. También han sido sustanciales los esfuerzos realizados hacia la automatización de la creación de bases de datos y la evaluación de la calidad del habla sintética para una variedad de tareas, que requieren no solamente de la transmisión de información sino también de la expresividad.

3.4.1. Generación de habla expresiva

La SHE cabría englobarla dentro del marco más general de la síntesis de voz con estilos de habla no neutros. Actualmente hay quien utiliza el término síntesis del habla "expresiva" para significar que el habla humana es muy rica en matices, y que las emociones solamente son una parte más de la gran variabilidad del habla humana (Campbell, 2003). Históricamente el estudio de las emociones se ha realizado desde ámbitos muy diversos, como la psicología o la fonética acústica, y no ha sido hasta hace poco que también se ha abordado desde el punto de vista de las tecnologías del habla.

La naturaleza compleja del habla es debida a que esta varía dependiendo del estilo de habla y emoción del interlocutor, dificultando su imitación por medio de la síntesis del habla. Para poder comparar habla sintética con habla real se utilizan tres características (Bulut et al., 2002):

- Inteligibilidad: es una medida del grado de entendimiento que tiene el habla para una persona.
- Naturalidad: se trata de la característica que muestra cómo de humana suena el habla generada por un sistema de CTH.
- Variabilidad: representa los cambios en la velocidad del habla, la VoQ y la naturalidad; siendo la falta de variabilidad la que ha impedido la aceptación general del habla sintética.

Por ello, debido al desconocimiento de cómo afectan a la señal de habla factores como el tipo de material leído, el comportamiento de la audiencia y la posición social del interlocutor, su actitud y emociones expresadas; han sido una de las mayores problemáticas a la hora de crear sistemas que suenen "humanos".

La necesidad de "humanizar" sistemas de CTH proviene del hecho que aplicaciones basadas en la interacción entre hombre y máquina pueden verse enormemente mejoradas y hacerlas más simples y convincentes. Desde el inicio del desarrollo de sistemas de CTH ha habido intentos de dotarlos de la capacidad para generar voces con emoción (Murray y Arnott, 1993; Schröder, 2001). Los primeros intentos de sintetizar habla expresiva, o emocionada en general, se realizaron sobre sistemas basados en reglas y, en particular, con sistemas de síntesis por formantes (Cahn, 1989; Murray y Arnott, 1995).

Actualmente, la mayoría de sistemas de CTH utilizan la concatenación de segmentos de voz extraídos de grandes bases de datos, seleccionando el segmento más apropiado en cada ocasión (Esquerra, 2006). Esto hace que el habla generada mantenga las características propias del locutor y en cierta medida del estilo en que fueron grabadas, por tanto, para poder generar otros tipos de habla es necesario disponer de corpus más extensos que recojan las variantes de habla deseadas (Black, 2003).

Schröder (2004) realizó una amplia investigación en el tema de las emociones y el habla, presentando el caso de la SHE. El modelado de la emoción en el habla depende principalmente de parámetros como la frecuencia fundamental (F_0), la VoQ o la

Parámetro		Autor
Prosodia	Reglas extraídas de la bibliografía	Cahn (1990) Murray y Arnott (1995) Rank y Pirker (1998) Murray et al. (2000) Stallo (2000)
	Reglas obtenidas del análisis de sus corpus	Montero et al. (1998) Mozziconacci y Hermes (1999) Iriondo et al. (2000) Campbell y Marumoto (2000) Boula de Mareüil et al. (2002)
	Reglas resultantes de valores perceptivos óptimos en la síntesis	Mozziconacci (1998) Burkhardt y Sendlmeier (2000)
VoQ		Cahn (1990) Murray y Arnott (1995) Rank y Pirker (1998) Iriondo et al. (2000) Murray et al. (2000) Campbell y Marumoto (2000) Burkhardt y Sendlmeier (2000)
Precisión articulatoria		Cahn (1990) Murray y Arnott (1995) Rank y Pirker (1998) Burkhardt y Sendlmeier (2000)
Categorías lingüísticas	Distinción entre el tempo de vocales y consonantes	Murray y Arnott (1995) Rank y Pirker (1998) Stallo (2000)
	Distinción entre el tempo de sílabas acentuadas y no acentuadas	Murray y Arnott (1995) Burkhardt y Sendlmeier (2000) Stallo (2000)
	Ubicación de las pausas	Cahn (1990)
	Prosódicas, como por ejemplo contornos de F_0	Mozziconacci y Hermes (1999) Burkhardt y Sendlmeier (2000)

Tabla 3.5: *Resumen del estudio realizado por Schröder (2004) sobre la propuesta de parámetros utilizados en síntesis del habla expresiva*

Parámetro	Autor
Prosodia	Garrido (1991) Murray y Arnott (1993) Banse y Scherer (1996) Gobl et al. (2002) Drioli et al. (2003) Bänziger y Scherer (2003) Yamagishi et al. (2004) Tesser et al. (2005) Yanushevskaya et al. (2005) Iriondo et al. (2007c) Iriondo (2008)
VoQ	Murray y Arnott (1993) Banse y Scherer (1996) Gobl et al. (2002) Gobl y Ní Chasaide (2003) Campbell y Mokhtari (2003) Drioli et al. (2003) Bänziger y Scherer (2003) Yamagishi et al. (2004) Tesser et al. (2005) Yanushevskaya et al. (2005) Iriondo (2008)

Tabla 3.6: *Relación entre trabajos y la utilización de prosodia y/o cualidad de la voz para la síntesis del habla expresiva*

precisión articulatoria; mientras que diferentes técnicas de síntesis proveen control sobre estos parámetros: síntesis por formantes, síntesis concatenativa por difonemas o síntesis concatenativa por selección de unidades. A pesar de estas metodologías, en los últimos años se han desarrollado trabajos donde se han empleado otras técnicas de síntesis, que permiten la modificación de parámetros pudiendo ser aplicados en SHE, como son la síntesis basada en HMM (Yamagishi et al., 2004) o HNM (haciendo uso únicamente de la parte armónica (Drioli et al., 2003) o bien mostrando las posibilidades que tiene la técnica en otras aplicaciones de transformación de habla (Stylianou et al., 1995)). En relación con los parámetros y su modificación usando las diferentes técnicas propuestas, en la Tabla 3.5 se presenta el resumen del estudio realizado por Schröder (2004), donde se relacionan los parámetros con los autores que los han utilizado en su trabajo. Complementariamente, en la Tabla 3.6 se presentan otros estudios en los que se han aplicado tanto parámetros prosódicos como de VoQ con el objetivo de desarrollar sistemas de CTH para SHE.

3.4.2. Evaluación de la síntesis del habla expresiva

La SHE puede ser evaluada bajo diferentes criterios, es decir, existen una serie de metodologías que han sido usadas en un elevado número de estudios, a pesar de que no hay una opinión unánime sobre cuál es el idóneo.

La forma más habitual de evaluar la calidad resultante de la SHE es a través de una prueba perceptiva, en la que se fuerza al oyente a dar una respuesta al estímulo presentado entre las categorías modeladas, correspondiendo estos estímulos a un reducido número de frases portadoras con contenido semántico neutro. Las ventajas de este tipo de prueba son que es relativamente sencillo tener una medida simple del reconocimiento relativo conseguido, y así poder establecer límites para poder llevar a cabo comparaciones entre estudios. Sin embargo, la calidad del estímulo presentado al oyente se desconoce en términos de naturalidad y verosimilitud, siendo esta la razón por la cual algunos estudios complementan la evaluación con medidas de naturalidad, verosimilitud o de preferencia global de la expresividad mostrada (por ejemplo mediante escalas de cinco puntos). Además, otras informaciones adicionales que pueden ser evaluadas son la intensidad de la expresividad o el nivel de inteligibilidad de la síntesis.

Una segunda posibilidad, adecuada para hallar fenómenos no esperados, son las pruebas con respuesta libre. Una vez se dispone de las respuestas, estas se agrupan según las clases más representativas usando listas validadas de palabras (Murray y Arnott, 1995).

Por último, una alternativa que puede ser interesante para llevar a cabo la evaluación es la siguiente. Primero se añaden un número de categorías que puedan “distraer” al evaluador, como por ejemplo la categoría de “otros”. Además, los textos utilizados tienen contenido semántico, tanto para el caso de expresividad neutra como para otras y, por ejemplo, en el supuesto de estar utilizando reglas prosódicas, estas se usan tanto con la síntesis neutra como la expresiva . La diferencia de reconocimiento entre la versión de prosodia para la síntesis neutra y expresiva se tomaría como la medida para medir el impacto perceptivo de las reglas prosódicas. Con este procedimiento, se evaluó en un contexto audiovisual la preferencia de los evaluadores ante el estímulo de una cabeza parlante con expresividades (Stallo, 2000), donde el habla era sintética neutra y expresiva. Se pidió puntuar la naturalidad e inteligibilidad entre otros, mostrando que el habla expresiva tenía una preferencia clara.

A partir de las diferentes modalidades de evaluación presentadas, en la Tabla 3.7 se resume el estudio que realiza Schröder (2004) relacionando las diferentes metodologías con trabajos que las utilizan.

Método de evaluación		Autor
Prueba perceptiva con respuesta forzada	Categoría	Cahn (1990) Vroomen et al. (1993) Heuft et al. (1996) Edgington (1997) Rank y Pirker (1998) Montero et al. (1998) Mozziconacci y Hermes (1999) Montero et al. (1999) Scherer (1999) Burkhardt y Sendlmeier (2000) Iida et al. (2000) Campbell y Marumoto (2000)
Prueba perceptiva con respuesta libre	+ naturalidad, verosimilitud o preferencia global	Cahn (1990) Rank y Pirker (1998) Scherer (1999) Iida et al. (2000)
	+ intensidad de la expresividad, o bien + inteligibilidad del habla sintetizada	Cahn (1990) Iida et al. (2000)
Prueba del impacto perceptivo		Murray y Arnott (1995) Scherer (1999) Murray y Arnott (1995) Stallo (2000)

Tabla 3.7: *Resumen del estudio presentado por Schröder (2004) sobre las metodologías de evaluación de la síntesis del habla expresiva*

CAPÍTULO 4

Corpus oral para el análisis y la síntesis del habla expresiva

En este capítulo se presenta el material de voz que se ha empleado en los experimentos de análisis y de síntesis del habla expresiva, pudiendo dividir la información según los siguientes puntos:

- Se hace una introducción a la temática de los corpus orales (Sección 4.1).
- A partir de la visión general sobre los corpus orales se muestran las consideraciones que se deben de tener en cuenta para su desarrollo (Sección 4.2).
- Se presenta el caso concreto del corpus oral expresivo del grupo de investigación GTM (Sección 4.3).
- Por último se describen dos corpus orales expresivos, empleados durante los experimentos destinados a ampliar los resultados obtenidos con el corpus del punto anterior (Sección 4.4).

4.1. Introducción

Existe una tendencia creciente a la utilización del habla en la interacción hombre-máquina. En este campo de trabajo, la inclusión de reconocimiento automático de emociones o la Síntesis del Habla Expresiva (SHE) pueden mejorar la comunicación por otorgarle una mayor naturalidad.

Uno de los mayores retos en el estudio del habla expresiva es el de disponer de corpus orales con un contenido emocional auténtico. La naturalidad de las locuciones depende de la estrategia elegida durante su obtención, centrándose el debate principalmente en la relación entre autenticidad y el grado de control durante la grabación. Campbell (2002) y Schröder (2004) propusieron cuatro fuentes de habla emocional:

- Habla **natural** recopilada a partir de la interacción espontánea. Presenta la mayor naturalidad a pesar de sus problemas debido a la falta de control de su contenido, es decir, insuficiente calidad en la señal de voz y la dificultad en su etiquetado.

- Habla emocional **inducida**. Las emociones se provocan en laboratorio, de forma que se compensan algunos de los problemas detectados en situaciones naturales, aunque el desarrollo de auténticas emociones es raramente conseguido.

- Habla emocional **estimulada**. El estado emocional se crea a partir de la lectura de textos con contenido verbal relacionado con el de la emoción deseada. El problema de comparar frases con diferentes textos se contrarresta incrementando el tamaño del corpus, y así poder usar métodos estadísticos para generar modelos acústicos de la emoción.

- Habla emocional de **actor**. En este caso las mismas frases son leídas con diferentes emociones, permitiendo la comparación directa de parámetros prosódicos y de VoQ. Sin embargo, el mayor obstáculo que presenta esta metodología es la falta de autenticidad en la emoción expresada.

Las bases de datos usadas normalmente en SHE están basadas en habla emocional de actor (Douglas-Cowie et al., 2003), donde un locutor profesional lee un conjunto de textos (con o sin contenido emocional) y simula las emociones que se desea generar.

Otro aspecto importante a tener en cuenta es el propósito de la investigación del habla y emoción. Tal y como señala Schröder (2004), es necesario diferenciar entre los procesos de percepción (centrado en el interlocutor) y aquellos de expresión (centrados en el oyente). El objetivo del primero es el de establecer la relación entre el estado emocional del interlocutor y los parámetros cuantificables del habla, que habitualmente se relacionan con el reconocimiento de emociones a partir de la señal de habla. De acuerdo con Devillers et al. (2005), uno de los retos en el análisis del habla real es la identificación de indicadores orales atribuibles al comportamiento emocional, sin que sean simplemente características propias del habla conversacional espontánea. El segundo de los propósitos, los de expresión centrados en el oyente, están focalizados en determinar los parámetros del habla en un esfuerzo de transmitir un cierto estado emocional, estando de este modo relacionados con la SHE. El tipo de descripción de los estados emocionales utilizado tiene un papel muy importante en los resultados obtenidos. Otro aspecto importante es la elección de los parámetros del habla que se van a modificar durante la simulación de emociones.

Las bases de datos orientadas a la SHE generalmente tienen las siguientes características:

- La **estrategia de grabación** suele consistir en utilizar un actor o un locutor profesional que lea un conjunto de textos con las emociones que se desean simular. Existen dos posibilidades en cuanto a la naturaleza de los textos:

i. Conjunto de textos neutros —sin contenido emocional— que se repiten para cada emoción.

ii. Textos con contenido emocional.

Mientras que el primer tipo de texto facilita la comparación entre estilos, ya que el contenido es el mismo para todas las emociones, el segundo facilita la simulación de la emoción por parte locutor.

- La **duración del corpus** suele ser de varias horas, especialmente en el caso de que la síntesis sea basada en corpus. En este tipo de síntesis se requieren diferentes subcorpus que contengan los diferentes estilos y que bien pueden utilizarse de forma independiente (*tiering*) bien pueden mezclarse (*blending*) permitiendo cambios graduales entre tipos de voz y estilos (Black, 2003).

Como ejemplo de bases de datos de habla emocional de actor existentes en español, orientadas a la SHE, se tienen las bases de datos "*Spanish Emotional Speech database*" (SES) (Montero et al., 1998), "*Interface Emotional Speech Synthesys DataBase*" (IESSDB) (Hozjan et al., 2002) y "Spanish Expressive Voices" (SEV) (Barra-Chicote et al., 2008), donde esta última tiene como objetivos, junto al de servir para SHE, los de ser usada en conversión de habla, reconocimiento de habla en campo lejano e identificación de emociones basada en vídeo. Estas bases de datos fueron validadas mediante pruebas subjetivas, considerando los resultados obtenidos suficientes como para garantizar la expresividad de las grabaciones.

La base de datos SES incluye 4 emociones (enfado, tristeza, alegría y sorpresa) junto con una expresividad neutra. En las pruebas subjetivas casi el $90\,\%$ de las emociones fueron identificadas correctamente, excepto en el caso de alegría, donde el porcentaje descendió hasta el $74\,\%$. Por otro lado, la base de dados IESSDB, fue grabada en francés, inglés, esloveno y castellano, y está formada por 6 emociones (enfado, tristeza, alegría, miedo, asco y sorpresa) más una expresividad neutra. Para la evaluación de la autenticidad del contenido emocional se llevó a cabo una prueba subjetiva, donde los sujetos podían elegir dos emociones para cada una de las frases presentadas. La identificación global fue de aproximadamente el $80\,\%$, si se tenía en cuenta la primera opción seleccionada por los participantes en la prueba, y de un $90\,\%$ si además se consideraba como correcta la segunda de ellas. Finalmente, la base de datos SEV incluye 6 emociones (enfado, tristeza, alegría, miedo, asco y sorpresa) más una expresividad neutra. De su evaluación se obtuvo una validación global media del $87\,\%$ ($89,6\,\%$ para la voz femenina y $84,3\,\%$ para la masculina).

4.2. Desarrollo de corpus orales

En el momento de crear un corpus oral, la primera etapa es la definición de las tareas asociadas y el diseño de elementos clave que determinarán la calidad final. El diseño de un corpus depende de los objetivos que se persigan y de las limitaciones que puedan permitirse. En este apartado se explican, en primer lugar, los objetivos generales relativos a la creación de un corpus oral orientado a la SHE y, en segundo lugar,

se concretan los pasos seguidos en el diseño del corpus del grupo de investigación GTM, utilizado en el presente trabajo.

4.2.1. Objetivos generales

Existen cuatro objetivos teóricos que deberían ser cubiertos por el corpus oral si lo que se pretende es que sea de utilidad en aplicaciones de SHE.

Naturalidad y calidad del habla. Asegurar la naturalidad, entendiéndose esta como la característica de transmitir el estado emocional del interlocutor, es la condición fundamental de un corpus de habla expresiva. A pesar de que las locuciones espontáneas son las consideradas como el tipo más natural de habla (Campbell, 2000), presentan dos importantes inconvenientes cuando son utilizadas en síntesis del habla:

1. Su contenido no puede ser predefinido.
2. Las condiciones ambientales de grabación pueden afectar a la calidad de la señal de voz.

Así pues, la calidad de la grabación es un objetivo prioritario, aunque este objetivo no tiene que ser un impedimento para que las grabaciones realizadas posean un contenido emocional suficiente como para conseguir una SHE de alta calidad.

Cobertura expresiva. Un corpus oral expresivo debería representar un amplio margen de emociones, actitudes y estados de ánimo, de uno o varios interlocutores. Este objetivo puede ser conseguido mediante la creación de un corpus de grandes dimensiones (sobre las 1000 horas) a partir del registro de situaciones cotidianas (Campbell, 2002), consiguiendo de este modo una síntesis próxima al habla natural (Campbell, 2005). Sin embargo, este tipo de grabaciones pueden necesitar de varios años para obtener un número suficiente de locuciones, sin tener en cuenta todo el proceso posterior de etiquetado. Sin embargo, un corpus de tamaño menor con la cobertura apropiada podría ser suficiente para aplicaciones de síntesis, reduciendo de esta manera tanto el coste económico como el tiempo de realización.

Cobertura fonética segmental y suprasegmental. Los sistemas de síntesis basados en corpus requieren de bases de datos de habla con un elevado número de unidades fonéticas y variaciones de características lingüísticas, para poder ser así reproducidas durante el proceso de síntesis (François y Boëffard, 2002). Es importante controlar el tamaño de la base de datos resultante, sin olvidar que se debe conseguir una cobertura adecuada tanto de unidades fonéticas (a nivel segmental) como de variabilidad prosódica (a nivel suprasegmental). La utilización de semifonemas simplifica el problema de la cobertura ya que su número es muy inferior al de difonemas o trifonemas. Sin embargo, su uso no es adecuado para una síntesis del habla de

alta calidad. En el caso de los trifonemas, la cobertura total es prácticamente imposible de conseguir debido al elevado número de combinaciones existentes (Bozkurt et al., 2003). Por esta razón es esencial cubrir los difonemas más habituales y algunos de los trifonemas, considerando la distribución de frecuencia fonética, es decir la frecuencia de aparición de los distintos fonemas para cada lengua específica. La cobertura prosódica debe alcanzarse de forma que se garantice una variedad específica de patrones de entonación en oraciones enunciativas, interrogativas y exclamativas (Iriondo et al., 2007c). Hay que resaltar que se trata de un aspecto muy dependiente de la lengua y que el diseño variará sustancialmente si se pretende desarrollar un sistema multilingüe.

Contenido semántico. Con el objetivo de crear habla emocional estimulada, es decir que lo que se dice ayude al locutor a cómo decirlo, el texto debe incluir tanto una buena cobertura fonética como prosódica y, además, un contenido semántico adecuado para ayudar a expresar los estilos expresivos deseados. Es preferible empezar el proceso de diseño de los textos a partir de un corpus textual rico, ya que un incremento del número de estilos lo complicará (Navas et al., 2006), pudiendo resultar esta estrategia menos costosa que la redacción expresa de los textos. Las fuentes textuales en estos casos pueden ser cuentos, novelas, textos publicitarios o cualquier texto que ayude al locutor a generar la expresividad deseada.

4.2.2. Etiquetado

Una vez el corpus está grabado, el siguiente paso es el de etiquetar la señal de voz, a partir de disponer únicamente de esta y de la transcripción ortográfica y fonética. El proceso básico de etiquetado, que nos servirá para poder realizar análisis más complejos, se divide en los siguientes dos grandes bloques:

- **Segmentación.** Identificación de los límites temporales de los diferentes fonemas, de forma que se puede extraer la información de duración segmental.
- **Marcado de *pitch*.** Se lleva a cabo un análisis de la frecuencia fundamental (F_0) para cada uno de los pseudoperiodos de la señal, obteniéndose de este modo las marcas de *pitch* y la F_0 media de cada uno de los fonemas.

4.2.2.1. Segmentación

La segmentación se trata de la identificación de los límites temporales de las diferentes unidades de interés (por ejemplo los fonemas), de forma que se puede extraer la información de su duración segmental. La metodología empleada puede diferir entre autores, existiendo diferentes alternativas como por ejemplo las basadas en HMM (Monzo et al., 2008a) o en entropía (Kempe, 1999). Un ejemplo de este proceso se muestra en la Figura 4.1, donde se observan los límites de cada uno de los fonemas identificados sobre la señal de voz.

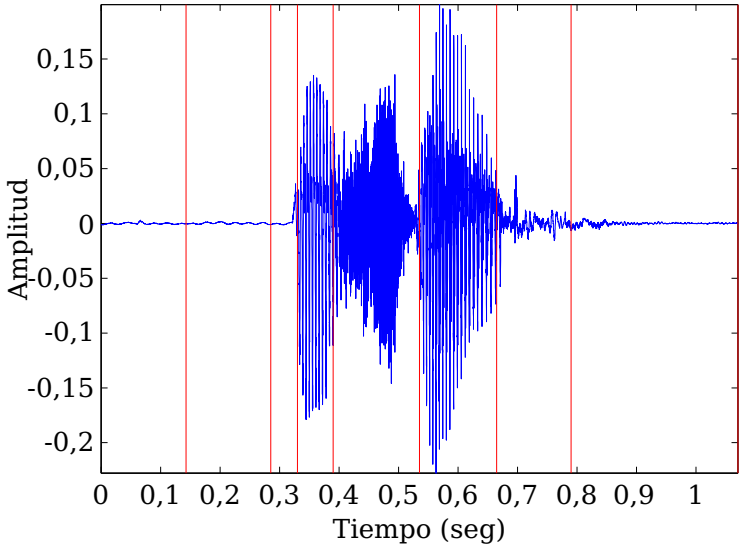


Figura 4.1: *Ejemplo de segmentación para la palabra "pisar"*

Dentro del grupo de investigación se ha trabajado en la evolución del proceso de segmentación. Desde el primer segmentador desarrollado (Alías y Iriondo, 2001b) se ha ido mejorando tanto en la calidad como en la usabilidad, mediante el hecho de hacerlo independiente de la lengua de interés, realizar la detección de las pausas en la señal de voz y el desarrollo de interfaces de usuario. Actualmente, tanto el proceso de entrenamiento como el proceso de etiquetado están basados en el uso de Modelo Oculto de Markov —*Hidden Markov Model*— (HMM). Para ello se dispone de código desarrollado en Matlab que se comunica a su vez con la herramienta *Hidden Markov Model Toolkit* (HTK)[1], automatizando el modelado y etiquetado, y aportando como valor añadido el control sobre los diferentes parámetros de configuración del HTK y de los corpus a etiquetar (Apéndice D.2).

El HTK es un conjunto de herramientas orientadas a la construcción y manipulación de HMM. Originariamente fue diseñado para ser utilizado en la investigación del reconocimiento del habla, aunque ha sido empleado en numerosas aplicaciones como la investigación de síntesis del habla, reconocimiento de caracteres y secuencias de ADN. Consiste en un conjunto de librerías y herramientas cuyo su código fuente, escrito en C, está disponible. Las herramientas proveen facilidades para el análisis del habla, entrenamiento de HMM, test y análisis de resultados. El software soporta HMM usando tanto densidades de probabilidad continuas de mezcla de Gaussianas como distribuciones discretas, y puede ser usado para construir sistemas HMM complejos. Fue originalmente desarrollado en el *Machine Intelligence Laboratory*, conocido formalmente como *Speech Vision and Robotics Group*, del *Cambridge University Engineering Department* (CUED). En 1993 *Entropic Research Laboratory Inc.* adquirió los derechos de venta del HTK. Su desarrollo fue plenamente traspasado a *Entropic* en 1995, cuando *Entropic Cambridge Research Laboratory Ltd.* fue

[1] http://htk.eng.cam.ac.uk/

establecida. En 1999 *Microsoft* compró *Entropic* y desde entonces CUED puede redistribuir HTK y proveer soporte de desarrollo vía web.

4.2.2.2. Marcado de *pitch*

El *pitch*, o tono, es un atributo perceptivo de los sonidos y el habla no sordos (Sección 2.3). Las marcas de *pitch*, juntamente con las marcas de segmentación, son uno de los elementos principales del etiquetado de un corpus.

Las marcas de *pitch* indican la ubicación temporal de los pseudoperiodos de la señal de voz (Figura 4.2). Se habla de pseudoperiodos por no ser perfectamente periódica esta señal. Estrictamente deberían denominarse marcas del periodo fundamental (T_0), ya que se trata del parámetro físico que etiquetan, aunque sin embargo, debido a que generalmente existe una relación directa entre *pitch* y T_0, se suele hablar de marcas de *pitch* (de Cheveigné y Kawahara, 2002). Las marcas de *pitch* son utilizadas en procesos que necesitan conocer la posición temporal de los periodos de la señal de voz (Veldhuis, 2000). En el contexto de un sistema de CTH destaca su utilidad en:

1. La **modificación** de la duración o el **tono** de la señal, como por ejemplo mediante PSOLA (Moulines y Charpentier, 1990).

2. Etiquetado **prosódico** sincrónico respecto a la periodicidad de la señal de voz (Black y Font Llitjós, 2002).

3. Etiquetado de parámetros de **VoQ** (Slyh et al., 1999).

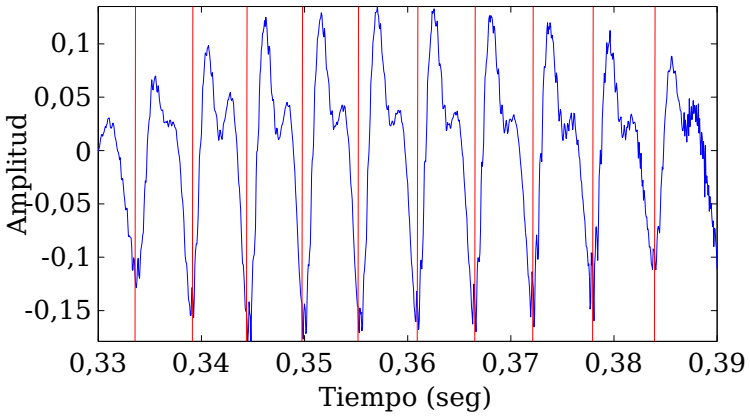


Figura 4.2: *Ejemplo de marcado de pitch (fonema /i/ en la palabra "pisar")*

El análisis automático de la periodicidad de la señal de voz pretende, por un lado, la detección o extracción de la periodicidad de la señal mediante un Algoritmo de Detección de *Pitch* —*Pitch Detection Algorithm*— (PDA), también conocido como algoritmo de seguimiento de *pitch* o *Pitch Tracking Algorithm* en inglés. Por otra parte, se quiere posicionar temporalmente las marcas de *pitch* identificando los diferentes pseudoperiodos de la señal de voz, utilizando para ello un Algoritmo de Marcado

de *Pitch —Pitch Marking Algorithm—* (PMA). Estos procesos son complejos, debido principalmente a los siguientes factores:

- El habla se genera a partir de un **proceso físico**, por lo tanto no es una señal perfectamente periódica sino cuasiestacionaria (Harbeck et al., 1995).
- Existen multitud de posibles estructuras o **patrones** en la señal de voz debido principalmente al locutor, al sonido que se esté articulando, la pronunciación o el estado emocional del interlocutor (Barner, 1996).
- La señal de excitación, o **pulso glótico**, no es una señal regular debido a la presencia de inestabilidades en el sistema de excitación (Sun, 2000) o a irregularidades del proceso de vibración glótica (de Cheveigné y Kawahara, 2002).

4.3. Corpus oral expresivo del grupo de investigación

El material de voz mayoritariamente utilizado en este trabajo se trata de un corpus oral expresivo en castellano, desarrollado en el grupo de investigación GTM. Este corpus fue desarrollado con dos propósitos principales. En primer lugar, para ser usado en el modelado acústico (prosodia y cualidad de la voz) de habla expresiva. En segundo lugar, para ser la base de datos que el sistema de CTH utilizaría.

4.3.1. Diseño del corpus

Para el desarrollo del corpus se contó con la ayuda de expertos del Laboratorio de Análisis Instrumental de la Comunicación (LAICOM) de la UAB. El corpus fue construido a partir de la lectura de textos cuyo contenido semántico ayudara a expresar el estilo deseado (habla estimulada). Estos textos, para cada una de los estilos expresivos, fueron leídos por una locutora profesional a lo largo de diferentes sesiones de grabación. Con esta estrategia se asume una disminución de la posibilidad de modelar habla informal espontánea, garantizando el control de las condiciones de grabación, el estilo definido y el diseño de los textos.

Para la selección de los textos, semánticamente relacionados con diferentes estilos expresivos, se hizo uso de una base de datos existente de publicidad, extraída de periódicos y revistas. En base a un estudio previo de publicidad audiovisual (Montoya, 1998), algunas de las categorías en las que estaba clasificada esta base de datos fueron consideradas como útiles a la hora de producir diferentes estilos expresivos.

Es importante mencionar que la locutora recibió previamente entrenamiento vocal sobre los patrones de cada uno de los estilos que se deseaba generar. Las características fonéticas, tanto segmentales como suprasegmentales, para estos patrones fueron definidas por los expertos del LAICOM. Gracias al uso de textos extraídos de cada una de las categorías publicitarias se ayudó a la locutora a mantener el estilo deseado a lo largo del total de sesiones de grabación y, de este modo, el estilo esperado no dependió en ninguna medida de su criterio. A pesar de que se realizaron

distintas sesiones de grabación, todos los textos relacionados con el mismo estilo fueron grabados consecutivamente en la misma, de forma que se minimizó la posibilidad de cambio de patrón. Como medida de control se realizó una supervisión de un experto durante todo el proceso para así evitar posibles desviaciones del estilo predefinido.

La asignación de las cinco categorías, seleccionadas a partir del corpus publicitario, a cinco estilos de habla expresivos se realizó como se indica a continuación:

- **Nuevas tecnologías**: estilo neutro (**NEU**) que transmite una cierta madurez.
- **Educación**: estilo alegre (**ALE**) que transmite una sensación de extraversión.
- **Cosmética**: estilo sensual (**SEN**) basado en una voz dulce.
- **Automóviles**: estilo agresivo (**AGR**) que transmite dureza.
- **Viajes**: estilo triste (**TRI**) que busca expresar cierto aire de melancolía.

La definición de estos cinco estilos pretendía proveer de la suficiente diversidad expresiva para poder realizar estudios en diferentes temas relacionados con el habla expresiva. A pesar de que no existía ninguna categoría en la base de datos que encajara con el estilo triste y, aunque la asignación a "viajes" fue un tanto artificial, se aceptó ya que la tristeza es de los estilos más sencillos de simular.

A partir del conjunto de textos disponibles, aquellos considerados como excepciones (palabras extranjeras, abreviaturas o palabras no estándares en general) fueron descartados para facilitar los posteriores procesos de transcripción fonética y etiquetado automático del corpus (Apéndice D.1). El conjunto de frases seleccionadas para cada una de las categorías fueron elegidas por medio de un algoritmo "voraz" (*greedy* en inglés) (François y Boëffard, 2002), que permitió la selección de frases fonéticamente balanceadas. La selección de frases similares fue penalizada por el algoritmo de forma que se maximizara la variabilidad. Para optimizar el proceso de selección los fonemas fueron ordenados según su porcentaje de aparición, permitiendo al algoritmo *greedy* minimizar el número de frases seleccionadas por empezar con aquellas que contienen los fonemas menos probables. De este modo se consiguió a su vez que la distribución fonética del corpus diseñado tendiera a la distribución fonética de la lengua de interés, en este caso el castellano. En Pérez (2003) se realiza un estudio de aparición de los fonemas que sirvió de base durante el proceso de selección, mostrando en las Tablas 4.1 y 4.2 la distribución fonética de vocales y consonantes del español comparando los porcentajes de aparición en el corpus diseñado y la media de los cinco estudios presentados en este. Por último, solamente aclarar que la notación utilizada es *Speech Assessment Methods Phonetic Alphabet* (SAMPA) (Wells, 1997).

El corpus, una vez segmentado, finalmente consta de 4638 frases y dura 5 horas y 27 minutos. Más concretamente se tienen 833 frases para el estilo neutro (50 min), 916 para el estilo alegre (56 min), 841 para el estilo sensual (51 min), 1048 para el estilo agresivo (84 min) y 1000 para el estilo triste (86 min). En cuanto a la duración de los estilos neutro, alegre y sensual se está trabajando en su ampliación para balancear los tamaños de todos los subcorpus.

(%)	/a/	/e/	/i/	/o/	/u/
Corpus diseñado	12,74	13,56	6,13	9,24	2,74
Media en Pérez (2003)	13,27	13,13	6,32	9,71	2,32

Tabla 4.1: *Comparación del porcentaje de aparición de las vocales en el total del corpus diseñado y el promedio de los cinco estudios presentados en Pérez (2003)*

(%)	/p/	/t/	/k/	/b/	/d/	/g/	/n/	/m/	/J/
Corpus diseñado	2,70	4,82	3,84	2,67	4,59	0,99	6,27	3,44	0,22
Media en Pérez (2003)	2,66	4,66	4,02	2,66	4,58	1,02	5,3	2,73	0,28
(%)	/s/	/x/	/C/	/T/	/r/	/R/	/l/	/L/	/f/
Corpus diseñado	7,51	0,85	0,24	1,83	5,72	0,92	4,99	0,32	0,81
Media en Pérez (2003)	8,72	0,65	0,34	1,89	4,48	0,69	4,86	0,57	0,74

Tabla 4.2: *Comparación del porcentaje de aparición de las consonantes en el total del corpus diseñado y el promedio de los cinco estudios presentados en Pérez (2003)*

Complementariamente a los textos seleccionados, con la finalidad de garantizar la aparición de todos los difonemas y los trifonemas utilizados durante la síntesis, se creó una lista de palabras portadoras que los contenían. Cada unidad de palabras portadoras está formada bien por una sola palabra si en su interior contiene la unidad requerida, bien por dos palabras si la unidad aparece por contacto del final de la primera palabra con el inicio de la segunda. Además de asegurar la presencia de los difonemas y trifonemas, esta lista de palabras permite realizar comparaciones directas entre los parámetros acústicos de los 5 estilos, aunque solamente sea a nivel segmental. Actualmente, esta lista contiene 1250 palabras. Un ejemplo de la aplicación de la lista de palabras que da cobertura al corpus la encontramos en el trabajo presentado por Alías et al. (2005).

Con la finalidad de comprobar el etiquetado llevado a cabo sobre cada uno de los estilos de habla expresivos de los que consta el corpus, se trabajó en el desarrollo de la herramienta presentada en el Apéndice D.4.

En la Tabla 4.3 se resume el número de palabras y de frases portadoras, juntamente con su duración, en función del corpus una vez este ha sido segmentado.

	Frases cantidad / duración	**Palabras portadoras cantidad / duración**
Neutro	833 / 50 min	1250 / 22 min
Alegre	916 / 56 min	1250 / 25 min
Sensual	841 / 51 min	1250 / 31 min
Agresivo	1048 / 84 min	1250 / 25 min
Triste	1000 / 86 min	1250 / 24 min

Tabla 4.3: *Resumen del número y duración de frases y palabras portadoras en el corpus expresivo una vez segmentado*

4.3.2. Grabación

La grabación del corpus oral fue llevada a cabo en el estudio de grabación de *Enginyeria i Arquitectura La Salle* de la URL (EALS-URL). Este consta de dos zonas diferenciadas: la sala de control, donde se encuentra el equipamiento de mezcla y producción, y la sala de grabación. Ambas salas están acústicamente acondicionadas para cumplir con las condiciones requeridas y ofrecer un alto grado de aislamiento.

La sala de grabación tiene forma irregular con una planta de 5 por 4 metros y una altura de $3,5$ metros. El tiempo de respuesta de la sala es de aproximadamente $0,8$ segundos aunque la situación relativa entre el locutor y el micrófono garantiza la ausencia de ecos audibles.

Para la captura de la señal de voz se utilizó un micrófono de condensador (AKG C-414) con una respuesta prácticamente plana (2 dB en el rango de $20 - 20000$ Hz) y una relación señal a ruido de 80 dBA SPL. La captura se realizó directamente en un disco duro mediante la plataforma digital Pro Tools 5.1 instalada en un ordenador Mac G5, utilizando una consola digital Yamaha 02R. Finalmente, para la digitalización de la señal de voz se utilizó una frecuencia de muestreo de 48 KHz y una cuantificación de 24 bits en ficheros de tipo WAV.

El protocolo de actuación usado durante toda la grabación, con el objetivo de optimizar los procesos de etiquetado que seguirían, implicaba la presencia de cuatro personas haciendo el seguimiento de la sesión:

- **Locutora profesional**: quien sería la encargada de leer los textos.
- **Ingeniero de audio**: responsable de realizar los ajustes de la plataforma de grabación y de la posición del micrófono y de la locutora.
- **Experto en comunicación audiovisual**: encargado de entrenar a la locutora en los distintos estilos expresivos y corregir posibles desviaciones de esta respecto del modelo deseado.
- **Técnico de control**: responsable de garantizar la correcta lectura de los textos.

4.3.3. Segmentación

El alineamiento de los diferentes fonemas o segmentación, es decir, la identificación temporal del inicio y final de cada uno de ellos, se ha realizado mediante alineamiento forzado utilizando la herramienta HTK (Sección 4.2.2.1). El alineamiento ha podido ser forzado por disponer de la transcripción fonética de cada una de las frases, de modo que dicha información se ha utilizado como gramática para guiar a los modelos de fonemas durante el proceso de alineamiento.

Una vez segmentados automáticamente cada uno de los archivos de audio del corpus, se revisó manualmente el resultado y corrigieron algunas frases que contenían errores de segmentación. Los errores principalmente eran debidos a la falta de coherencia entre el pausado realizado por la locutora y los signos de puntuación. En los casos en que aparecían silencios en el fichero de audio, sin el correspondiente

signo de puntuación, se modificó el texto y la transcripción fonética. De forma inversa, aquellas pausas no realizadas por la locutora y que, en cambio, estaban marcadas por un signo de puntuación, se corrigieron mediante la eliminación de este del texto y de la transcripción fonética.

El resultado de esta segmentación se puede utilizar para la extracción de parámetros acústicos segmentales. Estos parámetros pueden ser tanto de prosodia como de VoQ, y utilizados en aplicaciones donde se requiera un modelado de los mismos. Ejemplos de uso son el reconocimiento de emociones (análisis) y la síntesis del habla (análisis y síntesis).

4.3.4. Marcado de *pitch*

El marcado de *pitch* llevado a cabo para el etiquetado del corpus desarrollado, se basó en el marcador que utiliza el algoritmo *Robust Algorithm for Pitch Tracking* (RAPT) (Talkin, 1995), al que se le aplicó el Algoritmo de Filtrado de Marcas de *Pitch* —*Pitch Marks Filtering Algorithm*— (PMFA), desarrollado por Alías et al. (2006), para obtener las marcas de *pitch* finales. Este algoritmo fue desarrollado para mejorar la robustez de las marcas de *pitch*, optimizando así el etiquetado de los corpus de voz utilizados en el contexto de los sistemas de CTH. Con el objetivo de evaluar el marcado de *pitch* realizado, se trabajó en la evaluación de sistemas de marcado de *pitch* para medir la efectividad y la mejor configuración del PMFA (Apéndice D.3).

Se trata de un algoritmo que fue concebido inicialmente como un módulo de postprocesamiento para cualquier PMA (Alías y Iriondo, 2001a), pero que posteriormente se adaptó para ser utilizado también como algoritmo de posicionamiento temporal de las marcas de *pitch* dado un PDA cualquiera (Figura 4.3). A este algoritmo se le conoce como PMFA (Alías et al., 2006), ya que mejora la robustez del marcado de la señal de voz a partir de las marcas que recibe como información de entrada, filtrando las marcas espurias debidas a inserciones y/u omisiones (exceso o falta de marcas respectivamente). Por lo tanto, PMFA persigue ubicar con precisión las marcas de *pitch* sobre la señal de voz a partir de una secuencia inicial de marcas $m^i(n)$ (marcas de entrada) y $m^f(n)$ (marcas finales). En el caso de provenir la información de F_0 de un PDA se utilizará un Marcador Simple de *Pitch* —*simple Pitch Marking Algorithm*— (sPMA) para la ubicación inicial de las marcas. Por tanto, la secuencia de marcas de *pitch* iniciales $m^i(n)$ en el contexto del algoritmo desarrollado puede corresponder a:

1. Las marcas obtenidas por un PMA de entrada ($m^i(n) = PMA[x(n)]$),
2. Las marcas generadas mediante un algoritmo simple de marcado de *pitch* (sPMA) a partir del valor de la frecuencia fundamental por trama (F_0) obtenida por el PDA de entrada ($m^i(n) = sPMA[PDA[x(n)]]$).

Donde $x(n)$ corresponde al vector de $1 \leq n \leq M$ muestras de la señal de entrada y $m^i(n)$ representa el vector de marcas de *pitch* inicial de M muestras, con posiciones iguales a la unidad en las muestras correspondientes a las posiciones de las marcas de *pitch*, definición aplicable a $m^f(n)$. Si el PDA suministra los valores de cada trama

$F_0(t)$ en Hz, para $1 \leq t \leq T$ tramas de análisis, estos se convierten a información de periodicidad (muestras) mediante $T_0(t) = round(f_s/F_0(t))$, siendo t el índice de trama y f_s la frecuencia de muestreo utilizada.

El desarrollo del algoritmo está basado en la programación dinámica, inspirado en los trabajos presentados por Harbeck et al. (1995) y Goncharoff y Gries (1998). Utiliza como dato de entrada la secuencia de marcas de *pitch* obtenida de un PMA o los valores de F_0 entregados por un PDA.

Su ajuste se trata de un procedimiento sencillo. Se determina la variación máxima trama a trama y se fija el criterio local escogido según muestra la Figura 4.3. En cuanto a este criterio local, escogido para ubicar las marcas de *pitch*, se sitúan siguiendo los máximos en valor absoluto de la señal de voz dentro de las tramas periódicas. Por lo que se refiere a tramas aperiódicas, es decir silencios o tramos sordos, las marcas se distribuyen de forma que la transición es suave entre los valores de T_0 de los tramos sonoros anterior y posterior. Asimismo, PMFA no evalúa la sonoridad de la señal evitando incorporar los errores de este módulo durante la fase de posicionamiento temporal de las marcas.

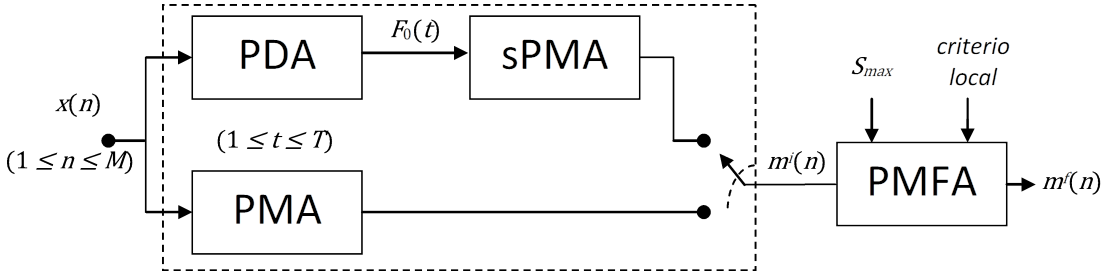


Figura 4.3: *Diagrama de bloques del sistema automático de marcado de pitch usando PMFA*

Partiendo del conjunto inicial de marcas $m^i(n)$, obtenidas a partir de la señal $x(n)$ utilizada, el PMFA obtiene las marcas de *pitch* finales $m^f(n)$ mediante un proceso dividido en dos fases:

1. **Filtrado de errores**: se eliminan los errores de $m^i(n)$ debidos a posiciones de marcas o valores espurios a partir de la primera aplicación del algoritmo de programación dinámica, restringido según la variación máxima de T_0 intertrama permitida (S_{max}).

2. **Ubicación temporal** de las marcas:

   a) Primera estimación de la posición temporal de las marcas a partir de los valores de T_0 filtrados con el proceso de optimización dinámica.

   b) Refinamiento de la posición temporal de las marcas según el criterio local escogido, es decir el máximo en valor absoluto de la señal dentro del periodo, mediante una segunda pasada del algoritmo de programación dinámica restringido.

Finalmente, una de las principales características de PMFA es la simplicidad:

- No necesita ninguna función de coste compleja, solamente ajustar los valores de S_{max}.
- Filtra los errores de PDA y PMA usando un esquema simple de votación binario.
- No intenta discriminar la sonoridad del habla, en zonas sordas con vecinas sonoras las transiciones se suavizan (Goncharoff y Gries, 1998).

4.4. Otros corpus orales expresivos utilizados

Junto con el corpus del grupo (Sección 4.3), mayoritariamente utilizado en los experimentos llevados a cabo en esta tesis, se hizo uso de otros corpus con los que pudieron ser complementados. Estos son dos corpus expresivos en alemán, el primero de ellos grabado en un entorno controlado (Sección 4.4.1) y el segundo de ellos todo lo contrario, se trata de un corpus de habla expresiva espontánea, donde no se impuso ningún tipo de control sobre los enunciados (Sección 4.4.2).

4.4.1. Corpus de habla expresiva en un entorno controlado

Este corpus de habla expresiva fue desarrollado por Burkhardt et al. (2005). Se trata de un corpus en alemán y en su creación participaron diez actores (5 mujeres y 5 hombres) simulando los diferentes estilos expresivos/emociones, produciendo 10 enunciados (5 frases cortas y 5 largas) que podrían aparecer en situaciones cotidianas y con la posibilidad de ser interpretadas en todos los estilos de habla expresivos.

Las grabaciones fueron realizadas en un cámara anecoica[2] con un equipo de grabación de alta calidad, donde además de la señal de voz también fue grabada la señal de EGG. El material de voz comprende unas 800 frases, 7 estilos de habla expresivos $\times$ 10 actores $\times$ 10 frases $+$ algunas segundas versiones.

Los estilos de habla expresivos utilizados fueron: neutro (*neutral*), enfado (*anger*), miedo (*fear*), alegría (*joy*), tristeza (*sadness*), asco (*disgust*) y aburrimiento (*boredom*). La totalidad del corpus fue evaluado en un test perceptivo en lo que concierne al reconocimiento de los estilos de habla expresivos y a su naturalidad, obteniendo los resultados de clasificación presentados en la Tabla 4.4: enfado ($96,9\,\%$), neutro ($88,2\,\%$), miedo ($87,3\,\%$), aburrimiento ($86,2\,\%$), alegría ($83,7\,\%$), tristeza ($80,7\,\%$) y asco ($79,6\,\%$). Aquellas realizaciones con un factor de reconocimiento superior al $80\,\%$, y juzgadas como naturales por más del $60\,\%$ de los oyentes, fueron fonéticamente etiquetadas con marcadores especiales para la VoQ, configuraciones fonatorias y articulatorias y características articulatorias, resultando en 535 de las 800 frases. El número de frases del que finalmente consta el corpus, por estilo expresivo y su duración, se muestran en la Tabla 4.4.

[2]Una cámara anecoica o anecoide es una sala especialmente diseñada para absorber el sonido que incide sobre las paredes, el suelo y el techo de la misma cámara, anulando los efectos de eco y reverberación del sonido.

	% reconocimiento	Número de frases	Duración (min)
Neutro	88,2	79	3,1
Enfado	96,9	127	5,6
Miedo	87,3	69	2,6
Alegría	83,7	71	3,0
Tristeza	80,7	62	4,2
Asco	79,6	46	2,6
Aburrimiento	86,2	81	3,8
Total	- -	*535*	*24,9*

Tabla 4.4: *Estilos de habla expresivos del corpus en alemán y su porcentaje de reconocimiento, número de frases que lo forman y duración*

4.4.2. Corpus de habla expresiva espontánea

El corpus de habla expresiva espontánea en cuestión se trata de la base de datos FAU AIBO (Steidl, 2009), utilizado en *INTERSPEECH 2009 Emotion Challenge* (Schuller et al., 2009).

Es un corpus con grabaciones de niñas y niños interactuando con Aibo, el robot mascota de Sony. El corpus consta de habla espontánea en alemán con contenido expresivo. A los niños se les hizo creer que Aibo respondía a sus órdenes cuando realmente era controlado por un operario. Aibo actuaba de una forma fija con una secuencia de acciones predeterminada, algunas veces de forma desobediente, causando así reacciones emocionales. La información se capturó en dos escuelas distintas, con un total de 51 niños de edades comprendidas entre 10 y 13 años, siendo 21 niños y 30 niñas; con una duración total de aproximadamente 9 horas de habla sin pausas.

Las grabaciones fueron segmentadas automáticamente en turnos usando una pausa de 1 segundo. Cinco personas (estudiantes avanzados de lingüística) escucharon los turnos en orden secuencial y anotaron cada palabra independientemente del resto como neutro (por defecto) o perteneciente a una de otras 10 posibles clases. Ya que la mayoría de los enunciados son solamente órdenes cortas y que pueden haber pausas largas entre palabras debido al tiempo de reacción de Aibo, el estado emocional o emoción del niño puede cambiar también dentro de los turnos, de ahí que la información se etiquetara a nivel de palabra. A partir del voto mayoritario de los cinco anotadores, la etiqueta se asignó a la palabra, obteniendo las siguientes asignaciones sobre las 10 posibles clases elegidas:

- Neutra (*neutral*): 39169
- Alegre (*joyful*): 101
- Sorprendido (*surprise*): 0
- Enfático (*emphatic*): 2528
- Impotente (*helpless*): 3
- Irritado (*touchy/irritated*): 225

- Enfadado (*angry*): 84
- Habla materna (*motherese*), tal y como habla un adulto a un niño pequeño: 1260
- Aburrido (*bored*): 11
- Reprimenda (*reprimanding*): 310
- Resto de clases, no neutras, que no pertenecen a las anteriores: 3
- No etiquetadas: 4707

Finalmente se realizó una agrupación de las diferentes clases, dejando 5 clases que caracterizaran el problema:

- **Neutro**
- **Enfado**: incluyendo enfadado, irritado y reprimenda
- **Enfático**
- **Positivo**: incluyendo alegre y habla materna.
- **Resto de clases**

Además, este corpus proporciona una única transcripción detallada del contenido locutado con límite de palabras, vocalizaciones no lingüísticas, etiquetas de emociones y análisis de unidades.

CAPÍTULO 5

Modelado de la cualidad de la voz

A lo largo de este capítulo se presentan las aportaciones realizadas en el modelado de la Cualidad de la Voz —*Voice Quality*— (VoQ) con el fin de ser aplicado en Síntesis del Habla Expresiva (SHE). Los experimentos realizados se agrupan en los siguientes puntos:

- Definición de la fases que se siguieron para el modelado de la VoQ (Sección 5.1).
- Una vez definidas las fases del modelado, el primer punto a tratar será la parametrización del habla (Sección 5.2).
- Con la detección de las limitaciones de los parámetros utilizados se proponen nuevas metodologías que los adecuen al reconocimiento de emociones y a la SHE (Sección 5.3).
- Seleccionados los parámetros de VoQ se presentan los experimentos acerca de la capacidad discriminatoria que estos tienen sobre los estilos de habla expresivos (Sección 5.4).
- A partir de los experimentos anteriores se dispone de la información necesaria para el modelado de la VoQ, de modo que se proponen las metodologías de trasnformación de los estilos de habla expresivos mediante la modificación de la VoQ para sistemas de SHE (Sección 5.5).
- El último de los puntos que se presenta, fruto de la colaboración con otros trabajos de investigación, es el papel que tuvo la VoQ en dichas aplicaciones (Sección 5.6).

Los resultados de cada uno de los experimentos realizados fueron justificados mediante un proceso de evaluación.

5.1. Fases del modelado de la cualidad de voz

En la búsqueda del conocimiento de qué parámetro o qué combinación de parámetros era la más idónea para modelar estilos de habla expresivos, se ha elaborado un plan de trabajo dividido en las siguientes 5 fases principales:

1. Determinar qué parámetros permiten identificar a cada uno de los estilos expresivos, de utilidad tanto en el reconocimiento de emociones como en la SHE.

2. Calcular las estadísticas de los parámetros de VoQ seleccionados en la fase anterior, para cada estilo de habla expresivo del corpus descrito en la Sección 4.3, de utilidad en la transformación de estilos de habla expresivos para la SHE y en el análisis objetivo de la calidad de la SHE o, en general, del habla generada por un sistema de CTH.

3. Proponer alternativas para la metodologías de análisis y modificación de los parámetros de VoQ *jitter* y *shimmer*, con el objetivo de mejorar su modelado en aplicaciones de SHE.

4. Aplicar, en el proceso de SHE, las relaciones halladas entre estilos de habla expresivos, los parámetros y los valores que mejor las modelan.

5.2. Los parámetros de cualidad de la voz

5.2.1. Elección de los parámetros de cualidad de la voz

La elección de los parámetros de VoQ, implicados en la representación de la expresividad de un mensaje oral, es el primer paso para trabajar en la SHE.

Tal y como se presenta en la Sección 2.4.2 son múltiples las opciones de las que se dispone para trabajar con parámetros de VoQ, debiendo de adecuar la decisión de cuáles se usarán en función de varios factores: la aplicación para la que serán utilizados, el material del que disponemos o los requisitos de computación necesarios.

En nuestro caso concreto estamos trabajando con un sistema de CTH en el contexto de SHE, con lo que desearemos disponer de parámetros que puedan ser calculados a partir de un corpus, analizados y modificados durante la síntesis del habla. Estos parámetros han sido utilizados para el modelado de la VoQ, dividido en dos procesos:

1. **Análisis**. Extracción de información a partir de medidas realizadas sobre el corpus. Esta información ha permitido conocer el comportamiento de los parámetros sobre ejemplos reales de habla, pudiendo así extraer reglas y relaciones para ser aplicadas durante el proceso de modificación.

2. **Síntesis**. A partir del proceso de análisis, se ha extraído la información sobre los parámetros que modelan el estilo expresivo durante el proceso de síntesis del habla.

En el caso que ocupa a esta tesis, los parámetros de VoQ han sido complementarios a la información de prosodia (Sección 2.3), sin entrar en discusión sobre cómo esta modelará los estilos expresivos disponibles en el corpus (Sección 4.3).

Pasemos a ver la selección de los parámetros de VoQ. Como se muestra en la Sección 2.4.2, existen diversas metodologías para la obtención de la información de VoQ. Basándonos en estas alternativas y en el hecho de que sería deseable depender únicamente de la señal acústica, para así evitar tener que acceder al cuerpo del interlocutor mediante hardware extra o transductores invasivos (Lugger y Yang, 2006b), las alternativas se redujeron a la extracción de parámetros a partir de la señal acústica. Asimismo, para facilitar la manipulación de estos parámetros y evitar usar un modelo de señal glótica, que puedan añadir errores debidos al propio modelo, los parámetros que finalmente fueron considerados son los parámetros acústicos extraídos de la señal acústica (Sección 2.4.2.4), de los que se seleccionaron, siguiendo la propuesta presentada por Drioli et al. (2003), los siguientes (Monzo et al., 2007):

- *Jitter*
- *Shimmer*
- *Harmonic-to-Noise Ratio* (HNR)
- *Glottal-to-Noise Excitation Ratio* (GNE)
- *Spectral Flatness Measure* (SFM)
- *Drop-off of Spectral Energy above 1000 Hz* (do1000)
- *Hammarberg Index* (HammI)
- *Relative Amount of Energy above 1000 Hz* (pe1000)

Como se puede observar, el parámetro NNE no forma parte de los parámetros seleccionados, principalmente por presentar similitudes con el HNR.

Los primeros experimentos tenían como característica que la parametrización se realizaba sobre las vocales, que como se explica en la Sección 5.4 son zonas sonoras estables en las que se obtiene el cálculo robusto de los parámetros de VoQ, para lo cual se hizo uso de la herramienta Praat (Boersma, 2001). En cambio, en el momento de plantearse la transformación de estilos de habla expresivos para la SHE mediante HNM (Sección 5.5), apareció la necesidad de un análisis de un mayor número de unidades, independientemente de si eran o no vocales, debiendo de adaptar el proceso de análisis al modelo de HNM usado. Además, se planteó la metodología de modificación de la VoQ para la síntesis del habla expresiva, todo ello desarrollado en Matlab®[1], pudiendo realizar prototipos, actualizaciones, nuevas propuestas y experimentos.

Realizados los primeros experimentos, se consideró la modificación de la medida del *jitter* y del *shimmer*, proponiendo también su modificación durante el proceso de

[1]Matlab®, abreviatura de *MATrix LABoratory*, es un lenguaje de alto nivel y un entorno interactivo que permite realizar tareas de cálculo complejas.

síntesis. Este cambio de metodología vino dado por la dependencia que tenían estos parámetros con los parámetros de prosodia: frecuencia fundamental (F_0) y energía, cuando de estilos de habla expresivos se trataba. La nueva medida de *jitter* y de *shimmer* pretende hacerlos independientes de la F_0 y de la energía respectivamente, obteniendo así valores sin interferencias debidas a la prosodia (Monzo et al., 2008b).

5.2.2. Análisis de los parámetros de cualidad de la voz

El análisis de la VoQ pasó por diferentes etapas en función de las necesidades de cada momento. Las dos herramientas de las que se hizo uso fueron Praat y Matlab a partir de la parametrización del habla basada en HNM.

La primera alternativa fue la medida en Praat de los parámetros presentados en la Sección 5.2.1, mediante los algoritmos y las funciones propias que este implementa (Boersma, 2001). El principal beneficio fue que disponía de funciones específicas para el análisis de varios de los parámetros de interés, permitiendo la realización de los primeros experimentos. Otra ventaja fue el hecho que es una herramienta ampliamente conocida, pudiendo presentar a la comunidad científica los resultados obtenidos con mayores garantías.

En segundo lugar se migró el código Praat hacia Matlab por varios motivos:

- Matlab presentaba las mejores características para el desarrollo de nuevas propuestas metodológicas para el modelado de la VoQ (Sección 5.3).
- Matlab permitiría integrar con mayor simplicidad el análisis de VoQ en aplicaciones en las que estaban trabajando otros miembros del grupo de investigación (Sección 5.4.4).
- La transformación de estilos de habla expresivos en SHE, usando las herramientas para el análisis del habla basado en HNM requirió analizar los parámetros siguiendo dicho modelo HNM, con lo cual fue el mejor medio para la integración de todas ellas en un único *toolbox* (Sección 5.5).

A pesar de las variantes en el cálculo de las parametrizaciones, las relaciones entre parámetros y entre diferentes estilos de habla expresivos se mantuvieron en gran medida constantes. Es evidente que el posible cambio de metodología de análisis, que se pudiera dar por basar la medida en HNM en lugar de los algoritmos de los que dispone Praat, puede llevar a resultados absolutos diferentes, pero el interés de los experimentos se centra en la relación entre los parámetros y los estilos expresivos que caracterizan.

Dado que el desarrollo de la metodología de análisis usando HNM se aleja de la definición dada en la Sección 2.4.2.4, ya que no se basa en el análisis de la señal acústica sino en un modelo de la misma, en las Secciones 5.2.2.1 a 5.2.2.8 se muestra el procedimiento a seguir para el análisis de cada uno de los parámetros de VoQ bajo los dos enfoques: medida basada en la definición dada en la Sección 2.4.2.4 (y el cambio de metodología del *jitter* y del *shimmer*) y en el uso de HNM.

En el Apéndice B se complementa la información presentada en esta sección, presentando los siguientes aspectos relacionados con el proceso de análisis:

- La descripción de la parametrización del habla basada en HNM.
- Las consideraciones necesarias a tener en cuenta para la parametrización de la VoQ utilizando HNM.
- Se define la notación seguida en las ecuaciones de análisis de cada uno de los parámetros de VoQ (Secciones 5.2.2.1 a 5.2.2.8).

5.2.2.1. *Jitter*

Para la medida del *jitter* presentada en la Sección 2.4.2.4 se hizo uso de las funciones propias de la herramienta Praat, que devuelven el valor del parámetro para un segmento de voz dado y el marcado de *pitch* realizado sobre este (a partir del método de correlación cruzada con la que cuenta Praat). Los valores asignados a la herramienta para el cálculo del *jitter* fueron los valores asignados por defecto, con una F_0 mínima de 75 Hz y una máxima de 600 Hz utilizados para el marcado de *pitch*. El valor final se expresó como un $\%$.

Debido a que el objetivo final del análisis es el modelado de la VoQ para la transformación de estilos de habla expresivos durante la SHE, la metodología finalmente propuesta se trata de la nueva metodología presentada en la Sección 5.3.2, en la que este parámetro se aleja de su definición original (Sección 2.4.2.4) para adaptarse a las características del habla expresiva. Para esta nueva metodología de análisis del *jitter*, el procedimiento de análisis es independiente del empleo de HNM, por no depender de los parámetros extraídos de este. Se extrae la información de F_0 de las frecuencias del modelo, información que de otro modo hubiera sido extraída usando un marcador o detector de *pitch*.

5.2.2.2. *Shimmer*

La medida del *shimmer* utilizando Praat se sirvió de las funciones propias de la herramienta Praat, las cuales implementan la medida del *shimmer* para un segmento de voz dado siguiendo la definición dada en Sección 2.4.2.4. Para poder realizar esta medida, la herramienta primero calculó las marcas de *pitch*, utilizando el método que implementa de la correlación cruzada. Los parámetros asignados durante este proceso de medida fueron los utilizados por defecto por la herramienta, con una F_0 mínima de 75 Hz y una máxima de 600 Hz fijada durante el marcado de *pitch*. El valor finalmente asignado a este parámetro se expresó como un $\%$.

La metodología finalmente empleada es la presentada en la Sección 5.3.2, difiriendo de la definición original mostrada debido a su adaptación para ser empleada en aplicaciones de análisis de habla expresiva. Del mismo modo que pasa para el *jitter*, el *shimmer* es independiente del empleo de HNM, con la única salvedad de que en el caso de ser usado con HNM la información de voz empleada está asociada a la componente determinista en lugar de a la señal de voz original, y la información de

pitch se obtiene a partir de las frecuencias del modelo, evitando así la influencia que el ruido pudiera ocasionar.

5.2.2.3. *Harmonic-to-Noise Ratio* (HNR)

La relación entre la energía armónica y del resto de la señal (considerada como ruido) se puede realizar mediante los dos enfoques siguientes. El primero utilizó la herramienta Praat, que implementa el algoritmo basado en el análisis de la correlación cruzada, que extrae este parámetro para un segmento de voz, siguiendo la definición dada en Sección 2.4.2.4 y expresando en dB el resultado. Debido a la implementación del parámetro por parte de la herramienta, el valor finalmente asignado se calculó como la media de la medida realizada en el segmento de voz en ventanas de 10 ms, fijando un valor de F_0 mínima de 75 Hz.

En segundo lugar se definió el análisis del HNR adaptado al uso de la parametrización del habla basada en HNM. Se define como el cociente de energías de la parte armónica y ruidosa de la voz en un determinado segmento de voz (o periodo de integración).

$$HNR = 10\log\left(\frac{E_s}{E_r}\right) = 10\log\left(\frac{\sum_{k=1}^{K} E_k^s}{\sum_{j=1}^{J} E_j^r}\right) = 10\log\left(\frac{\sum_{k=1}^{K} T_k^s \sum_{i=1}^{I_k} \frac{A_{ik}^2}{2}}{\sum_{j=1}^{J} T^r \sigma_j^2}\right) \quad [dB] \qquad (5.1)$$

Pudiendo ser aproximada la expresión de la Ecuación 5.1 por:

$$HNR \approx 10\log\left(\frac{J}{K}\frac{\sum_{k=1}^{K}\sum_{i=1}^{I_k} \frac{A_{ik}^2}{2}}{\sum_{j=1}^{J} \sigma_j^2}\right) \quad [dB] \qquad (5.2)$$

Siendo ambas expresiones idénticas en el caso que las tramas de la componente determinista tengan longitud constante (sería el caso $T_k^s = T^s$).

5.2.2.4. *Glottal-to-Noise Excitation Ratio* (GNE)

El parámetro GNE fue calculado únicamente a partir del análisis con la herramienta Praat, haciendo uso del algoritmo que esta implementa, que calcula el valor de este parámetro para un segmento de voz dado, expresando el resultado entre 0 y 1. Los valores utilizados por el algoritmo para la medida fueron extraídos a partir del estudio realizado por Michaelis et al. (1997):

- Paso de ventana: 80 Hz
- Ancho de banda: 3000 Hz

- Frecuencia mínima: 1500 Hz
- Frecuencia máxima: 4500 Hz

Por otro lado, este parámetro no fue implementado cuando se empleó HNM para la parametrización del habla. En su lugar se utilizó el parámetro HNR, que podía ser calculado directamente a partir de las componentes determinista y estocástica.

5.2.2.5. *Spectral Flatness Measure* (SFM)

En el caso del parámetro SFM, se implementó su análisis utilizando la herramienta Praat hasta que surgió la necesidad de disponer de un procedimiento de análisis acorde a la parametrización del habla basada en HNM.

Con Praat, el análisis siguió la definición dada en la Sección 2.4.2.4, pero en este caso no existió una función propia de la herramienta que lo calculara. Por tanto, a partir de un segmento de habla, se extrajo la densidad espectral de potencia mediante el análisis LTAS que Praat implementa, sobre la cual se calcularon la media geométrica y la aritmética de la energía en la banda bajo análisis, expresando el resultado en dB.

Para el caso del análisis utilizando el modelo HNM, el SFM se define como una medida promediada, a lo largo de todo un segmento de voz, del cociente entre la media geométrica y la media aritmética del perfil de energía espectral calculada de forma independiente para cada trama dentro del segmento.

$$SFM_k = \frac{\sqrt[I_k]{\prod_{i=1}^{I_k} E_{ik}^s}}{\frac{1}{I_k}\sum_{i=1}^{I_k} E_{ik}^s} = \frac{\sqrt[I_k]{\prod_{i=1}^{I_k} \frac{A_{ik}^2}{2} T_k^s}}{\frac{1}{I_k}\sum_{i=1}^{I_k} \frac{A_{ik}^2}{2} T_k^s} = \frac{T_k^s \sqrt[I_k]{\prod_{i=1}^{I_k} \frac{A_{ik}^2}{2}}}{\frac{T_k^s}{I_k}\sum_{i=1}^{I_k} \frac{A_{ik}^2}{2}} = \frac{\sqrt[I_k]{\prod_{i=1}^{I_k} \frac{A_{ik}^2}{2}}}{\frac{1}{I_k}\sum_{i=1}^{I_k} \frac{A_{ik}^2}{2}} \tag{5.3}$$

$$SFM = 10\log\left(\frac{1}{K}\sum_{k=1}^{K} SFM_k\right) \qquad [dB] \tag{5.4}$$

5.2.2.6. *Drop-off of Spectral Energy above 1000 Hz* (do1000)

Este parámetro representa la pendiente del espectro de energía de la señal de voz por encima de 1000 Hz, y se calcula a partir de una aproximación por mínimos cuadrados de esta banda del espectro. Para su medida se realizaron dos enfoques, el primero utilizando Praat y el segundo, para poder trabajar en SHE, definiendo el procedimiento teniendo en cuenta la parametrización del habla basada en HNM.

En Praat no existió un algoritmo que hiciera su medida directamente, con lo que se extrajo la densidad espectral de potencia mediante el análisis LTAS que Praat implementa, y sobre estos valores se calculó la aproximación por mínimos cuadrados de la pendiente espectral.

Para el caso considerando el modelo HNM, en el que se utiliza únicamente la componente armónica de la señal, esta aproximación se realiza para cada trama dada k del segmento de voz analizado, y a partir de las amplitudes y frecuencias de los armónicos con índices $i \in [i_k^{1000}+1, i_k^{1000}+2, \ldots, I_k]$, llegando, por lo tanto, a aproximar la pendiente espectral en la banda efectiva de frecuencias $f \in (1000Hz, 5000Hz]$. Según la definición de do1000 dada en Sección 2.4.2.4 y estas consideraciones, se obtienen las siguientes ecuaciones de estimación:

$$\overline{F}_k^{1000} = \left(\frac{1}{I_k - i_k^{1000}}\right) \sum_{i=i_k^{1000}+1}^{I_k} F_{ik} \tag{5.5}$$

$$\begin{aligned} \overline{E}_k &= \left(\frac{1}{I_k - i_k^{1000}}\right) \sum_{i=i_k^{1000}+1}^{I_k} E_{ik}^s \\ &= \left(\frac{1}{I_k - i_k^{1000}}\right) \sum_{i=i_k^{1000}+1}^{I_k} \frac{A_{ik}^2}{2} T_k^s \\ &= \left(\frac{T_k^s}{I_k - i_k^{1000}}\right) \sum_{i=i_k^{1000}+1}^{I_k} \frac{A_{ik}^2}{2} \end{aligned} \tag{5.6}$$

$$do1000_k = \frac{\sum_{i=i_k^{1000}+1}^{I_k} \left(E_{ik}^s - \overline{E}_k\right)\left(F_{ik} - \overline{F}_k^{1000}\right)}{\sum_{i=i_k^{1000}+1}^{I_k} \left(F_{ik} - \overline{F}_k^{1000}\right)^2} \tag{5.7}$$

Donde:

- $\overline{F}_k^{1000}$: es la frecuencia media de los armónicos de la trama $k-$ésima dentro de la banda $f \in (1000Hz, 5000Hz]$.
- $\overline{E}_k$: es la energía media de los armónicos de la trama $k-$ésima dentro de la banda $f \in (1000Hz, 5000Hz]$.

Una vez calculada la pendiente espectral para cada trama del segmento ($do1000_k$), se realiza un promediado para las K tramas consideradas en dicho segmento. No obstante, a diferencia de los parámetros HNR, HammI y pe1000, este no depende de una relación de energías. Esto introduce una clara dependencia de la medida con la longitud del segmento de voz considerado (cuanto más largo sea el segmento, el parámetro do1000 dará valores mayores, por ser mayor el periodo de integración energética). Con el objetivo de independizar la medida de este parámetro de la longitud de las tramas y del segmento, en la ecuación de cálculo final se ha introducido el tiempo

total del segmento ($T_{seg} = \sum_{k=1}^{K} T_k^s$) como factor de normalización, y se ha realizado una aproximación con la finalidad de reducir el coste de cálculo:

$$
\begin{aligned}
do1000 &= \frac{1}{T_{seg}} \sum_{k=1}^{K} do1000_k = \frac{1}{T_{seg}} \sum_{k=1}^{K} \frac{\sum_{i=i_k^{1000}+1}^{I_k} \left(E_{ik}^s - \overline{E}_k\right) \left(F_{ik} - \overline{F}_k^{1000}\right)}{\sum_{i=i_k^{1000}+1}^{I_k} \left(F_{ik} - \overline{F}_k^{1000}\right)^2} \\
&= \frac{1}{T_{seg}} \sum_{k=1}^{K} \frac{\sum_{i=i_k^{1000}+1}^{I_k} \left(\frac{A_{ik}^2}{2} T_k^s - \left(\frac{T_k^s}{I_k - i_k^{1000}}\right) \sum_{i=i_k^{1000}+1}^{I_k} \frac{A_{ik}^2}{2}\right) \left(F_{ik} - \overline{F}_k^{1000}\right)}{\sum_{i=i_k^{1000}+1}^{I_k} \left(F_{ik} - \overline{F}_k^{1000}\right)^2} \\
&= \frac{1}{T_{seg}} \sum_{k=1}^{K} T_k^s \left(\frac{\sum_{i=i_k^{1000}+1}^{I_k} \left(\frac{A_{ik}^2}{2} - \left(\frac{1}{I_k - i_k^{1000}}\right) \sum_{i=i_k^{1000}+1}^{I_k} \frac{A_{ik}^2}{2}\right) \left(F_{ik} - \overline{F}_k^{1000}\right)}{\sum_{i=i_k^{1000}+1}^{I_k} \left(F_{ik} - \overline{F}_k^{1000}\right)^2} \right) \\
&\approx \frac{1}{2K} \sum_{k=1}^{K} \left(\frac{\sum_{i=i_k^{1000}+1}^{I_k} \left(A_{ik}^2 - \left(\frac{1}{I_k - i_k^{1000}}\right) \sum_{i=i_k^{1000}+1}^{I_k} A_{ik}^2\right) \left(F_{ik} - \overline{F}_k^{1000}\right)}{\sum_{i=i_k^{1000}+1}^{I_k} \left(F_{ik} - \overline{F}_k^{1000}\right)^2} \right) \qquad (5.8)
\end{aligned}
$$

Siendo ambas expresiones idénticas en el caso que las tramas de la parte armónica tengan longitud constante (sería el caso $T_k^s = T^s$).

5.2.2.7. *Hammarberg Index* (HammI)

A partir de la definición de la Sección 2.4.2.4, tanto para el análisis usando la herramienta Praat como en base al modelo HNM, se calcula el valor de HammI.

Primero, usando Praat no se disponía de una función propia que realizara el cálculo de forma directa, con lo que se hizo uso del análisis LTAS que implementa para la extracción de la densidad espectral de potencia. Con este análisis se dispone de la información necesaria para encontrar el máximo de energía en cada una de las bandas de frecuencia bajo estudio y su relación expresándola en dB.

Para el caso de trabajar con el modelo HNM, el valor del parámetro HammI se obtiene como el promedio de los valores de la relación entre las energías máximas

en las bandas de $f \in [0Hz, 2000Hz]$ y $f \in (2000Hz, 5000Hz]$, halladas para todas las tramas del segmento de voz analizado. El análisis de la energía máxima se ha realizado considerando la diferencia entre energías de cada componente armónica, que sería equivalente a haber usado un promediado con un filtro de ancho de banda igual a la mínima frecuencia de *pitch*.

$$\begin{aligned} HammI_k &= \frac{\text{máx}\left(E_{ik}^s \mid i \in \left[0, i_k^{2000}\right]\right)}{\text{máx}\left(E_{ik}^s \mid i \in \left[i_k^{2000}+1,\ I_k\right]\right)} \\ &= \frac{\text{máx}\left(\frac{A_{ik}^2}{2}T_k^s \;\middle|\; i \in \left[0, i_k^{2000}\right]\right)}{\text{máx}\left(\frac{A_{ik}^2}{2}T_k^s \;\middle|\; i \in \left[i_k^{2000}+1,\ I_k\right]\right)} \\ &= \frac{\text{máx}\left(A_{ik}^2 \mid i \in \left[0, i_k^{2000}\right]\right)}{\text{máx}\left(A_{ik}^2 \mid i \in \left[i_k^{2000}+1,\ I_k\right]\right)} \end{aligned} \tag{5.9}$$

$$HammI = 10\log\left(\frac{1}{K}\sum_{k=1}^{K} HammI_k\right) \qquad [dB] \tag{5.10}$$

5.2.2.8. *Relative Amount of Energy above 1000 Hz* (pe1000)

La medida del parámetro pe1000 basada en la herramienta Praat se calcula como el promedio de los valores de relación entre las energías acumuladas en las bandas de $f \in [0Hz, 1000Hz]$ y $f \in (1000Hz, 5000Hz]$.

Debido a que la herramienta Praat no disponía de un algoritmo específico para el cálculo de este parámetro de forma directa, lo que se hizo fue utilizar el análisis LTAS que esta implementa para la extracción de la densidad espectral de potencia. A partir de esta información se calculó la energía total para cada una de las bandas de frecuencia implicadas y su relación expresada en dB.

Usando el modelo HNM, el valor del parámetro pe1000 se calcula como el promedio de los valores de relación entre las energías acumuladas en las bandas de $f \in [0Hz, 1000Hz]$ y $f \in (1000Hz, 5000Hz]$, halladas para todas las tramas del segmento de voz analizado. La energía acumulada en una banda se calcula como la suma de las energías de los armónicos presentes en ella.

$$pe1000_k = \frac{\sum_{i=i_k^{1000}+1}^{I_k} E_{ik}^s}{\sum_{i=1}^{i_k^{1000}} E_{ik}^s} = \frac{\sum_{i=i_k^{1000}+1}^{I_k} \frac{A_{ik}^2}{2}T_k^s}{\sum_{i=1}^{i_k^{1000}} \frac{A_{ik}^2}{2}T_k^s} = \frac{\sum_{i=i_k^{1000}+1}^{I_k} A_{ik}^2}{\sum_{i=1}^{i_k^{1000}} A_{ik}^2} \tag{5.11}$$

$$pe1000 = 10\log\left(\frac{1}{K}\sum_{k=1}^{K} pe1000_k\right) \qquad [dB] \tag{5.12}$$

5.2.3. Modificación de los parámetros de cualidad de la voz

En esta sección se propone la modificación de los parámetros de VoQ con el objetivo de mejorar la SHE a partir del modelado de cada estilo de habla expresivo. Se hace uso de la modificación de los parámetros de VoQ definidos en la Sección 2.4.2.4 para transformar las características del habla, cambiando así su percepción. A diferencia de lo presentado para el análisis de la VoQ (Sección 5.2.2), la modificación se vincula únicamente al uso de la parametrización del habla basada en HNM, que permite la modificación de la mayor parte de los parámetros de VoQ de una forma flexible, manteniendo la calidad del habla transformada (en la Sección 5.5.4.2 se muestra la evaluación de los experimentos de transformación de estilos de habla expresivos usando la modificación presentada en esta sección).

Por tanto, esta sección se centra en la modificación de los parámetros de VoQ utilizando HNM, de la que se ha hecho uso durante los experimentos de transformación de estilos de habla expresivos llevados a cabo (Sección 5.5), dejando así una propuesta metodológica para trabajar en SHE. No todos los parámetros han sido modificados, ya que el GNE y SFM fueron descartados para las primeras tentativas de transformación debido a las siguientes razones:

- En primer lugar, el GNE tiene muchas similitudes con el HNR en cuanto a lo que representa, pero el HNR tiene como ventaja, al utilizar HNM, que hace uso de la flexibilidad que le aporta conocer la componente armónica y la de ruido de la señal de voz.

- En segundo lugar, analizando el papel del SFM en la caracterización del habla, se observó que otros parámetros como son el HammI y el pe1000 podían ocupar su lugar. A esta conclusión se llegó a partir de los resultados de discriminación (Sección 5.4, y en especial en la Sección 5.4.4.3) y debido a las limitaciones que se detectaron para su modificación debidas a la propia naturaleza del parámetro (la modificación simultánea de la media geométrica y aritmética de los valores de amplitud de los armónicos). Se trabajó en varias propuestas de modificación, como una modificación iterativa lineal y una más compleja basada en *Simulated Annealing* (SA)[2], sin conseguir resultados de calidad suficiente como para justificar su utilización, de modo que se decidió descartar este parámetro por el momento.

La modificación de los parámetros de VoQ para la transformación de estilos de habla expresivos durante la SHE (Secciones 5.2.3.1 a 5.2.3.6) ha sido diseñada para aceptar tanto un valor destino del parámetro como un factor de modificación lineal (β) (Apéndice B.4). Esta segunda alternativa presenta la ventaja de poder modificar una voz sin tener un modelo destino prefijado, pudiendo establecer ajustes manualmente y realizar el diseño de multitud de voces (*presets*), ya sea en aplicaciones de SHE o de transformación de locutor en general. Todos los parámetros, excepto el *jitter* y el

[2] *Simulated Annealing* o "recocido simulado" es un algoritmo de búsqueda meta-heurística para problemas de optimización global, es decir, encontrar una buena aproximación al óptimo global de una función en un gran espacio de búsqueda.

shimmer, utilizan internamente el factor de modificación β, con lo cual, en el caso de indicar un valor destino, el factor β equivalente se calcula a partir del proceso de análisis, dando lugar a un valor β_k que puede ser diferente para cada trama k. Para el caso del *jitter* y del *shimmer*, se trabaja con el valor de destino directamente por motivos de implementación (Sección 5.3.3), con lo que en el caso de indicarse un factor de modificación, este se convierte al valor destino a partir del proceso de análisis (Sección 5.2.2).

La energía global del habla resultante se ha mantenido invariante durante las modificaciones que han afectado a la energía de distintas bandas de frecuencia. Para ello, se utiliza el factor de corrección de energía α_k (Apéndice B.5) correspondiente a la trama $k-$ésima. Esta corrección en la energía se ha aplicado sobre los parámetros HNR, HammI y pe1000. En el caso del *jitter* este ajuste no tiene sentido porque este parámetro no implica la modificación de la energía de la señal, mientras que para el *shimmer* y el do1000 no se realiza ninguna corrección por ser precisamente la variación de la energía lo que les confiere sus características.

Debido a la definición de cada uno de los parámetros, hay que tener en cuenta que la modificación de un parámetro afecta a otros que modelan características compartidas, especialmente en el caso de todos los que implican modificaciones directamente sobre bandas espectrales. Por ejemplo, el caso de modificar el parámetro HammI manipulando las bandas por debajo y por encima de 1000 Hz puede provocar cambios en el parámetro pe1000, que trabaja en las bandas de frecuencia alrededor de 2000 Hz. Para minimizar el impacto que pudiera suceder entre ellos, y teniendo en cuenta qué modificación requiere cada uno, se propone el siguiente orden de modificación: (1) *Jitter*, (2) HNR, (3) do1000, (4) pe1000, (5) HammI, (6) *Shimmer*.

En las Secciones 5.2.3.1 a 5.2.3.6 se presenta el proceso de modificación de cada uno de los parámetros de VoQ, mientras que la notación que se ha seguido se muestra en el Apéndice B.3. Por último, los detalles de su utilización durante la transformación de estilos de habla expresivos para la SHE se amplían en la Sección 5.5.

5.2.3.1. *Jitter*

La modificación del *jitter* sigue la propuesta presentada en la Sección 5.3.3. Como se explicó para su análisis (Sección 5.2.2.1), esta metodología se adapta a las características del habla expresiva, aplicando la nueva propuesta a la SHE. Esta metodología es independiente del algoritmo empleado por el CTH (p. ej. PSOLA o HNM) por ser la única información que necesita la de F_0.

5.2.3.2. *Shimmer*

El caso del *shimmer*, tal y como pasa durante su análisis (Sección 5.2.2.1), sigue los mismos pasos del *jitter*. Para su modificación se propone la adaptación de la definición del parámetro *shimmer* presentada en la Sección 5.3.3, debido a que la nueva propuesta realizada tiene en cuenta que el ámbito de aplicación es el habla expresiva, siendo por tanto la metodología que mejor se adapta a la SHE. Esta metodología es independiente del algoritmo de síntesis empleado por el CTH, con la particularidad

de que, si por ejemplo se usa PSOLA, este parámetro se genera a partir de la información de la señal de voz y, en el caso del HNM, como es el de la presente sección, se utiliza únicamente la componente determinista y la información de *pitch* indicada por las frecuencias del modelo.

5.2.3.3. *Harmonic-to-Noise Ratio* (HNR)

Partiendo de las frecuencias y amplitudes de la componente determinista, de la varianza de la componente de ruido y del factor de modificación β, se modifican los parámetros del HNM (nuevas amplitudes A'_{ik} y nueva varianza de ruido $(\sigma_j^2)'$) que permiten la resíntesis de la nueva señal de voz (Ecuaciones 5.13 a 5.18).

$$E_k^s = \sum_{i=1}^{I_k} \frac{A_{ik}^2}{2} T_k^s = T_k^s \sum_{i=1}^{I_k} \frac{A_{ik}^2}{2} \tag{5.13}$$

$$\overline{E}^s = \frac{1}{T_{seg}} \sum_{k=1}^{K} E_k^s \approx \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{I_k} \frac{A_{ik}^2}{2} \tag{5.14}$$

$$\overline{E}^r = \frac{1}{T_{seg}} \sum_{j=1}^{J} E_j^r = \frac{1}{J} \sum_{j=1}^{J} \sigma_j^2 \tag{5.15}$$

$$\alpha = \frac{\overline{E}^s + \overline{E}^r}{\sqrt{\beta} \cdot \overline{E}^s + \frac{1}{\sqrt{\beta}} \cdot \overline{E}^r} \tag{5.16}$$

$$A'_{ik} = \sqrt{\alpha} \cdot \sqrt[4]{\beta} \cdot A_{ik} \tag{5.17}$$

$$(\sigma_j^2)' = \alpha \cdot \frac{1}{\sqrt{\beta}} \cdot \sigma_j^2 \tag{5.18}$$

Donde:

- $\overline{E}^s$: energía media de la componente determinista para el total de la señal de voz analizada.
- $\overline{E}^r$: energía media de la componente estocástica para el total de la señal de voz analizada.

5.2.3.4. *Drop-off of Spectral Energy above 1000 Hz* (do1000)

La modificación de este parámetro se muestra en las Ecuaciones 5.19 a 5.28. Para cada trama k, utilizando las frecuencias y amplitudes de la componente determinista, se obtienen las frecuencias y las energías por encima de 1000 Hz hasta el valor del mayor armónico situado en el modelo HNM en 5000 Hz. Con esto se calcula la aproximación por mínimos cuadrados de la pendiente espectral siguiendo el trabajo de Abdi (2003), dando lugar a las nuevas energías debido a la aproximación lineal ($\tilde{E}_k^{1000}$). El siguiente paso es calcular las nuevas energías, restando al original la aproximación lineal, valor sobre el que se suma de nuevo esta aproximación lineal donde su pendiente ha sido modificada según el factor β_k. Finalmente se calculan las nuevas

amplitudes A'_{ik}, donde únicamente se han visto modificados los valores por encima de 1000 Hz. En este caso, el factor de corrección de la energía α_k no se aplica, ya que el efecto que se busca es precisamente la percepción de la variación de la energía en la banda de interés. Debido a que la máxima frecuencia de la componente armónica es un parámetro del modelo, no tiene por qué coincidir con la mitad de la frecuencia de muestreo.

$$F_k^{1000} = F_{ik} \quad | \, i \in [i_k^{1000}+1, I_k] \tag{5.19}$$

$$\overline{F}_k^{1000} = \left(\frac{1}{I_k - i_k^{1000}}\right) \sum_{i=i_k^{1000}+1}^{I_k} F_{ik} \tag{5.20}$$

$$\overline{E}_k = \left(\frac{1}{I_k - i_k^{1000}}\right) \sum_{i=i_k^{1000}+1}^{I_k} E_{ik}^s = \left(\frac{T_k^s}{I_k - i_k^{1000}}\right) \sum_{i=i_k^{1000}+1}^{I_k} \frac{A_{ik}^2}{2} \tag{5.21}$$

$$\tilde{\overline{E}}_k = \frac{1}{T_{seg}} \cdot \overline{E}_k \approx \frac{1}{K(I_k - i_k^{1000})} \sum_{i=i_k^{1000}+1}^{I_k} \frac{A_{ik}^2}{2} \tag{5.22}$$

$$\tilde{E}_{ik}^s = \frac{1}{T_{seg}} \cdot E_{ik}^s = \frac{T_k^s \frac{A_{ik}^2}{2}}{T_{seg}} \approx \frac{A_{ik}^2}{2K} \tag{5.23}$$

$$\tilde{E}_k^{1000} = a_k + b_k \cdot F_{ik} \quad | \, i \in [i_k^{1000}+1, I_k] \tag{5.24}$$

$$b_k = \frac{\sum_{i=i_k^{1000}+1}^{I_k} \left(\tilde{E}_{ik}^s - \tilde{\overline{E}}_k\right) \cdot \left(F_{ik} - \overline{F}_k^{1000}\right)}{\sum_{i=i_k^{1000}+1}^{I_k} \left(F_{ik} - \overline{F}_k^{1000}\right)^2} \tag{5.25}$$

$$a_k = \tilde{\overline{E}}_k - b_k \cdot \overline{F}_k^{1000} \tag{5.26}$$

$$\left(\tilde{E}_{ik}^s\right)' = \tilde{E}_{ik}^s - \tilde{E}_k^{1000} + \left(a_k + \beta_k \cdot b_k \cdot F_k^{1000}\right) \quad | \, i \in [i_k^{1000}+1, I_k] \tag{5.27}$$

$$A'_{ik} = \begin{cases} A_{ik} & | \, i \in [1, i_k^{1000}] \\ \sqrt{2K\left(\tilde{E}_{ik}^s\right)'} & | \, i \in [i_k^{1000}+1, I_k] \end{cases} \tag{5.28}$$

Donde:

- F_k^{1000}: frecuencias en la banda $f \in (1000Hz, 5000Hz]$ de la trama $k-$ésima.
- $\overline{F}_k^{1000}$: es la frecuencia media de los armónicos de la trama $k-$ésima dentro de la banda $f \in (1000Hz, 5000Hz]$.
- $\overline{E}_k$: es la energía media de los armónicos de la trama $k-$ésima dentro de la banda $f \in (1000Hz, 5000Hz]$.
- $\tilde{\overline{E}}_k$: energía promediada $\overline{E}_k$ normalizada a la duración total del segmento.

- $\tilde{E}^s_{ik}$: energía E^s_{ik} normalizada a la duración total del segmento.
- $\tilde{E}^{1000}_k$: aproximación por mínimos cuadrados de la pendiente espectral en la banda $f \in (1000Hz, 5000Hz]$ de la trama $k-$ésima.
- a_k: ordenada en el origen de la aproximación lineal en la trama k.
- b_k: pendiente de la aproximación lineal en la trama k.
- $(\tilde{E}^s_{ik})'$: energías normalizadas resultantes de la modificación de $\tilde{E}^s_{ik}$.

5.2.3.5. *Hammarberg Index* (HammI)

Las modificaciones se hacen para cada trama k, calculando la energía en las bandas de frecuencia por debajo y por encima de 2000 Hz (hasta el máximo fijado en 5000 Hz), asegurando con el factor α_k que la energía total de la señal de voz se mantiene invariante después de la modificación y, obteniendo finalmente, las nuevas amplitudes A'_{ik} (Ecuaciones 5.29 a 5.30). El valor finalmente modificado es el correspondiente a la frecuencia de la máxima amplitud en cada una de las bandas (de baja y de alta frecuencia respectivamente) para cada trama k. La modificación se realiza de esta manera debido a que si fuera modificada toda la banda, en lugar de los valores puntuales, asegurando que el valor de frecuencia asociado al máximo de amplitud no pudiera cambiar, la modificación provocaría cambios más bruscos en parámetros como el pe1000, al depender este de la relación de energía en diferentes bandas.

$$\begin{aligned}
\alpha_k &= \frac{(E^s_k \mid i \in [1, i^{2000}_k]) + (E^s_k \mid i \in [i^{2000}_k + 1, I_k])}{\sqrt{\beta_k} \cdot (E^s_k \mid i \in [1, i^{2000}_k]) + \frac{1}{\sqrt{\beta_k}} \cdot (E^s_k \mid i \in [i^{2000}_k + 1, I_k])} \\
&= \frac{T^s_k \sum_{i=1}^{i^{2000}_k} \frac{A^2_{ik}}{2} + T^s_k \sum_{i=i^{2000}_k+1}^{I_k} \frac{A^2_{ik}}{2}}{\sqrt{\beta_k} \left(T^s_k \sum_{i=1}^{i^{2000}_k} \frac{A^2_{ik}}{2} \right) + \frac{1}{\sqrt{\beta_k}} \left(T^s_k \sum_{i=i^{2000}_k+1}^{I_k} \frac{A^2_{ik}}{2}) \right)} \\
&= \frac{\sum_{i=1}^{i^{2000}_k} \frac{A^2_{ik}}{2} + \sum_{i=i^{2000}_k+1}^{I_k} \frac{A^2_{ik}}{2}}{\sqrt{\beta_k} \left(\sum_{i=1}^{i^{2000}_k} \frac{A^2_{ik}}{2} \right) + \frac{1}{\sqrt{\beta_k}} \left(\sum_{i=i^{2000}_k+1}^{I_k} \frac{A^2_{ik}}{2}) \right)}
\end{aligned} \tag{5.29}$$

$$A'_{ik} = \begin{cases} \sqrt{\alpha_k} \cdot \sqrt[4]{\beta_k} \cdot A_{ik} & \mid i = \underset{i \in [1, i^{2000}_k]}{\arg\max} A_{ik} \\ \sqrt{\alpha_k} \cdot \frac{1}{\sqrt[4]{\beta_k}} \cdot A_{ik} & \mid i = \underset{i \in [i^{2000}_k+1, I_k]}{\arg\max} A_{ik} \end{cases} \tag{5.30}$$

5.2.3.6. *Relative Amount of Energy above 1000 Hz* (pe1000)

Para cada trama k, se calcula la energía total en las bandas de frecuencia por debajo y por encima de 1000 Hz con un límite superior dado por el modelo HNM de 5000 Hz, asegurando que la energía total de la señal de voz se mantiene invariante mediante el factor α_k y, obteniendo finalmente, las nuevas amplitudes A'_{ik} (Ecuaciones 5.31 y 5.32).

$$\begin{aligned}
\alpha_k &= \frac{(E_k^s \mid i \in [i_k^{1000}+1, I_k]) + (E_k^s \mid i \in [1, i_k^{1000}])}{\sqrt{\beta_k} \cdot (E_k^s \mid i \in [i_k^{1000}+1, I_k]) + \frac{1}{\sqrt{\beta_k}} \cdot (E_k^s \mid i \in [1, i_k^{1000}])} \\
&= \frac{T_k^s \sum_{i=i_k^{1000}+1}^{I_k} \frac{A_{ik}^2}{2} + T_k^s \sum_{i=1}^{i_k^{1000}} \frac{A_{ik}^2}{2}}{\sqrt{\beta_k}\left(T_k^s \sum_{i=i_k^{1000}+1}^{I_k} \frac{A_{ik}^2}{2}\right) + \frac{1}{\sqrt{\beta_k}}\left(T_k^s \sum_{i=1}^{i_k^{1000}} \frac{A_{ik}^2}{2})\right)} \\
&= \frac{\sum_{i=i_k^{1000}+1}^{I_k} \frac{A_{ik}^2}{2} + \sum_{i=1}^{i_k^{1000}} \frac{A_{ik}^2}{2}}{\sqrt{\beta_k}\left(\sum_{i=i_k^{1000}+1}^{I_k} \frac{A_{ik}^2}{2}\right) + \frac{1}{\sqrt{\beta_k}}\left(\sum_{i=1}^{i_k^{1000}} \frac{A_{ik}^2}{2})\right)}
\end{aligned} \tag{5.31}$$

$$A'_{ik} = \begin{cases} \sqrt{\alpha_k} \cdot \frac{1}{\sqrt[4]{\beta}} \cdot A_{ik} & \mid i \in [1, i_k^{1000}] \\ \sqrt{\alpha_k} \cdot \sqrt[4]{\beta_k} \cdot A_{ik} & \mid i \in [i_k^{1000}+1, I_k] \end{cases} \tag{5.32}$$

5.3. Propuesta metodológica para el *jitter* y el *shimmer*

5.3.1. Introducción

La metodología llevada a cabo por herramientas usadas en el análisis de la señal de voz, como Praat o MDVP, que implementan la definición de *jitter* y de *shimmer* presentada en la Sección 2.4.2.4, presenta el inconveniente de no tener en cuenta el estilo de habla expresivo con el que el interlocutor está hablando, asumiendo que parámetros como los de prosodia no afectan a su medida. En aplicaciones acerca del estudio de patologías de voz, el análisis de la VoQ se hace independiente de la prosodia (Núñez et al., 2004) debido a que no existe contenido expresivo en el mensaje oral. En cambio, en estudios relacionados con las tecnologías del habla, como los de Slyh et al. (1999) y Verma y Kumar (2005), se muestra como la medida de estos parámetros se realiza sin considerar la variación de la prosodia existente en el enunciado a causa del estilo expresivo transmitido, a pesar de que en trabajos como el de Swerts y Veldhuis (2001) ya se observa su efecto sobre la VoQ.

En el método aquí presentado, se plantea considerar el efecto de la prosodia para así tratar de cancelarlo, obteniendo una medida de la VoQ sin interferencias, para

aplicaciones tanto de reconocimiento de emociones como de síntesis del habla. Para ello, se ha trabajado en un nuevo procedimiento de análisis mediante la redefinición de los parámetros *jitter* y *shimmer* (Sección 5.3.2), manteniendo el objetivo último de ambos, es decir, extraer información de las variaciones del periodo fundamental y de las amplitudes pico a pico en periodos consecutivos de la señal de voz. Además de realizar la medida del *jitter* y del *shimmer* de una forma más fiable, se fija como objetivo presentar la metodología para la modificación de estos parámetros para ser utilizados en la SHE (Sección 5.3.3).

Conviene aclarar que las metodologías de análisis y de modificación presentadas son comunes para cualquier algoritmo de generación de habla utilizado por el sistema de CTH. La única diferencia que esto supone, y que no afecta a la metodología, es el origen de la información a utilizar. En el caso de usar PSOLA, la información acústica y de *pitch* que se emplea es la que se encuentra disponible en la base de datos del corpus, o bien la que se genere después del módulo de predicción prosódica. Por otro lado, si se usa parametrización del habla basada en HNM, la señal de voz y la información de *pitch* están respectivamente asociadas a las amplitudes y las frecuencias de la componente determinista del modelo (Apéndice B.1).

Junto con el nuevo procedimiento de análisis y de modificación, se realizó una prueba perceptiva (Sección 5.3.4) mediante un CMOS (ITU-P.800, 1996) sobre cuatro estilos de habla expresivos: agresivo, alegre, sensual y triste. Para llevar a cabo las pruebas de la nueva metodología, se ha utilizado el corpus de frases en castellano descrito en la Sección 4.3. La modificación de estos parámetros de VoQ se aplicó sobre muestras de habla sintetizada usando el sistema de CTH del grupo (Alías y Iriondo, 2002; Monzo et al., 2008a), utilizando el algoritmo PSOLA con modelado de la prosodia (Iriondo et al., 2007c).

Finalmente, antes de presentar las metodologías de análisis y de modificación para el *jitter* y el *shimmer*, veamos la nomenclatura común que siguen ambos parámetros:

- z: índice del tramo bajo análisis.
- Z: número total de tramos analizados.
- p: índice del periodo de *pitch* en el tramo analizado (toda la señal de habla o bien un tramo z concreto).
- P_z: valor máximo de p, correspondiendo al total de periodos de *pitch* en el tramo z analizado.
- t^p: instante (expresado como muestra) correspondiente a la marca de *pitch* p.
- $\overline{p}_z$: valor medio de los valores de p en el tramo z.
- a_z: ordenada en el origen de la aproximación lineal en el tramo z.
- b_z: pendiente de la aproximación lineal en el tramo z.

5.3.2. Metodología de análisis

En esta sección se expone el nuevo procedimiento para el análisis del *jitter* y del *shimmer*. Para cada uno de estos parámetros, se presenta su descripción y la propuesta realizada. El cálculo habitual de estos parámetros (Sección 2.4.2.4) no tiene en cuenta las variaciones de F_0 ni de energía, presentes en el habla y que en general son más importantes en el habla expresiva, ya que sus orígenes están ligados al análisis de la VoQ en el estudio de patologías de la voz, en el que tal y como se muestra en la Sección 2.2.4.2 la tendencia ha sido trabajar sobre vocales sostenidas (Lee et al., 2008).

5.3.2.1. *Jitter*

El nuevo procedimiento de análisis propuesto parte de la información de F_0, obtenida a partir de las marcas de *pitch* calculadas por un PMA o bien por la parametrización del habla usando HNM. A partir de la curva de F_0 se lleva a cabo una transformación logarítmica utilizando semitonos (Fant et al., 2002), consiguiéndose así una normalización relativa al tono medio y una mejor representación de la percepción subjetiva de las variaciones de tono. La transformación de hercios a semitonos, y su inversa, se muestra en las Ecuaciones 5.33 y 5.34 respectivamente, donde F_{ref} es la frecuencia de referencia correspondiente a la F_0 media del locutor para el estilo expresivo deseado. Durante los experimentos, esta frecuencia de referencia se calculó como la F_0 media del locutor para el correspondiente corpus de síntesis: neutro (188 Hz), alegre (294 Hz), triste (184 Hz), sensual (154 Hz) y agresivo (278 Hz).

$$Hz = 2^{st/12} \cdot F_{ref} \tag{5.33}$$

$$St = 12 \cdot [\ln(Hz/F_{ref})/\ln 2] \tag{5.34}$$

Por tanto, a partir de la curva de F_0 original, con un valor de F_0 para cada periodo de *pitch* (F_0^p), se obtiene la curva transformada a semitonos $\mathring{F}_0^p$ para cada una de las marcas de *pitch* p (Ecuación 5.35).

$$\mathring{F}_0^p = 12 \cdot \frac{\ln(F_0^p/F_{ref})}{\ln 2} \tag{5.35}$$

Una vez se dispone de los valores transformados en semitonos, se realiza una detección de los tramos de crecimiento y de decrecimiento de la curva $\mathring{F}_0^p$ a partir del análisis de la pendiente, tal y como se muestra a continuación:

1. Se realiza la derivada discreta de la curva de $\mathring{F}_0^p$ a partir de la resta de dos valores consecutivos de la curva indicados por p (Ecuación 5.36).

$$\Delta\mathring{F}_0^p = \mathring{F}_0^{p+1} - \mathring{F}_0^p \tag{5.36}$$

2. Se almacena el signo de cada uno de los valores de la derivada, asociando un -1 a los valores negativos y 1 al resto.

3. Se aplica un filtro de mediana[3] sobre los signos de la derivada, centrados en una ventana de longitud igual al valor indicado como parámetro de entrada a la función de análisis del parámetro. A la salida del filtro se dispone de la información de crecimiento o decrecimiento de la curva, evitando gracias al uso de la mediana la toma de decisiones locales por cambios rápidos debidos precisamente al parámetro de VoQ, ya que se tiene en cuenta el valor actual de los datos analizados y sus vecinos. El tamaño de este filtro de mediana fue fijado durante los experimentos a un valor igual a 5 muestras.

4. Finalmente, se realiza la decisión de tramos de crecimiento y decrecimiento de la curva de $\mathring{F}_0$. Esto se lleva a cabo a partir de los cambios de signo resultantes del filtrado, ya que la salida del filtro informa sobre el signo de la pendiente de la curva en cada momento, correspondiendo el instante de cambio de tendencia en la curva con el cambio de signo a la salida de este filtro.

Como se ha visto, a partir del análisis del signo de la pendiente se deciden los tramos donde la curva mantiene su tendencia creciente o decreciente, siendo a lo que nos referiremos como "tramo" (z). Por lo tanto, a cada uno de estos tramos se le asocia una curva de F_0 transformada a semitonos ($\mathring{F}^p_{0z}$), donde a cada frecuencia se le asigna un índice p correspondiente a un periodo de *pitch* en ese tramo. El siguiente paso es que para cada tramo se calcula la aproximación lineal (Abdi, 2003) de la curva de $\mathring{F}^p_{0z}$, representando la variación de la $\mathring{F}^p_{0z}$ debida a la prosodia, dando como resultado $\tilde{F}^p_{0z}$ (Ecuaciones 5.37 a 5.39). Este valor se resta de la curva inicial $\mathring{F}^p_{0z}$ con el fin de anular el efecto de la prosodia, obteniendo $\breve{F}^p_{0z}$ (Ecuación 5.42). A modo de ejemplo, este procedimiento se representa en la Figura 5.1.

$$\overline{p}_z = \frac{1}{P_z} \cdot \sum_{p=1}^{P_z} p \tag{5.37}$$

$$\overline{\mathring{F}}_{0z} = \frac{1}{P_z} \cdot \sum_{p=1}^{P_z} \mathring{F}^p_{0z} \tag{5.38}$$

$$\tilde{F}^p_{0z} = a_z + b_z \cdot p \tag{5.39}$$

$$b_z = \frac{\sum_{p=1}^{P_z} \left(\mathring{F}^p_{0z} - \overline{\mathring{F}}_{0z} \right) \cdot (p - \overline{p}_z)}{\sum_{p=1}^{P_z} (p - \overline{p}_z)^2} \tag{5.40}$$

$$a_z = \overline{\mathring{F}}_{0z} - b_z \cdot \overline{p}_z \tag{5.41}$$

$$\breve{F}^p_{0z} = \mathring{F}^p_{0z} - \tilde{F}^p_{0z} \tag{5.42}$$

Para terminar, eliminado el efecto de la prosodia sobre la curva $\mathring{F}^p_{0z}$ ($\breve{F}^p_{0z}$), la Ecuación 5.43 muestra el cálculo de la variación en periodos de *pitch* consecutivos

[3] El filtro de mediana realiza un filtrado por el cual, para cada valor de entrada al filtro, la salida corresponde a la mediana de los valores contenidos en este.

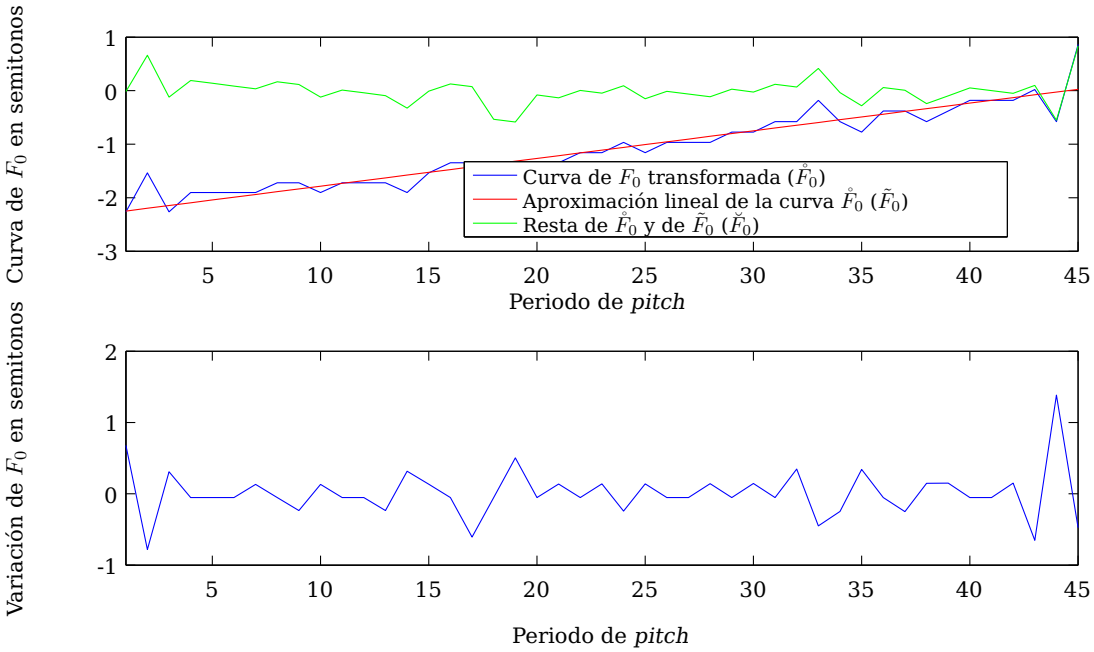


Figura 5.1: *Ejemplo de la extracción de la variabilidad de frecuencia fundamental*

de $\breve{F}^p_{0z}$, dando como resultado $\Delta\breve{F}^p_{0z}$. Finalmente, se obtiene el valor de *jitter* para cada uno de los tramos bajo análisis como la potencia de la señal $\Delta\breve{F}^p_{0z}$ (Ecuación 5.44). El valor de *jitter* final, para el total de la señal de voz analizada, se corresponde con el valor medio del *jitter* para cada uno de los tramos (Ecuación 5.45).

$$\Delta\breve{F}^p_{0z} = \breve{F}^{p+1}_{0z} - \breve{F}^p_{0z} \tag{5.43}$$

$$jitter_z = \frac{1}{P_z - 1} \cdot \sum_{p=1}^{P_z-1} \left(\Delta\breve{F}^p_{0z}\right)^2 \tag{5.44}$$

$$jitter = \frac{1}{Z} \cdot \sum_{z=1}^{Z} jitter_z \tag{5.45}$$

A continuación se presenta la nomenclatura que se ha seguido para el análisis del *jitter* (Ecuaciones 5.37 a 5.45), que complementa a la común mostrada en la Sección 5.3.1:

- $\mathring{F}^p_{0z}$: curva de F_0 transformada a semitonos para el tramo z y un valor para cada periodo de *pitch* p en ese tramo.
- $\overline{\mathring{F}}_{0z}$: valor medio de $\mathring{F}^p_{0z}$.
- $\tilde{F}^p_{0z}$: aproximación lineal de la curva $\mathring{F}^p_{0z}$.
- $\breve{F}^p_{0z}$: valores de frecuencia, después de anular el efecto de la prosodia, para cada tramo z.

- $\Delta\breve{F}_{0z}^{p}$: variaciones de $\mathring{F}_{0z}^{p}$, asociadas al *jitter*, entre periodos de *pitch* p consecutivos del tramo z.

- $jitter_z$: valor del *jitter* para cada uno de los tramos z.

5.3.2.2. *Shimmer*

El nuevo procedimiento propuesto para el *shimmer* está inspirado en el desarrollado para el *jitter* (Sección 5.3.2.1). Se parte de igual forma de la información de F_0 (y sus equivalentes marcas de *pitch*), con lo que para las muestras de la señal de voz ($x(n)$) contenidas en cada periodo de *pitch* (p), se calcula la curva de amplitudes pico a pico máximas (U) siguiendo la Ecuación 5.46. Por último, se lleva a cabo una transformación logarítmica aplicando el logaritmo natural (Ecuación 5.47) obteniendo $\mathring{U}$, ya que del mismo modo que para el caso de la F_0, queda mejor representada la percepción humana del sonido (Sección 3.2.2).

$$U^p = \left|\text{máx}(x(n) \mid n \in [t^p, t^{p+1}]) - \text{mín}(x(n) \mid n \in [t^p, t^{p+1}])\right| \quad (5.46)$$

$$\mathring{U}^p = \ln(U^p) \quad (5.47)$$

Una vez se dispone de los valores transformados, se elimina el efecto que tiene la prosodia sobre la energía. Para ello, se realiza una detección de los tramos de crecimiento y de decrecimiento de la curva $\mathring{U}^p$, repitiendo los pasos 1 al 4 seguidos durante el análisis del *jitter* (Sección 5.3.2.1), sustituyendo la curva $\mathring{F}_0^p$ por la $\mathring{U}^p$.

A partir del análisis del signo de la pendiente, se deciden los tramos donde la curva mantiene su tendencia creciente o decreciente, a lo que nos referiremos como "tramo" (z). A cada uno de ellos se le asigna una curva de amplitudes pico a pico transformada ($\mathring{U}_z^p$), donde a cada punto de la curva se le asocia un índice p correspondiente a un periodo de *pitch* en ese tramo. El siguiente paso es aplicar una aproximación lineal de la curva de amplitudes pico a pico ($\tilde{U}_z^p$) (Ecuaciones 5.48 a 5.50), que se resta de la curva de amplitudes inicial $\mathring{U}_z^p$ (Ecuación 5.53) con el fin de anular el efecto de la prosodia ($\breve{U}_z$). En la Figura 5.2 se ejemplifica gráficamente el proceso descrito.

$$\overline{p}_z = \frac{1}{P_z} \cdot \sum_{p=1}^{P_z} p \quad (5.48)$$

$$\overline{\mathring{U}}_z = \frac{1}{P_z} \cdot \sum_{p=1}^{P_z} \mathring{U}_z^p \quad (5.49)$$

$$\tilde{U}_z^p = a_z + b_z \cdot p \quad (5.50)$$

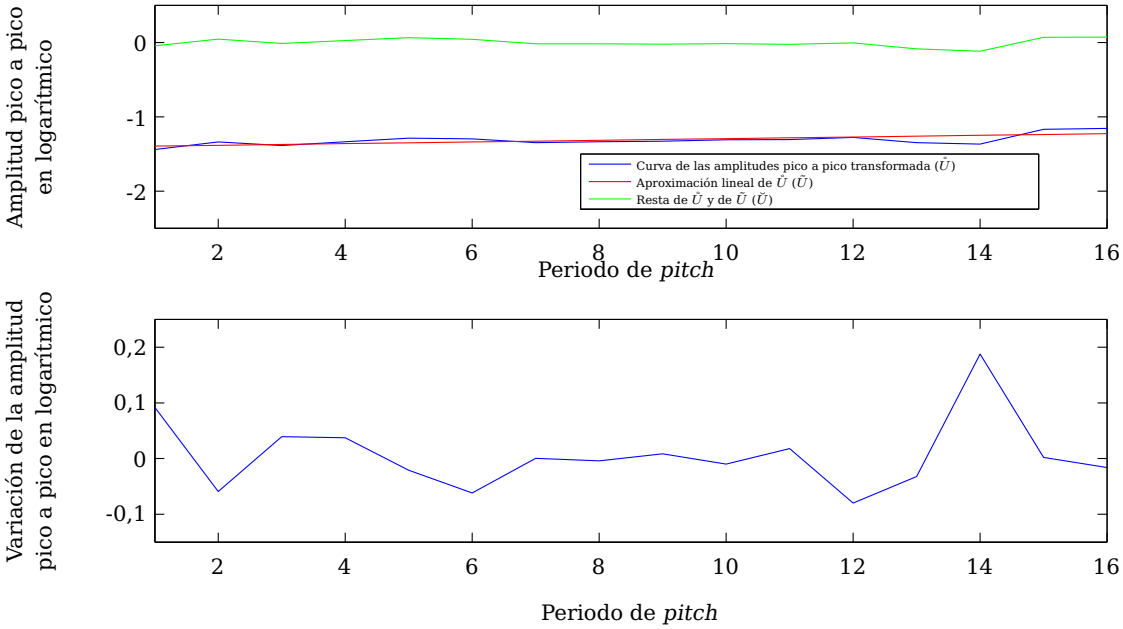


Figura 5.2: *Ejemplo de la extracción de la variabilidad de amplitud*

$$b_z = \frac{\sum_{p=1}^{P_z} \left(\mathring{U}_z^p - \overline{\mathring{U}}_z\right) \cdot (p - \overline{p}_z)}{\sum_{p=1}^{P_z} (p - \overline{p}_z)^2} \tag{5.51}$$

$$a_z = \overline{\mathring{U}}_z - b_z \cdot \overline{p}_z \tag{5.52}$$

$$\breve{U}_z^p = \mathring{U}_z^p - \tilde{U}_z^p \tag{5.53}$$

La Ecuación 5.54 muestra el cálculo de la variación en periodos de *pitch* consecutivos p de $\breve{U}_z^p$, dando como resultado $\Delta\breve{U}_z^p$. Por último, se obtiene el valor de *shimmer* para cada uno de los tramos bajo análisis como la potencia de la señal $\Delta\breve{U}_z^p$ (Ecuación 5.55). El valor de *shimmer* final, para el total de la señal de voz analizada, se corresponde con el valor medio del *shimmer* para cada uno de los tramos (Ecuación 5.56).

$$\Delta\breve{U}_z^p = \breve{U}_z^{p+1} - \breve{U}_z^p \tag{5.54}$$

$$shimmer_z = \frac{1}{P_z - 1} \cdot \sum_{p=1}^{P_z - 1} \left(\Delta\breve{U}_z^p\right)^2 \tag{5.55}$$

$$shimmer = \frac{1}{Z} \cdot \sum_{z=1}^{Z} shimmer_z \tag{5.56}$$

Por terminar, se muestra la nomenclatura propia del análisis del *shimmer* empleada en las Ecuaciones 5.48 a 5.56:

- $\mathring{U}_z^p$: curva de amplitudes pico a pico transformadas para el tramo z y un valor para cada periodo de *pitch* p en ese tramo.
- $\overline{\mathring{U}}_z$: valor medio de $\mathring{U}_z^p$.
- $\tilde{U}_z^p$: aproximación lineal de la curva $\mathring{U}_z^p$.
- $\breve{U}_z^p$: valores de amplitud pico a pico, después de anular el efecto de la prosodia, para cada tramo z.
- $\Delta\breve{U}_z^p$: variaciones de la amplitud pico a pico, asociadas al *shimmer*, entre periodos de *pitch* consecutivos p del tramo z.
- $shimmer_z$: valor del *shimmer* para cada uno de los tramos z.

5.3.3. Metodología de modificación

A partir de las metodologías de análisis para el *jitter* y el *shimmer*, se plantea el enfoque de cómo deben de ser modificados estos parámetros durante la síntesis del habla.

El primer paso para el *jitter* es el de transformar a semitonos la curva de F_0 usando el valor de referencia F_{ref} ($\mathring{F}_0$) y, para el *shimmer*, a logarítmico (logaritmo natural) la curva de amplitudes pico a pico de la señal de voz ($\mathring{U}$). Sobre estas curvas, tal y como se hace para el proceso de análisis (Sección 5.3.2), se realiza una detección de los tramos de crecimiento y de decrecimiento a partir del análisis la pendiente de sus derivadas, dando lugar a cada uno de los "tramos" z, que llevan asociadas las curvas $\mathring{F}_{0z}^p$ y $\mathring{U}_z^p$ respectivamente. Para cada tramo z, se realiza la aproximación lineal de las curvas $\mathring{F}_{0z}^p$ o $\mathring{U}_z^p$ según se trate del *jitter* o del *shimmer*, obteniendo respectivamente las curvas $\tilde{F}_{0z}^p$ y $\tilde{U}_z^p$, siendo considerado como el efecto de la prosodia.

El siguiente paso, que sigue el mismo procedimiento para ambos parámetros, se trata de la inserción de ruido blanco Gaussiano sobre $\tilde{F}_{0z}^p$ para el *jitter*, o sobre $\tilde{U}_z^p$ para el *shimmer*. La elección de ruido blanco Gaussiano se basa en trabajos previos como los de Cabral y Oliveira (2006) y Ruinskiy y Lavner (2008), que utilizan componentes aleatorias para la generación de estos parámetros. Asimismo, se realizó un estudio sobre la autocorrelación, la densidad espectral de potencia (PSD) y la distribución de los valores, para las variaciones de frecuencia y de amplitud una vez eliminado el efecto de la prosodia.

Este estudio se realizó sobre los valores analizados del estilo neutro, utilizando el enunciado con habla natural y este mismo donde se sintetizaron el *jitter* y el *shimmer*, por separado, con la metodología presentada en esta sección. Con estos dos casos, se pudieron comprobar las tres características bajo análisis.

El primero de los resultados obtenidos fue la no Gaussianidad de los datos, comprobada a partir del test de Lilliefors (1967), tanto sobre el caso natural como el sintetizado, hecho que no fue del todo sorprendente si se tiene en cuenta que los procesos fisiológicos tienen un margen de funcionamiento acotado (no se puede generar cualquier frecuencia ni cualquier amplitud), y el número de muestras podía

ser insuficiente durante la modificación como para tener tramos con una distribución realmente de ruido blanco Gaussiano.

A pesar de este primer resultado, la hipótesis de usar un ruido blanco Gaussiano se mantuvo, ya que a pesar de lo dicho, la mayor parte de los valores sí que situaban alrededor de un valor medio con una cierta dispersión, con lo cual, la comparativa se realizó respecto del histograma de los valores resultantes.

En la Figura 5.3 se muestran los resultados para los parámetros *jitter* (Figuras 5.3a a 5.3f) y *shimmer* (Figuras 5.3g a 5.3l), analizados sobre el enunciado originalmente neutro y sobre el mismo con ambos parámetros modificados, dando lugar a la posibilidad de comparación de la metodología de modificación. En cuanto a la autocorrelación, las líneas horizontales indican los límites por los cuales el margen de valores dentro de ellos se considera 0.

Para el caso del *jitter*, se comprueba que la autocorrelación (Figuras 5.3a y 5.3d) sigue un patrón similar, mostrando que los valores prácticamente no están correlacionados. Por su parte, la densidad espectral de potencia (Figuras 5.3b y 5.3e) también presenta características similares, y donde básicamente no existe una clara distribución de los valores de potencia o "coloreamiento". En todo caso, es en la modificación donde aparece un cierto "coloreamiento" en las bajas frecuencias, hecho que se puede asociar a una falta de muestras de ruido generadas, no siendo por tanto suficientes para generar el ruido blanco esperado. Por último, observando el histograma (Figuras 5.3c y 5.3f), se ve la agrupación de los valores entre -1 y 1 tanto para el análisis del enunciado natural como el del sintetizado, dando pie a que los valores de *jitter* se muevan en los mismos límites.

Por lo que hace referencia al *shimmer*, la autocorrelación (Figuras 5.3g y 5.3j), al igual que para el *jitter*, presenta un grado de incorrelación similar y dentro de los márgenes considerados 0. Las densidad espectral de potencia (Figuras 5.3h y 5.3k) se asemejan, no presentando un patrón claro de "coloreamiento" de los valores obtenidos, tendiendo por tanto a características de ruido blanco. Finalmente, en cuanto al histograma (Figuras 5.3i y 5.3l), presenta similitudes en la distribución para el enunciado natural y el sintetizado, dando a entender que los valores generados mediante el ruido blanco Gaussiano siguen una tendencia próxima a lo que ocurre con el caso del habla natural.

Con los resultados presentados, se considera que la modificación del *jitter* y del *shimmer*, usando ruido blanco Gaussiano de potencia el valor esperado del parámetro, queda justificada por obtener comportamientos similares a los del habla natural.

En cuanto a la cantidad de ruido que se introduce, este es de potencia igual al valor esperado de *jitter* y de *shimmer*, adecuado al estilo expresivo destino. Los cálculos se repiten para cada uno de los tramos, sumando el ruido generado a la aproximación lineal de la curva de F_0 o de amplitudes pico a pico transformadas. En cuanto a la asignación del valor de *jitter* y de *shimmer* a aplicar, esta se puede hacer desde dos enfoques, tal y como se muestra en la Sección 5.2.3:

1. Los valores destino pueden ser indicados tras el proceso de análisis de los estilos de habla expresivos de los que consta el corpus utilizado. Los valores destino

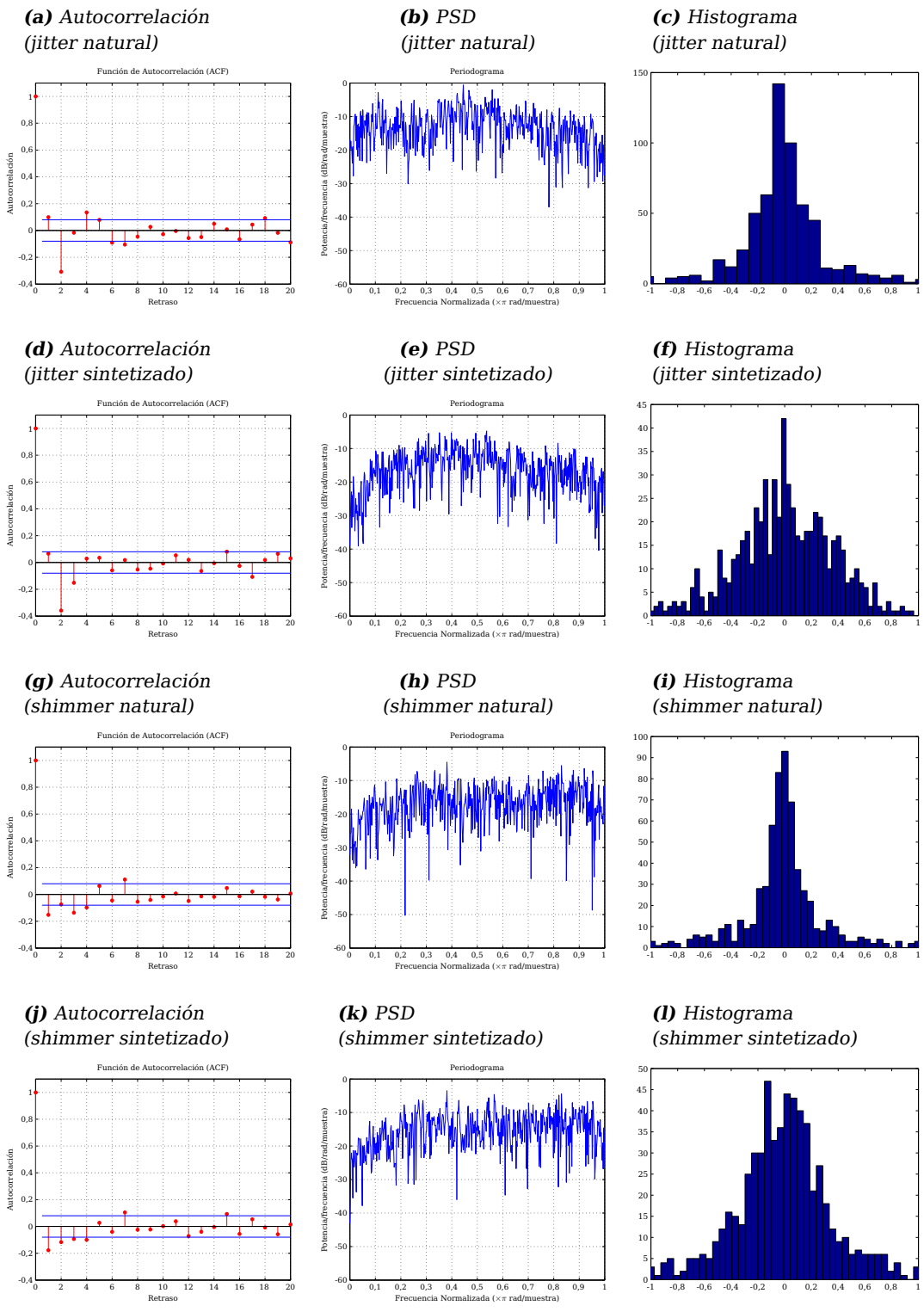


Figura 5.3: *Autocorrelación, densidad espectral de potencia e histograma de las variaciones de F_0 expresadas en semitonos y de las de amplitud pico a pico en logarítmico, durante el análisis del jitter y del shimmer para un enunciado natural y sintetizado con un estilo expresivo neutro.*

pueden ser definidos usando la metodología de análisis presentada en las Secciones 5.3.2.1 y 5.3.2.2 y a partir de la estadística descriptiva extraída sobre cada uno de ellos. Por ejemplo, para los experimentos aquí presentados, la Figura 5.4 muestra los valores de mediana extraídos del análisis, eligiendo en este caso la mediana para evitar valores atípicos (*outlier* en inglés) que puedan desviar el valor medio de la medida.

2. Puede indicarse un factor de modificación lineal del parámetro original (β). Indicarse un factor de modificación implica la realización de un proceso de análisis, para extraer el valor del parámetro origen sobre el que se va a aplicar la modificación. Como puede observarse, la utilización del factor de modificación β para los parámetros *jitter* y *shimmer* difiere respecto de cómo lo emplean el resto de parámetros de VoQ (Apéndice B.4).

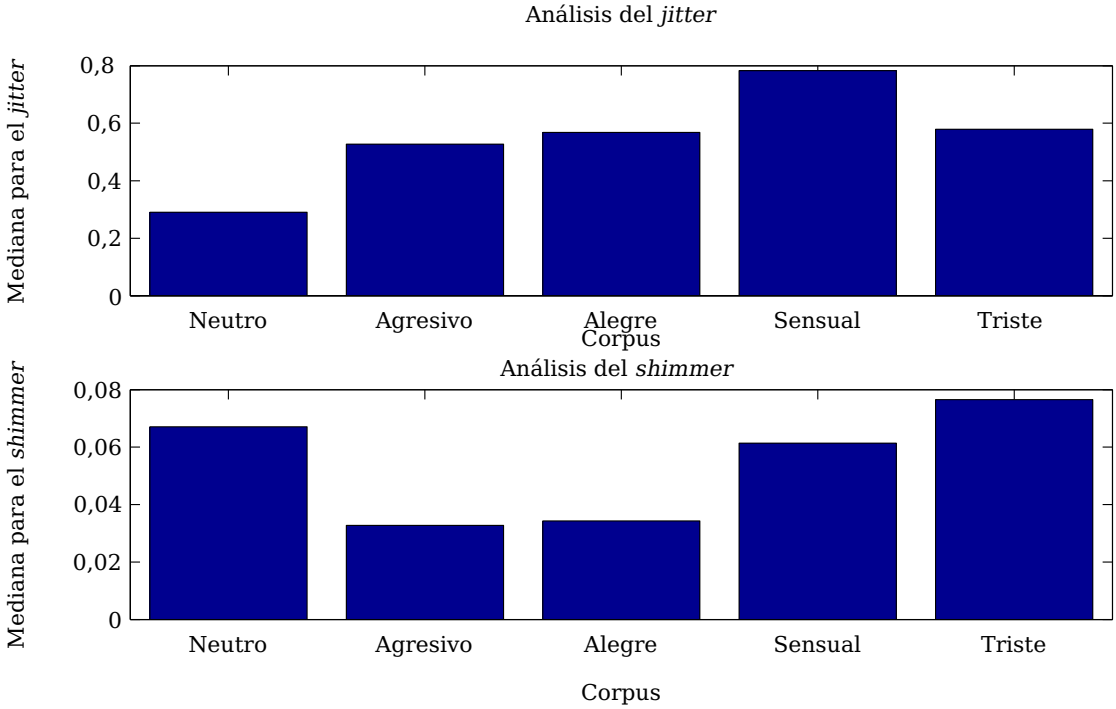


Figura 5.4: *Análisis del jitter y del shimmer sobre los cinco estilos de habla expresivos del corpus*

A continuación se presenta el desarrollo analítico para ambos parámetros y cada uno de los pasos a seguir. La nomenclatura utilizada, común tanto para el *jitter* como para el *shimmer* y para el proceso de análisis, se presenta en la Sección 5.3.1.

5.3.3.1. Jitter

Para el caso del *jitter*, del mismo modo que en el proceso de análisis (Sección 5.3.2.1), para cada tramo z se calcula la aproximación lineal de la curva de F_0 transformada a semitonos ($\mathring{F}^p_{0z}$), dando lugar a $\tilde{F}^p_{0z}$. Sobre esta aproximación lineal se

añade el ruido blanco Gaussiano (B_z^p) de potencia el valor esperado de *jitter* ($jitter_z$), obteniendo así la nueva curva de $\mathring{F}_{0z}^p$ ($(\mathring{F}_{0z}^p)'$). En el caso de que el modelo de VoQ indique el valor de *jitter* destino, este se corresponderá directamente con la variable $jitter_z$ para todos los tramos. Por el contrario, si se indica un factor de modificación lineal β, se analiza el tramo de interés obteniendo el valor de *jitter* original ($sjit_z$), con lo que aplicando la Ecuación 5.57 se calcula el valor destino correspondiente para ese tramo ($jitter_z$).

$$jitter_z = \beta \cdot sjit_z \tag{5.57}$$

Por último, se deshace la transformación de semitonos de la nueva curva $(\mathring{F}_{0z}^p)'$, obteniendo así $(F_{0z}^p)'$ (Ecuaciones 5.58 y 5.59). Este proceso se realiza para todos los tramos z, de forma que la nueva curva de F_0 se corresponde con la concatenación de cada una de las curvas $(F_{0z}^p)'$.

$$(\mathring{F}_{0z}^p)' = \tilde{F}_{0z}^p + \sqrt{jitter_z} \cdot B_z^p \tag{5.58}$$

$$(F_{0z}^p)' = 2^{((\mathring{F}_{0z}^p)'/12)} \cdot F_{ref} \tag{5.59}$$

5.3.3.2. Shimmer

En cuanto al *shimmer*, la modificación se realiza siguiendo el mismo procedimiento que para el proceso de análisis (Sección 5.3.2.1). En cada tramo z se calcula la aproximación lineal de la curva $\mathring{U}_z^p$ ($\tilde{U}_z^p$), sobre la que se añade el ruido blanco Gaussiano (B_z^p) de potencia el valor esperado de *shimmer* ($shimmer_z$), obteniendo así la nueva curva de $\mathring{U}_z^p$ ($(\mathring{U}_z^p)'$). Este valor destino puede venir dado por el modelo de VoQ, que se corresponde directamente con la variable $shimmer_z$ para todos los tramos. Por el contrario, si se indica un factor de modificación lineal β, se analiza el tramo de interés para obtener el valor de *shimmer* original ($sshim_z$), de modo que aplicando la Ecuación 5.60 se calcula el valor destino correspondiente para ese tramo ($shimmer_z$).

$$shimmer_z = \beta \cdot sshim_z \tag{5.60}$$

Como último paso queda deshacer la transformación logarítmica de la nueva curva $(\mathring{U}_z^p)'$, dando como resultado $(U_z^p)'$ (Ecuaciones 5.61 y 5.62). Este proceso se realiza para todos los tramos z, de forma que la nueva curva de amplitudes pico a pico se corresponde con la concatenación de cada una de las curvas $(U_z^p)'$.

$$(\mathring{U}_z^p)' = \tilde{U}_z^p + \sqrt{shimmer_z} \cdot B_z^p \tag{5.61}$$

$$(U_z^p)' = e^{(\mathring{U}_z^p)'} \tag{5.62}$$

La modificación del *jitter* y del *shimmer* trae consigo la respectiva modificación de la F_0 y de las amplitudes de la señal de voz originales. En el caso de emplear por ejemplo el algoritmo PSOLA para la generación del habla, el *jitter* provoca cambios en la posición de las marcas de *pitch*, mientras que si se emplea HNM, el cambio de F_0 quedaría reflejado directamente en las frecuencias del modelo. En el caso de las amplitudes, la modificación se realiza independientemente del algoritmo de generación

de habla, con lo que las muestras de la señal de voz origen $x(n)$, para cada tramo z y periodo de *pitch* p, se ven multiplicadas por un factor γ_z^p calculado a partir del resultado de la modificación del *shimmer* para cada tramo (Ecuación 5.63). No obstante, existe una pequeña diferencia en cuanto al origen de esta señal de voz, mientras que en el caso de PSOLA esta es directamente la señal de voz original, para el HNM se trata de la asociada a la componente determinista, evitando así la influencia del ruido debida a la componente estocástica.

$$\gamma_z^p = \frac{(U_z^p)'}{\text{máx}\,(x_z(n)) - \text{mín}\,(x_z(n))} \tag{5.63}$$

Este factor γ_z^p representa un valor por periodo de *pitch*, con lo que para modificar todas las muestras de la señal de voz se propone una interpolación lineal de cada uno de estos factores situados en los máximos de amplitud. A cada una de las muestras de la señal de voz en cada tramo z se le asigna un factor $(\gamma_z^n)'$ que la multiplica, dando como resultado la señal de voz destino $x'_z(n)$ (Ecuación 5.64).

$$x'_z(n) = (\gamma_z^n)' \cdot x_z(n) \tag{5.64}$$

Donde:

- n: índice de la muestra de la señal de voz.
- $x_z(n)$: información de la señal de voz original correspondiente al tramo z.
- $x'_z(n)$: nueva señal de voz generada, a partir del $shimmer$ indicado, asociada al tramo z.
- γ_z^p: factor de modificación de la amplitud, en periodos consecutivos de *pitch* p y por tramo z.
- $(\gamma_z^n)'$: interpolación lineal de γ_z^p, correspondiendo a un factor de modificación de la amplitud para cada muestra de la señal de voz en el tramo z.

5.3.4. Evaluación de la nueva metodología

Dado el nuevo procedimiento de análisis (Secciones 5.3.2.1 y 5.3.2.2) y de modificación (Sección 5.3.3) descrito para el *jitter* y el *shimmer*, se evaluó cómo pueden contribuir a la SHE, ya sea de forma independiente o bien conjunta.

La prueba realizada partió de 5 enunciados sintetizados, usando el sistema de CTH del grupo (Alías y Iriondo, 2002; Monzo et al., 2008a), con 4 estilos de habla expresivos diferentes: agresivo, alegre, sensual y triste. Durante la síntesis del habla se aplicaron modelos de prosodia predicha mediante Razonamiento Basado en Casos —*Case Based Reasoning*— (CBR) (Iriondo et al., 2007c), obteniendo el contorno de F_0, el contorno de energía y la duración segmental para cada fonema en el estilo de destino, sobre un enunciado originalmente neutro. Una vez generadas las frases con la predicción de prosodia, se les aplicó la modificación del *jitter* y del *shimmer* para su posterior evaluación.

La evaluación se llevó a cabo mediante una prueba perceptiva CMOS (ITU-P.800, 1996), usando la plataforma web TRUE (Planet et al., 2008), desarrollada con el propósito de evaluar sistemas multimedia. Los enunciados (Apéndice C.1) se mostraron en parejas, comparando la frase original sintetizada usando modelos prosódicos, con la modificada donde al modelo de prosodia se le sumó el *jitter*, el *shimmer*, o bien ambos parámetros al mismo tiempo, dando lugar a 60 comparaciones (5 enunciados, generados con 4 estilos de habla expresivos y usando 3 configuraciones). Con esto se pudo comparar el efecto de añadir los parámetros de VoQ sobre la prosodia.

Los voluntarios, 13 en total, fueron 10 hombres ($77\,\%$) y 3 mujeres ($23\,\%$) de edades comprendidas entre los 24 y los 48 años, combinando expertos ($54\,\%$) y no expertos ($46\,\%$) en tecnologías del habla. A cada uno de ellos se le preguntó si la intensidad de la expresividad de la señal de voz presentada era "mucho más", "más", "poco más" o "igual" que la del otro, usando una puntuación de: 3, 2, 1, 0, -1, -2 y -3.

Una vez llevadas a cabo las pruebas, se realizó su análisis utilizando los *boxplots* —Apéndice F.4— de la Figura 5.5, en la que los valores positivos se reservaron para aquellos casos donde, por usar VoQ, la expresividad se mostraba con mayor intensidad, mientras que los negativos indicaban que era la original con solo modelado prosódico.

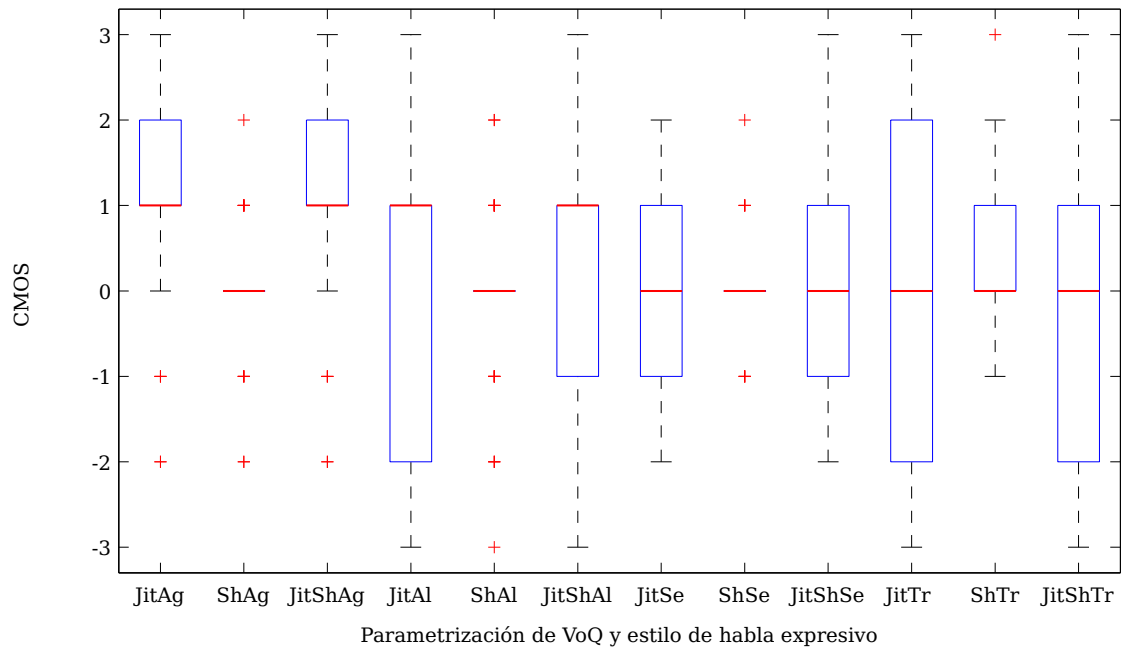


Figura 5.5: *Resultados de la prueba CMOS sobre el jitter (Jit) y el shimmer (Sh) usando cuatro estilos de habla expresivos: agresivo (Ag), alegre (Al), sensual (Se) y triste (Tr). Por ejemplo, JitShSe indica que en la transformación del estilo neutro al sensual se vieron modificados el jitter y el shimmer*

Por otro lado, el valor de CMOS medido, calculado como su valor medio, se presenta en la Tabla 5.1, indicando en *cursiva* aquellos estilos de habla expresivos donde la VoQ (*jitter* y *shimmer*) mejoró la percepción del estilo expresivo, y en **negrita** el valor donde esta mejora fue mayor.

	Agresivo	**Alegre**	**Sensual**	**Triste**
Jitter	*1, 06*	0, 00	−0, 12	−0, 06
Shimmer	*0, 12*	−0, 06	*0, 08*	*0, 29*
Jitter + Shimmer	***1, 14***	*0, 18*	−0, 03	−0, 31

Tabla 5.1: *CMOS para 3 configuraciones y 4 estilos de habla expresivos (en cursiva aquellos estilos de habla expresivos donde el jitter y el shimmer mejoran la percepción de la expresividad y en* ***negrita*** *el valor donde esta mejora es mayor)*

Como se puede observar en la Figura 5.5, el estilo expresivo para el que el efecto de la VoQ intensificó en mayor grado su percepción fue el agresivo, con un valor superior a 1 en la escala CMOS (Tabla 5.1). Además, el alegre presenta buenos resultados, mediana igual a 1, a pesar de su mayor dispersión. Para el resto de estilos expresivos, sensual y triste, se obtienen resultados con una elevada dispersión, hecho que hace pensar en su utilidad solamente en ciertos casos. Estos resultados son interesantes en tanto que agresivo y alegre daban los peores resultados en estudios donde únicamente se utilizaba prosodia (Iriondo, 2008). Por tanto, los parámetros *jitter* y *shimmer*, analizados y sintetizados con la metodología propuesta, complementan a la prosodia a la hora de generar estos estilos de habla expresivos durante la SHE.

Finalmente, se puede deducir de los resultados obtenidos (Figura 5.5 y Tabla 5.1) que con la modificación del *jitter* y del *shimmer* se ha conseguido incrementar la intensidad con la que se percibieron los estilos expresivos evaluados.

5.4. Capacidad discriminatoria de la cualidad de la voz

En esta sección se presenta el estudio que ha permitido conocer la capacidad de los parámetros de VoQ escogidos (Sección 5.2.1) para discriminar estilos de habla expresivos. Esta información es de utilidad para cualquier aplicación relacionada con el habla expresiva, ya que conocer su comportamiento es fundamental ya sea en aplicaciones de reconocimiento de emociones, de SHE o de medida objetiva de su calidad.

Los experimentos presentados en esta sección han sido llevados a cabo con dos objetivos principales:

- Conocer si los parámetros de VoQ permiten discriminar entre estilos de habla expresivos.
- Averiguar qué configuración de parámetros identifica mejor a cada uno de los estilos de habla expresivos, pudiendo utilizar esta información en reconocimiento de emociones y en la SHE. Para el foco de interés de esta tesis, la de la SHE

realizada por el sistema de CTH, de estos experimentos se pueden extraer las reglas de transformación entre estilos de habla expresivos.

5.4.1. Discriminación de estilos de habla expresivos

El primer paso fue averiguar si los parámetros de VoQ podrían ser de utilidad en el reconocimiento de emociones y para nuestro objetivo último de SHE. Ya se conocía la utilidad de la VoQ en la caracterización del contenido emocional o expresivo del habla (Sección 3.2.2), pero se decidió comprobar el comportamiento de los parámetros propuestos juntamente con los estilos de habla expresivos de los que se disponía, para poder afrontar nuevos experimentos con mayor seguridad.

En este experimento acerca de la capacidad discriminatoria de la VoQ, se partió del corpus en castellano descrito en la Sección 4.3, del que se hizo uso únicamente de la lista de palabras portadoras, para así tener realizaciones de cinco estilos expresivos alineados (cada uno de los enunciados del corpus estaba representado en cada uno de los estilos de habla expresivos), pudiendo realizar comparaciones sin que hubiera ningún tipo de sesgo hacia ninguno de ellos. Recordemos que este corpus consta de cinco estilos de habla expresivos: neutro (NEU), alegre (ALE), triste (TRI), sensual (SEN) y agresivo (AGR). A partir de su análisis se puede describir cada uno de los estilos expresivos respecto del estilo neutro, según la velocidad del habla y tipo de fonación (Tabla 5.2), pudiendo ser relacionada la parametrización de VoQ con estas características. De todas las unidades disponibles nos centramos únicamente en las vocales (Drioli et al., 2003; Keller, 2005): /a/, /e/, /i/, /o/ y /u/; ya que estas poseen zonas sonoras estables en las que se obtiene el cálculo robusto de los parámetros de VoQ.

Estilo	Velocidad del habla	Tipo de fonación
Neutro	- -	*Modal*
Alegre	Lenta	*Harsh* medio
Triste	Lenta	*Whispery* medio
Sensual	Muy lenta	*Whispery*
Agresivo	Lenta	*Harsh*

Tabla 5.2: *Velocidad del habla y tipo de fonación empleada en cada uno de los cinco estilos de habla expresivos del corpus*

El cálculo de los parámetros se realizó mediante la herramienta Praat, considerando toda la vocal como la ventana de información sobre la que realizar la medida. La distribución de los valores para cada uno de los parámetros se presenta en la Figura 5.6, y la metodología que se siguió para su cálculo en la Sección 5.2.2.

El análisis de los datos, para la discriminación de los diferentes estilos expresivos, se llevó a cabo en tres partes (Monzo et al., 2007):

1. A partir de estadística descriptiva se analizaron las distribuciones de los parámetros sobre los estilos expresivos (Sección 5.4.1.1).

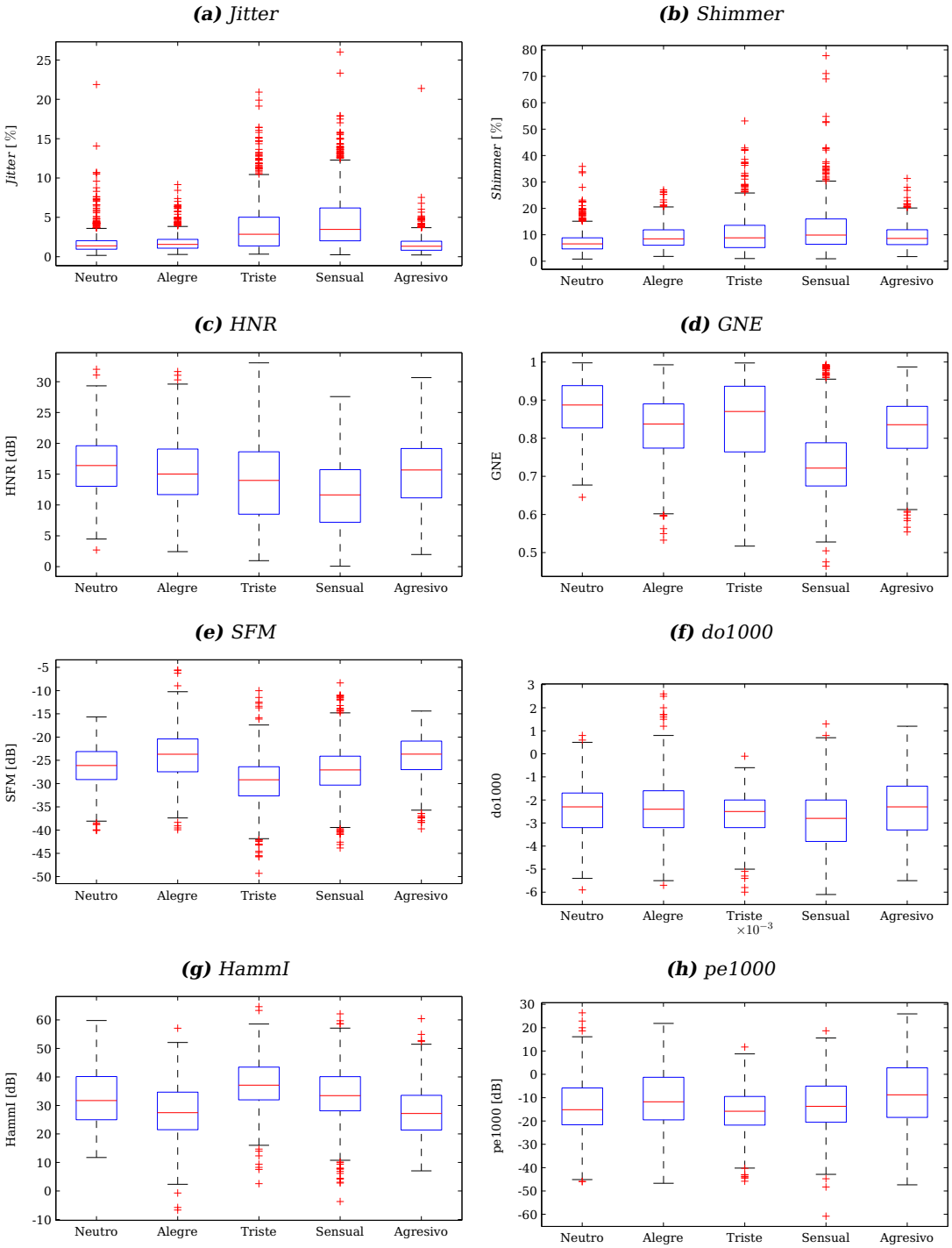


Figura 5.6: *Distribución de los valores de los parámetros de cualidad de la voz resultantes del análisis de las vocales del corpus de palabras en castellano*

2. Comprobadas mediante un estudio estadístico las posibilidades de discriminación, se pasó a aplicar un clasificador automático (Análisis Discriminante Lineal —*Linear Discriminant Analysis*— (LDA) (Sección 5.4.1.2)).

3. Finalmente, se validaron estadísticamente los resultados de la clasificación por medio de un $t-$test (Sección 5.4.1.3).

5.4.1.1. Estadística descriptiva sobre las distribuciones

Obtenidos los parámetros de VoQ, mediante estadística descriptiva (Apéndice F.3) se extrajeron estadísticas que representaran las distribuciones de valores de los mismos, pudiendo ver las posibilidades de discriminación de estilos expresivos para cada uno de ellos.

La Figura 5.7 muestra el valor absoluto de las medias y desviaciones estándar normalizadas de los parámetros de VoQ, mostrando las distribuciones de este modo para facilitar la presentación de los resultados. Como puede observarse, los parámetros de VoQ presentan distribuciones particulares para cada estilo expresivo, con lo que aquellas más separadas entre sí son a priori buenas candidatas a ser discriminadas, mientras que la elevada dispersión de algunos parámetros podría ser un problema para la discriminación, por estar excesivamente solapados los valores para diferentes estilos expresivos. Llegados a este punto, vistas las posibilidades y posibles limitaciones de la discriminación, se aplicó un clasificador automático que ayudara en este proceso de discriminación (Sección 5.4.1.2).

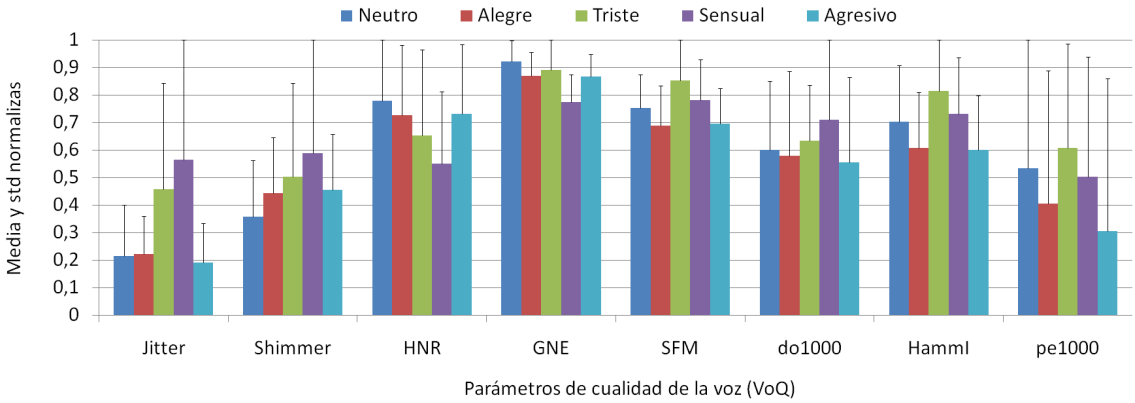


Figura 5.7: *Valor absoluto normalizado de la media y de la desviación estándar de los parámetros de cualidad de la voz*

5.4.1.2. Clasificación automática

Analizadas las estadísticas extraídas de los parámetros de VoQ (Sección 5.4.1.1), se pasó a aplicar una técnica estadística para la clasificación automática de los datos. Esta técnica es el LDA (Apéndice F.8), capaz de clasificar objetos en grupos excluyentes y exhaustivos basados en un conjunto mensurable de sus características. Las principales razones por las que se utilizó LDA son:

1. Ha sido una técnica ampliamente utilizada en estudios que tratan de la parametrización de la VoQ (Drioli et al., 2003; Lugger y Yang, 2006a).

2. No hay necesidad de una larga fase de entrenamiento (Lugger y Yang, 2006a), haciendo el proceso más sencillo.

El método de clasificación no era lo más importante en este experimento, ni tan solo el valor absoluto de clasificación, sino extraer relaciones entre parámetros y estilos expresivos y el conocimiento de cuál modela mejor a estos. De esta forma se podrían haber utilizado otras metodologías de clasificación, como por ejemplo el clasificador Bayesiano usado por Lugger y Yang (2007), o el amplio abanico de clasificadores empleados en los experimentos de clasificación de Iriondo et al. (2009).

El entrenamiento del clasificador se llevó a cabo mediante un proceso de entrenamiento supervisado. La información de entrada al clasificador fue cada uno de los parámetros de VoQ (*jitter*, *shimmer*, HNR, GNE, SFM, do1000, HammI y pe1000), calculados sobre las vocales de cada uno de los estilos de habla expresivos de los que consta el corpus (neutro, alegre, sensual, agresivo y triste), siendo cada uno de estos estilos la información asociada a la clase de salida del clasificador. Se realizó un $10-$*fold cross validation*[4] como proceso de validación, usando como medida de evaluación del comportamiento del clasificador la medida $F1$ (Sebastiani, 2002), ya que combina tanto la precisión como la cobertura en un sola medida, favoreciendo una actuación equilibrada de ambas (Apéndice F.7).

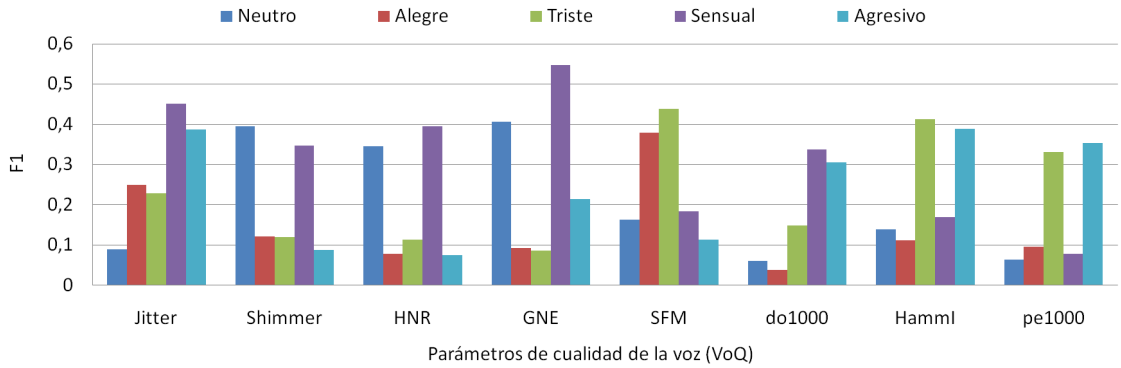


Figura 5.8: *Medida $F1$ para la clasificación de estilos de habla expresivos usando parámetros de cualidad de la voz*

Después del entrenamiento del clasificador se procedió a la realización de los experimentos. En la Figura 5.8, resumida en la Tabla 5.3, se muestra en términos de medida $F1$ el comportamiento que presenta el LDA para la discriminación de estilos de habla expresivos mediante la parametrización de la VoQ. Teniendo en cuenta que solamente se usó un parámetro cada vez en la discriminación de los cinco estilos de

[4] Proceso por el cual el conjunto de datos se parte en 10 subconjuntos de tamaño una décima parte del total, escogiendo 9 de estos subconjuntos como datos de entrenamiento y dejando el restante como de test, repitiendo este proceso 10 veces para tomar como resultado final el valor promedio de la medida realizada.

	Neutro	**Alegre**	**Triste**	**Sensual**	**Agresivo**	***F*1 *media***
Jitter	$0,09$	$0,25$	$0,23$	$0,45$	$\mathbf{0,39}$	$\boldsymbol{0,28}$
Shimmer	$0,40$	$0,12$	$0,12$	$0,35$	$0,09$	$0,22$
HNR	$0,35$	$0,08$	$0,11$	$0,40$	$0,07$	$0,20$
GNE	$\mathbf{0,41}$	$0,09$	$0,09$	$\mathbf{0,55}$	$0,21$	$0,27$
SFM	$0,16$	$\mathbf{0,38}$	$\mathbf{0,44}$	$0,18$	$0,11$	$0,25$
do1000	$0,06$	$0,04$	$0,15$	$0,34$	$0,30$	$0,18$
HammI	$0,14$	$0,11$	$0,41$	$0,17$	$\mathbf{0,39}$	$0,24$
pe1000	$0,06$	$0,10$	$0,33$	$0,08$	$0,35$	$0,18$

Tabla 5.3: *Medida de $F1$ para cada parámetro de cualidad de la voz y estilo de habla expresivo (en **negrita** se indica el valor máximo por columna)*

habla expresivos, se consideró que valores de $F1$ por encima de $0,3$ eran representativos de la capacidad discriminatoria del parámetro, resumiéndose en la Tabla 5.4 los parámetros que mejor discriminan a los estilos de habla expresivos.

Parámetro de VoQ	**Estilo de habla expresivo**
Jitter	SEN, AGR
Shimmer	NEU, SEN
HNR	NEU, SEN
GNE	NEU, SEN
SFM	ALE, TRI
do1000	SEN, AGR
HammI	TRI, AGR
pe1000	TRI, AGR

Tabla 5.4: *Parámetro de cualidad de la voz y estilos de habla expresivos mejor discriminados*

Con el objetivo de completar la Tabla 5.4, en la Tabla 5.5 se muestran los parámetros de VoQ más relevantes por estilo de habla expresivo. Por una parte, nótese como todos los parámetros son importantes en la discriminación de estilos de habla expresivos. Los resultados presentados en la Figura 5.8 y Tabla 5.4 muestran como sensual y agresivo son los estilos mejor discriminados. El estilo sensual, complicado de identificar mediante únicamente el uso de parámetros de prosodia (Iriondo et al., 2007a, 2009), puede ser en cambio claramente discriminado usando sobre todo los parámetros *jitter* y GNE. Por otro lado, el estilo alegre es el más difícil de caracterizar usando exclusivamente VoQ, ya que como se observa, solamente con el parámetro SFM se presenta una medida $F1$ por encima del valor umbral elegido.

Asimismo, los resultados obtenidos también pueden ser interpretados desde el punto de vista del tipo de fonación (Sección 2.1.3). Por ejemplo, el estilo sensual se caracteriza por una fonación de tipo *whispery*, mientras que el neutro se caracteriza por ser *modal*, por lo tanto, un parámetro de VoQ como es el GNE, relacionado con la medida de ruido, hace posible la discriminación entre ellos.

Estilo de habla expresivo	Parámetros de VoQ
Neutro	*Shimmer*, HNR, GNE
Alegre	SFM
Triste	SFM, HammI, pe1000
Sensual	*Jitter*, *Shimmer*, HNR, GNE, do1000
Agresivo	*Jitter*, do1000, HammI, pe1000

Tabla 5.5: *Relaciones más relevantes entre estilos de habla expresivos y parámetros de cualidad de la voz*

5.4.1.3. Validación estadística de los resultados de la clasificación

Como tercer y último paso realizado en el proceso de análisis de la capacidad discriminatoria de los parámetros de VoQ, se presenta el proceso de validación estadística de los resultados de la clasificación.

A partir de los resultados obtenidos del experimento de discriminación mediante el uso del LDA (Sección 5.4.1.2), se consideró necesaria la validación estadística de los resultados obtenidos. Esta validación se trataba de analizar como las distribuciones del parámetro de interés, para cada una de las parejas de estilos de habla expresivos bajo estudio, eran significativamente diferentes, permitiendo así su discriminación. Para ello se aplicó el $t-$test (Apéndice F.5) sobre la distribución de los valores de cada uno de los parámetros (Figura 5.7), para cada una de las parejas de estilos expresivos (Tabla 5.6). Con esta prueba podemos analizar si las medias de las distribuciones de un parámetro, para cada uno de los dos estilos de habla expresivos bajo estudio, son estadísticamente diferentes y, de este modo, conocer si la capacidad discriminatoria resultante de aplicar un clasificador automático como ha sido el LDA no es fruto de la casualidad y por tanto queda validada.

De esta prueba se obtuvo un valor de significación (p), de tal forma que aquellas parejas con un valor de p inferior al umbral de significación de $p < 0,05$, fueron consideradas significativamente diferentes y, por tanto, su discriminación posible.

	NEU ALE	**NEU TRI**	**NEU SEN**	**NEU AGR**	**ALE TRI**	**ALE SEN**	**ALE AGR**	**TRI SEN**	**TRI AGR**	**SEN AGR**
Jitter	$0,25$	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$
Shimmer	$*$	$*$	$*$	$*$	$*$	$*$	$0,31$	$*$	$*$	$*$
HNR	$*$	$*$	$*$	$*$	$*$	$*$	$0,63$	$*$	$*$	$*$
GNE	$*$	$*$	$*$	$*$	$*$	$*$	$0,44$	$*$	$*$	$*$
SFM	$*$	$*$	$*$	$*$	$*$	$*$	$0,31$	$*$	$*$	$*$
do1000	$0,09$	$*$	$*$	$*$	$*$	$*$	$0,08$	$*$	$*$	$*$
HammI	$*$	$*$	$*$	$*$	$*$	$*$	$0,45$	$*$	$*$	$*$
pe1000	$*$	$*$	$0,11$	$*$	$*$	$*$	$*$	$*$	$*$	$*$

Tabla 5.6: *Nivel de significación p, para la comparación de parejas de estilos de habla expresivos, por parámetro de cualidad de la voz ("$*$" para niveles por debajo del umbral de $p < 0,05$)*

Fijémonos como los resultados de discriminación más claros, obtenidos del análisis del LDA (Tabla 5.5), están por debajo del umbral de significación. Por lo tanto, la discriminación entre estilos de habla expresivos, usando LDA y parámetros de VoQ, queda validada. Sin embargo, existe una clara excepción entre los estilos alegre y agresivo, donde los niveles de significación para los parámetros SFM, do1000 y HammI, que habían sido considerados como discriminatorios de estos estilos, indican que no puede asegurarse su utilidad en la discriminación. Para los valores de *jitter* y do1000 entre los estilos neutro y alegre y de pe1000 entre neutro y sensual, los niveles de significación tampoco aseguran que puedan ser usados en la discriminación, tal y como ya se vio en la Figura 5.8, donde estos parámetros no daban niveles de $F1$ suficientes como para ser considerados de utilidad en la discriminación de esos estilos.

5.4.2. Parámetros y modelado de estilos de habla expresivos

En esta sección se muestra el trabajo realizado con el fin de extraer aquella configuración de parámetros que mejor modela a cada uno de los estilos expresivos bajo análisis: neutro (NEU), alegre (ALE), triste (TRI), sensual (SEN) y agresivo (AGR). Este mejor modelado de la VoQ significa buscar las mejores combinaciones de parámetros que maximicen la discriminación de estilos de habla expresivos, siendo de utilidad para el reconocimiento de emociones y la SHE. La información de partida utilizada fue la misma que se empleó en la Sección 5.4.1, un corpus alineado de palabras en castellano con 5 estilos de habla expresivos, sobre cuyas vocales se realizó la parametrización de la VoQ. Para hallar estas configuraciones se hizo una búsqueda exhaustiva, utilizando LDA como estrategia de clasificación, comparando los resultados obtenidos con los de referencia, presentados en la Sección 5.4.1.2.

En este experimento, a diferencia del presentado en la Sección 5.4.1.2, donde únicamente se utilizó un parámetro de VoQ para la discriminación de los 5 estilos de habla expresivos, se vieron implicados todos los parámetros de VoQ (Sección 5.2.1) como entrada al clasificador LDA, siendo todos los estilos de habla expresivos las clases de salida conocidas. Cada uno de los parámetros se calculó sobre las vocales del corpus (entrada al clasificador), conociendo el estilo al que este valor correspondía (salida del clasificador). La información de entrenamiento y de test se obtuvo mediante un $10-$*fold cross validation*, siendo la medida $F1$ la utilizada para la evaluación del comportamiento del clasificador. En este experimento, por tanto, se buscó extraer la combinación de parámetros de VoQ que maximizara la medida $F1$, demostrando como la correcta combinación de parámetros aumenta la capacidad discriminatoria de la VoQ, dejando constancia de aquellos parámetros que pueden ser usados en reconocimiento de emociones y descubriendo las reglas para el modelado de la VoQ en SHE.

Como punto de partida, y que sirve como referencia para presentar los resultados de la combinación de parámetros, se utilizó el experimento presentado en la en

la Sección 5.4.1.2, donde se mostró el comportamiento del LDA durante la discriminación de los 5 estilos de habla expresivos usando únicamente un parámetro de VoQ como entrada al clasificador (Tabla 5.3).

El siguiente paso fue el de hallar la configuración óptima de parámetros de VoQ que maximizara la medida $F1$ durante la clasificación realizada por el LDA. Para ello se realizó un análisis exhaustivo para todas las combinaciones posibles de los parámetros de VoQ, llegando al resultado presentado en la Tabla 5.7, donde se muestra la relación entre el valor máximo de $F1$ y los parámetros de VoQ asociados al proceso de discriminación.

	Neutro	**Alegre**	**Triste**	**Sensual**	**Agresivo**	***Global***
Jitter	●	- -	- -	- -	- -	●
Shimmer	●	- -	- -	- -	●	●
HNR	●	- -	●	●	●	●
GNE	●	- -	●	●	●	●
SFM	●	●	●	- -	- -	●
do1000	- -	- -	●	●	●	- -
HammI	●	- -	●	●	●	●
pe1000	- -	- -	●	●	●	●
***Medida F*1**	$0,52$	$0,38$	$0,47$	$0,66$	$0,54$	$0,48$

Tabla 5.7: *Configuraciones óptimas para los parámetros de cualidad de la voz involucrados en la discriminación de estilos de habla expresivos (indicando con "●" la implicación del parámetro)*

El análisis exhaustivo de las posibles configuraciones de VoQ provee de una detallada relación de los parámetros de VoQ implicados en la discriminación de estilos de habla expresivos, obteniendo la configuración óptima que maximiza los resultados de la clasificación (Tabla 5.7). Los parámetros mostrados en la Tabla 5.3, que daban el máximo de $F1$, se mantienen en la Tabla 5.7, excepto para el estilo agresivo usando *jitter*, donde esta dependencia ha cambiado. Además, la Tabla 5.7 presenta la configuración óptima global (columna 'Global'), es decir, la configuración para la cual el promedio de $F1$ es máximo usando la mejor combinación de parámetros para cada estilo expresivo. Por último, en cuanto a los valores de $F1$ conseguidos a partir de esta configuración global, sobre cada uno de los 5 estilos de habla expresivos implicados en el experimento, se muestra en la Tabla 5.8.

	Neutro	**Alegre**	**Triste**	**Sensual**	**Agresivo**	***Media***
Medida *F*1	$0,52$	$0,32$	$0,46$	$0,64$	$0,48$	$0,48$

Tabla 5.8: *Valores de $F1$ para cada estilo de habla expresivo en la configuración global, que maximiza el promedio de $F1$ calculado*

A partir de los valores de la configuración óptima, se pueden comparar estos con los valores de referencia para obtener los porcentajes de mejora en la discriminación de los estilos de habla expresivos (Tabla 5.9). Considerando el porcentaje de mejora usando la configuración óptima obtenida, se tiene por un lado que los estilos

neutro ($29,19\,\%$), sensual ($19,87\,\%$) y agresivo ($38,20\,\%$) consiguen las mejoras de $F1$ más importantes. Por otro lado, el estilo triste ($7,94\,\%$) consigue una buena mejora, mientras que para alegre ($0\,\%$) esta no existe, evidenciando las limitaciones de la VoQ para su discriminación. Finalmente, la configuración óptima para el análisis global muestra una importante mejora cuando se combinan los parámetros de VoQ ($71,4\,\%$).

	Neutro	**Alegre**	**Triste**	**Sensual**	**Agresivo**	***Media***
***F*1 referencia**	$0,41$	$0,38$	$0,44$	$0,55$	$0,39$	*0,28*
***F*1 conf. ópt.**	$0,52$	$0,38$	$0,47$	$0,66$	$0,54$	*0,48*
% *de mejora*	*29,19*	*0*	*7,94*	*19,87*	*38,20*	*71,4*

Tabla 5.9: *Porcentaje de mejora, en la discriminación de estilos de habla expresivos, de la configuración óptima de parámetros de cualidad de la voz respecto de su referencia*

Analizada la capacidad discriminatoria de los parámetros de VoQ, sobre los estilos de habla expresivos, se fue un paso más allá realizando nuevos experimentos. Se analizaron las parejas de estilos de habla expresivos, de tal manera que se extrajeron las dependencias existentes entre ellos y los parámetros de VoQ. Esta propuesta metodológica se elaboró con la finalidad de conocer aquellas propiedades del habla relacionadas con la VoQ que caracterizaran a cada una de las parejas de estilos de habla expresivos, permitiendo su discriminación (p. ej. en aplicaciones de reconocimiento de emociones) y su control (p. ej. en aplicaciones de SHE durante la transformación de estilos de habla expresivos).

El primer paso fue establecer un valor de referencia para el uso de la configuración óptima de los parámetros de VoQ. Para ello se repitió la clasificación mediante el LDA usando cada uno de los parámetros de VoQ independientemente como entrada al clasificador, pero esta vez para cada una de las parejas de estilos de habla expresivos, dando como resultado los valores de $F1$ presentados en la Tabla 5.10. El valor de referencia finalmente utilizado es el marcado en **negrita** para cada columna de la Tabla 5.10, correspondiendo con la $F1$ máxima obtenida por pareja de estilo de habla expresivo.

	NEU ALE	**NEU TRI**	**NEU SEN**	**NEU AGR**	**ALE TRI**	**ALE SEN**	**ALE AGR**	**TRI SEN**	**TRI AGR**	**SEN AGR**
Jitter	$0,52$	$\mathbf{0,69}$	$0,71$	$0,49$	$0,68$	$0,71$	$0,53$	$0,54$	$0,70$	$\mathbf{0,73}$
Shimmer	$0,60$	$0,63$	$0,64$	$0,61$	$0,53$	$0,57$	$0,50$	$0,54$	$0,53$	$0,56$
HNR	$0,55$	$0,59$	$0,68$	$0,53$	$0,54$	$0,63$	$0,50$	$0,56$	$0,55$	$0,63$
GNE	$\mathbf{0,61}$	$0,53$	$\mathbf{0,80}$	$\mathbf{0,62}$	$0,54$	$\mathbf{0,73}$	$0,51$	$\mathbf{0,72}$	$0,55$	$\mathbf{0,73}$
SFM	$0,58$	$0,64$	$0,55$	$0,60$	$\mathbf{0,71}$	$0,63$	$0,49$	$0,59$	$\mathbf{0,72}$	$0,64$
do1000	$0,49$	$0,52$	$0,57$	$0,50$	$0,50$	$0,56$	$0,51$	$0,57$	$0,52$	$0,58$
HammI	$0,57$	$0,61$	$0,54$	$0,59$	$0,70$	$0,63$	$0,51$	$0,59$	$\mathbf{0,72}$	$0,64$
pe1000	$0,55$	$0,52$	$0,52$	$0,59$	$0,57$	$0,53$	$\mathbf{0,54}$	$0,54$	$0,61$	$0,57$

Tabla 5.10: *Valores de $F1$ para la clasificación del LDA, usando un sólo parámetro de cualidad de la voz por pareja de estilos de habla expresivos (máximo valor por columna en **negrita**)*

Una vez definido el valor de referencia (máximo valor por columna de la Tabla 5.10), se pasó a establecer la configuración óptima por parejas de estilos de habla expresivos. Para ello se repitió el proceso de búsqueda exhaustiva usando clasificadores LDA, considerando todas las combinaciones posibles de los parámetros de VoQ como entradas. La configuración óptima se seleccionó como aquella que obtuvo el valor de $F1$ máximo durante el proceso de clasificación (Tabla 5.11).

	NEU ALE	**NEU TRI**	**NEU SEN**	**NEU AGR**	**ALE TRI**	**ALE SEN**	**ALE AGR**	**TRI SEN**	**TRI AGR**	**SEN AGR**
Jitter	•	•	- -	•	•	•	•	- -	•	•
Shimmer	•	- -	- -	•	- -	•	•	- -	•	•
HNR	•	•	•	•	•	•	•	•	- -	•
GNE	•	•	•	•	•	•	•	•	•	•
SFM	•	•	•	- -	- -	- -	•	- -	- -	- -
do1000	•	•	- -	•	•	•	•	- -	•	•
HammI	•	- -	- -	•	•	•	•	•	•	•
pe1000	•	- -	•	•	•	•	- -	- -	•	•
***Medida F*1**	$0,73$	$0,75$	$0,87$	$0,81$	$0,81$	$0,87$	$0,63$	$0,74$	$0,86$	$0,89$

Tabla 5.11: *Configuración óptima de parámetros de cualidad de la voz por parejas de estilos de habla expresivos ("•" indica la implicación del parámetro de cualidad de la voz)*

A partir de las configuraciones óptimas por parejas de estilos de habla expresivos y de los valores de referencia, la Tabla 5.12 muestra los porcentajes de mejora que reporta la combinación de los parámetros de VoQ respecto al uso de un único parámetro que daba la máxima $F1$. Considerando los valores máximos tomados de la Tabla 5.10 y el valor de $F1$ de la Tabla 5.11, se aprecian interesantes mejoras especialmente para los casos de 'NEU-AGR' ($31,88\,\%$) y 'SEN-AGR' ($22,02\,\%$), donde la combinación de parámetros demuestra ser muy beneficiosa. Además, los valores absolutos de $F1$ para ambos casos muestran un elevado nivel de clasificación ($0,81$ y $0,89$ respectivamente), dando pie a que en futuros estudios, estudios sobre métodos de clasificación puedan mejorar aún más los resultados conseguidos.

	NEU ALE	**NEU TRI**	**NEU SEN**	**NEU AGR**	**ALE TRI**	**ALE SEN**	**ALE AGR**	**TRI SEN**	**TRI AGR**	**SEN AGR**
***F*1 referencia**	$0,61$	$0,69$	$0,80$	$0,62$	$0,71$	$0,73$	$0,54$	$0,72$	$0,72$	$0,73$
***F*1 conf. ópt.**	$0,73$	$0,75$	$0,87$	$0,81$	$0,81$	$0,87$	$0,63$	$0,74$	$0,86$	$0,89$
% de mejora	$19,42$	$9,44$	$8,58$	$31,88$	$13,18$	$18,33$	$16,53$	$2,71$	$19,66$	$22,02$

Tabla 5.12: *Porcentaje de mejora, en la discriminación de parejas de estilos de habla expresivos, de la configuración óptima de parámetros de cualidad de la voz respecto de su referencia*

De los resultados obtenidos a partir del análisis de la discriminación mediante parejas de estilos de habla expresivos, se desprende una conclusión interesante. La VoQ tenía dificultades para discriminar el estilo alegre (Tabla 5.9) cuando se combinaba con el resto de estilos de habla expresivos al mismo tiempo, pero eso cambia

cuando se combina por parejas, momento en el cual se consiguieron buenos resultados. Este hecho nos permitirá reconocer y modificar el habla con respecto a este estilo de habla expresivo.

	Parámetros de cualidad de la voz implicados	p
NEU-ALE	Jitter, Shimmer, HNR, GNE, SFM, do1000, HammI, pe1000	$*$
NEU-TRI	Jitter, HNR, GNE, SFM, do1000	$*$
NEU-SEN	HNR, GNE, SFM, pe1000	$*$
NEU-AGR	Jitter, Shimmer, HNR, GNE, do1000, HammI, pe1000	$*$
ALE-TRI	Jitter, HNR, GNE, do1000, HammI, pe1000	$*$
ALE-SEN	Jitter, Shimmer, HNR, GNE, do1000, HammI, pe1000	$*$
ALE-AGR	Jitter, Shimmer, HNR, GNE, SFM, do1000, HammI	$0,38$
TRI-SEN	HNR, GNE, HammI	$*$
TRI-AGR	Jitter, Shimmer, GNE, do1000, HammI, pe1000	$*$
SEN-AGR	Jitter, Shimmer, HNR, GNE, do1000, HammI, pe1000	$*$

Tabla 5.13: *Nivel de significación p, para la comparación de parejas de estilos de habla expresivos, por parámetro de cualidad de la voz ("$*$" indica nivel por debajo del umbral $p < 0,05$)*

Por último, del mismo modo que se hizo en la Sección 5.4.1.3, se validaron estadísticamente los resultados obtenidos para las diferentes configuraciones con los parámetros de VoQ y las parejas de estilos de habla expresivos. Para ello se usó el análisis Análisis Multivariante de la Varianza —*Multivariate ANalysis Of VAriance*— (MANOVA)[5], utilizando el valor de significación (p) con un valor umbral de $0,05$, calculado sobre las distribuciones de los datos extraídas de la parametrización del corpus (Sección 5.4.1.1) para cada configuración de parámetros de VoQ y a través de todas las parejas de estilos de habla expresivos (Tabla 5.13). Nótese como la validación de la discriminación muestra los problemas que tiene la relación entre los estilos alegre y agresivo ($p = 0,38$) usando los parámetros de VoQ. El resto de las parejas de estilos de habla expresivos son significativamente diferentes y, por tanto, la capacidad de la VoQ para ser usada en la discriminación de estilos de habla expresivos queda validada. Además, los resultados presentados en las Tablas 5.11 y 5.13, nos muestran como el estilo de habla expresivo alegre tiene un buen comportamiento cuando es usado en pareja con otros estilos, excepto para el ya comentado caso de agresivo.

[5]Análisis de Análisis de Varianza —*ANalysis Of VAriance*— (ANOVA) donde hay más de una variable dependiente a ser analizada

5.4.3. Ampliación de la capacidad discriminatoria

Tal y como se presenta en las Secciones 5.4.1 y 5.4.2, la capacidad para la discriminación de voz expresiva con parámetros de VoQ para el corpus estudiado ha quedado probada. A partir de los resultados obtenidos se pueden extraer las reglas generales y la dependencias existentes entre los estilos de habla expresivos, de los que consta el corpus de voz usado para la SHE (Sección 4.3), y los parámetros de VoQ. Aun así, no se ha analizado el comportamiento en situaciones más reales, con frases más largas que provoquen una mayor variabilidad de los parámetros analizados, o la dependencia con la lengua utilizada. De este modo, se considera de utilidad mostrar los resultados obtenidos para el caso de usar las frases de las que consta el corpus en castellano del grupo GTM (Sección 4.3) y un corpus de habla expresiva en alemán (Sección 4.4.1), donde cambió la lengua y que era multilocutor, pudiendo realizar comparativas en el comportamiento de los parámetros de VoQ.

Se partió de los mismos parámetros de VoQ empleados en las Secciones 5.4.1 y 5.4.2 (*jitter*, *shimmer*, HNR, GNE, SFM, do1000, HammI y pe1000) y del mismo corpus de palabras alineadas en castellano (Sección 4.3). Frente a esta información, mediante el análisis únicamente de las vocales, se añadió la parametrizaron de las frases de este mismo corpus en castellano (Sección 4.3) y del corpus de habla expresiva multilocutor en alemán (Sección 4.4.1).

Debido a que se deseaba comparar los resultados entre los tres corpus: corpus de palabras en castellano, corpus de frases en castellano y corpus en alemán; los estilos de habla expresivos que se vieron envueltos en el experimento fueron los mismos: neutro, alegre, sensual, agresivo y triste. Para el caso del corpus en alemán no existía el estilo sensual, de modo que no se pudo hacer la comparativa para este, mientras que los estilos agresivo (castellano) y enfado (alemán) se consideraron similares y de esta manera fueron usados en las comparaciones.

Con los corpus parametrizados, se realizó una comparación sobre qué conjunto de parámetros reportaba una mejor discriminación de las parejas de estilos de habla expresivos implicados. Como clasificador se usó el LDA, y la medida $F1$ como referencia para realizar las comparaciones. Las configuraciones usadas para la clasificación fueron todas las combinaciones posibles de parámetros (búsqueda exhaustiva), aplicadas sobre la discriminación de cada una de las parejas de estilos de habla expresivos.

A partir de los resultados obtenidos, se consideraron como las mejores configuraciones de parámetros aquellas en el margen entre el valor máximo menos la desviación estándar de los valores de $F1$. Se contabilizó la aparición de cada uno de los parámetros en estas configuraciones, expresando este valor como un $\%$ del total de veces que aparecía el parámetro en el margen de $F1$ seleccionado, de modo que se pudiera comparar su contribución entre los distintos corpus implicados en el experimento (Tabla 5.14). Se decidió utilizar esta estrategia, debido a que usando únicamente el valor del máximo de $F1$ como en la Sección 5.4.2 podían estar descartándose parámetros de utilidad simplemente por no estar asociados a la configuración de la máxima $F1$. Asimismo, de esta forma puede detectarse la utilidad de algún parámetro claramente sobre el resto, siendo de utilidad a la hora de aplicar reglas de transformación entre estilos de habla expresivos en la SHE.

La Tabla 5.14 muestra los porcentajes de aparición, ordenados por frecuencia de aparición en orden descendente, de cada uno de los parámetros usados en la discriminación de cada una de las parejas de estilos de habla expresivos: neutro (NEU), alegre (ALE), triste (TRI), sensual (SEN) y agresivo (AGR). Con estos resultados se muestra la relación de los parámetros con cada una de las parejas de estilos de habla expresivos y se puede analizar si se mantiene entre diferentes tipos de enunciados (palabras/frases), lenguas (castellano/alemán) y locutores (uno o varios en el corpus). Para el caso del corpus en alemán, no todas las parejas de estilos expresivos han podido ser estudiadas por no existir el estilo sensual en el corpus, indicándose estos casos con el símbolo "- -".

	Corp. castellano (palabras)		**Corp. castellano (frases)**		**Corp. alemán**	
	Pár. VoQ	**% uso**	**Pár. VoQ**	**% uso**	**Pár. VoQ**	**% uso**
NEU-ALE	GNE	18	GNE	14	HammI	23
	HammI	17	*Shimmer*	14	*Jitter*	11
	Shimmer	13	HammI	14	*Shimmer*	11
	HNR	12	HNR	13	HNR	11
	SFM	11	SFM	13	GNE	11
	Jitter	10	do1000	11	SFM	11
	pe1000	10	*Jitter*	11	do1000	11
	do1000	9	pe1000	10	pe1000	11
NEU-TRI	*Jitter*	15	HammI	17	GNE	20
	SFM	15	SFM	13	pe1000	12
	HNR	13	GNE	12	SFM	12
	GNE	12	*Jitter*	12	HammI	12
	HammI	12	*Shimmer*	12	*Jitter*	11
	do1000	11	do1000	12	*Shimmer*	11
	pe1000	11	HNR	11	HNR	11
	Shimmer	11	pe1000	11	do1000	11
NEU-SEN	GNE	23	GNE	23	- -	- -
	Jitter	11	*Jitter*	11	- -	- -
	Shimmer	11	*Shimmer*	11	- -	- -
	HNR	11	HNR	11	- -	- -
	SFM	11	SFM	11	- -	- -
	do1000	11	do1000	11	- -	- -
	HammI	11	HammI	11	- -	- -
	pe1000	11	pe1000	11	- -	- -

Continúa en la página siguiente

Tabla 5.14 – continúa de la página anterior

	Corp. castellano (palabras)		Corp. castellano (frases)		Corp. alemán	
	Pár. VoQ	**% uso**	**Pár. VoQ**	**% uso**	**Pár. VoQ**	**% uso**
NEU-AGR	GNE	18	GNE	15	HammI	18
	HammI	17	HammI	15	SFM	13
	Shimmer	13	*Shimmer*	14	*Jitter*	12
	SFM	11	*Jitter*	13	do1000	12
	pe1000	11	HNR	12	*Shimmer*	12
	HNR	11	SFM	11	HNR	11
	Jitter	10	pe1000	11	pe1000	11
	do1000	9	do1000	9	GNE	10
ALE-TRI	HammI	16	HammI	16	HammI	19
	SFM	14	SFM	15	*Jitter*	12
	Jitter	13	*Jitter*	12	GNE	12
	HNR	13	GNE	12	SFM	12
	do1000	11	*Shimmer*	12	HNR	12
	pe1000	11	do1000	11	*Shimmer*	11
	Shimmer	11	HNR	11	do1000	11
	GNE	11	pe1000	11	pe1000	11
ALE-SEN	GNE	18	HammI	16	- -	- -
	Jitter	13	GNE	13	- -	- -
	HNR	13	*Jitter*	13	- -	- -
	HammI	12	HNR	13	- -	- -
	SFM	12	SFM	12	- -	- -
	do1000	11	*Shimmer*	11	- -	- -
	pe1000	11	do1000	11	- -	- -
	Shimmer	10	pe1000	11	- -	- -
ALE-AGR	*Jitter*	19	*Jitter*	21	*Jitter*	15
	do1000	16	*Shimmer*	14	HNR	15
	Shimmer	13	HNR	13	pe1000	14
	GNE	11	SFM	11	*Shimmer*	14
	pe1000	11	do1000	11	HammI	12
	HammI	10	HammI	10	SFM	11
	HNR	10	pe1000	10	do1000	10
	SFM	10	GNE	10	GNE	9

Continúa en la página siguiente

Tabla 5.14 – continúa de la página anterior

	Corp. castellano (palabras)		Corp. castellano (frases)		Corp. alemán	
	Pár. VoQ	**% uso**	**Pár. VoQ**	**% uso**	**Pár. VoQ**	**% uso**
TRI-SEN	GNE	23	SFM	15	- -	- -
	Jitter	11	GNE	15	- -	- -
	Shimmer	11	HammI	14	- -	- -
	HNR	11	*Jitter*	12	- -	- -
	SFM	11	HNR	12	- -	- -
	do1000	11	*Shimmer*	11	- -	- -
	HammI	11	pe1000	11	- -	- -
	pe1000	11	do1000	10	- -	- -
TRI-AGR	HammI	15	HammI	19	HammI	18
	Jitter	14	*Jitter*	13	SFM	13
	pe1000	14	SFM	13	do1000	12
	do1000	13	*Shimmer*	11	*Jitter*	12
	SFM	13	HNR	11	GNE	12
	Shimmer	11	GNE	11	HNR	11
	HNR	11	pe1000	11	pe1000	11
	GNE	9	do1000	11	*Shimmer*	11
SEN-AGR	GNE	15	*Jitter*	17	- -	- -
	pe1000	14	HammI	15	- -	- -
	Jitter	13	GNE	14	- -	- -
	do1000	13	HNR	12	- -	- -
	HNR	12	pe1000	11	- -	- -
	HammI	12	do1000	11	- -	- -
	SFM	11	*Shimmer*	10	- -	- -
	Shimmer	10	SFM	10	- -	- -

Tabla 5.14: *Porcentaje de uso de los parámetros de cualidad de la voz, que mejor discriminan parejas de estilos de habla expresivos, sobre distintos corpus de habla expresiva*

De la Tabla 5.14 se desprende que, en gran medida, los parámetros de VoQ con porcentajes de aparición más altos se corresponden para todos los corpus bajo estudio. En especial se aprecia claramente este comportamiento para las parejas 'NEU-SEN' (únicamente para corpus en castellano), 'ALE-TRI', 'ALE-AGR' y 'TRI-AGR'. En el resto de casos existe una relación entre parámetros y corpus, pero no es tan importante. Es interesante ver como la importancia de cada parámetro, sobre la discriminación de cada pareja de estilos de habla expresivos, se mantiene aún cuando existe un cambio en las características de los corpus. Se comprueba como entre las dos partes del corpus en castellano las variaciones son menores, mientras que para el alemán aparecen mayores divergencias, especialmente para el estilo neutro, debidas a la gran

variabilidad existente en el propio corpus y a la baja cantidad de información (baja duración) de la que consta (Sección 4.4.1). Por el contrario, para el resto de estilos de habla expresivos se tiene una mejor correspondencia con el corpus en castellano, quedando mejor caracterizadas estas parejas de estilos expresivos por los parámetros de VoQ.

Finalmente, la Tabla 5.14 puede ser utilizada como herramienta de modelado de la VoQ para su aplicación en el reconocimiento y la síntesis del habla expresiva, ya que se dispone de un resumen de la dependencia entre los estilos de habla expresivos y los parámetros que los caracterizan.

5.4.4. Discriminación en habla expresiva espontánea

En esta sección se presenta el último de los experimentos relacionados con la discriminación de estilos de habla expresivos. A partir de la capacidad de discriminación demostrada en las Secciones 5.4.1 a 5.4.3, el último aspecto en el que se ha trabajado es el de la discriminación de habla expresiva espontánea, es decir, aquellas situaciones donde nos encontremos situaciones reales en las que no existe control sobre los enunciados utilizados.

En este tipo de contextos, donde no existe control sobre los enunciados, aparecen importantes novedades con respecto a los experimentos que se habían aplicado hasta el momento:

- El habla es espontánea y no se basa en la lectura de unos textos bien diseñados (p. ej. debido a la interacción del locutor con su entorno).
- Al ser habla espontánea se pueden encontrar artefactos en la señal de voz ajenos al mensaje oral (p. ej. ruido al cerrarse una puerta).
- La calidad de la señal de voz disponible es inferior por no grabarse en un estudio de grabación profesional (p. ej. no se mantiene la distancia locutor-micrófono).
- La transcripción fonética puede no coincidir fielmente con el enunciado leído (p. ej. utilizando un reconocedor automático del habla), o bien puede directamente no existir (p. ej. no existe un guión a seguir a partir del diseño de los textos).
- El/la locutor/ra puede no ser profesional (p. ej. niños).

Para el caso genérico de habla espontánea (Sección 5.4.4.2) se realizaron adaptaciones de las medidas de VoQ, ya que entre otros problemas que se pueden encontrar, el desconocimiento de los fonemas asociados a la señal de voz imposibilitan un análisis centrado en ciertas unidades, tal y como se hacía hasta ese momento (Sección 5.4.1). La mayor dificultad que aparece en estas situaciones son la calidad de la señal de voz, ya que se pueden encontrar artefactos como ruidos, risas u onomatopeyas, que pueden dificultar la parametrización.

5.4.4.1. Descripción

Este trabajo se realizó como parte de un estudio sobre reconocimiento de emociones conjuntamente con Planet et al. (2009), participando en el *INTERSPEECH 2009 Emotion Challenge* (Schuller et al., 2009) dentro de la categoría de *Feature Sub-Challenge*.

Antes de pasar a describir el trabajo desarrollado, lo ubicaremos en el contexto de este *challenge*. El campo del reconocimiento de emociones a partir de la voz ha ido ganando recientemente un considerable interés en la comunicación hombre-máquina, comunicación hombre-robot y recuperación de información multimedia. Son numerosos los estudios que en las últimas décadas han intentado mejorar los clasificadores usados. Sin embargo, a diferencia de las tareas relacionadas con el procesado del habla y el reconocimiento automático del habla y de locutor, prácticamente no existen corpus y condiciones de test estandarizadas para comparar rendimientos bajo las mismas condiciones, y el empleo de multiplicidad de estrategias de evaluación, sin una definición consensuada, impiden su idéntica reproducción. Además, para poder realizar casos de uso más realistas, se necesita de información más espontánea y menos prototípica.

Es en este contexto donde el *INTERSPEECH 2009 Emotion Challenge* ayuda a unir la investigación realizada en reconocimiento de emociones a partir de habla y la baja compatibilidad de resultados. La organización proveyó el corpus de habla expresiva FAU Aibo (Sección 4.4.2), de habla espontánea multilocutor en alemán con contenido emocional, y los resultados de referencia de dos de los enfoques de arquitectura más populares. El hecho de que este corpus disponga de aproximadamente 9 horas de habla, con 51 niñas y niños de 2 escuelas distintas, permite la definición de particiones de entrenamiento y de test incorporando independencia de locutor tal y como se necesita en situaciones de la vida real.

En el *Feature Sub-Challenge*, los participantes debían de utilizar sus mejores características individuales por unidad de análisis, con un máximo de 100. Estas características fueron testadas por los organizadores, con configuraciones equivalentes en una tarea de clasificación, y puestos en común en un proceso de selección de características. En particular, la petición realizada fue la de utilizar características novedosas, de alto nivel o de percepción adecuada.

5.4.4.2. Contribución

Se contribuyó a partir de la inclusión de parámetros de VoQ que pudieran ayudar a la mejor clasificación emocional del corpus. Para ello, además de la propuesta de parámetros de VoQ, se realizó la adaptación de la metodología para su obtención, manteniendo la metodología para su análisis, ya que la información de la que se disponía para el entrenamiento y para el test estaría limitada. Por tanto, el procedimiento que se definió difiere del utilizado en el estudio de la capacidad discriminatoria de los parámetros de VoQ (Secciones 5.4.1 a 5.4.3), donde recordemos que se analizaban únicamente zonas vocálicas. En este caso en cambio, solamente se disponía de la

información de voz con la que trabajar, de modo que no se disponía de ninguna información de los fonemas bajo estudio, siendo por tanto necesario definir la metodología para la obtención de los parámetros de VoQ para este tipo de casos:

1. La medida de los parámetros de VoQ se realizó sobre las partes sonoras del habla, detectadas en base al uso de un marcador de *pitch* basado en el algoritmo RAPT (Talkin, 1995), que a partir de detección de sonoridad únicamente coloca marcas en estas zonas.

2. Sobre las zonas sonoras, los parámetros de VoQ se calcularon utilizando una ventana de tamaño de 40 ms y un paso de 20 ms. Sobre esta ventana de 40 ms se llevaron a cabo las medidas de VoQ del mismo modo que cuando se trataba la información de una vocal, no existiendo diferencias en la metodología de análisis del parámetro en cuestión.

Debido a la falta de información de los fonemas, la VoQ fue calculada tanto en fonemas vocálicos como no vocálicos, hecho que también constituye diferencia respecto de los experimentos de discriminación realizados hasta el momento (Secciones 5.4.1 a 5.4.3), y que aún da mayor importancia a este experimento, que como se puede ver se realiza en unas condiciones más complejas en cuanto al control de los valores de los parámetros obtenidos.

En lo que se refiere a los parámetros que finalmente se vieron envueltos en el experimento, se excluyeron los parámetros de ruido HNR y GNE, principalmente por ya ser usado el HNR por los organizadores y por ser similar la información que aporta el GNE respecto al primero (recordemos que el número de características que podían ser utilizadas estaba limitada a 100). De esta manera, los parámetros de VoQ finalmente propuestos en este trabajo fueron: *jitter*, *shimmer*, SFM, do1000, HammI y pe1000; donde *jitter* y *shimmer* fueron calculados siguiendo la metodología descrita en la Sección 5.3, pudiendo así probar su comportamiento en situaciones reales, y el resto mediante la definición presentada en la Sección 2.4.2.4.

A diferencia del procedimiento seguido en las Secciones 5.4.1 a 5.4.3, donde se utilizaba la distribución de valores extraídos de la parametrización de la VoQ durante el análisis del corpus, en este experimento se adaptó la información de los parámetros a la propuesta realizada en los trabajos previos de Iriondo et al. (2007a, 2009), donde para cada enunciado y parámetro implicado en el experimento se calcularon una serie de estadísticas. En concreto fueron las siguientes 11 estadísticas: media, desviación estándar, mediana, valor máximo, valor mínimo, rango, primer cuartil, tercer cuartil, rango intercuartílico, asimetría (*skewness*) y curtosis (*kurtosis*). Finalmente, considerando tanto los parámetros de VoQ, que son los que nos interesan en este momento, como el resto involucrados en el experimento (prosodia, espectrales y de VoQ proporcionados por los organizadores (Schuller et al., 2009)), fueron un total de 454 características de las que se partió excluyendo la etiqueta con la emoción, y de las cuales recordemos que solamente 100 podían ser empleadas en el *Feature Sub-Challenge*.

Parámetro	Estadística aplicada	Orden
Jitter	Mediana	1
	Desviación estándar	5
	Valor mínimo	6
	Valor máximo	65
	Primer cuartil	67
	Media	70
	Rango intercuartílico	75
	Curtosis	81
Shimmer	Desviación estándar	9
	Primer cuartil	10
	Rango intercuartílico	19
	Valor mínimo	20
	Media	22
	Rango	27
	Tercer cuartil	34
	Valor máximo	45
HNR	Valor mínimo	3
	Curtosis	16
SFM	- -	- -
do1000	Rango intercuartílico	54
HammI	Curtosis	91
pe1000	Primer cuartil	51
	Media	62
	Curtosis	89

Tabla 5.15: *Estadísticas de los parámetros de cualidad de la voz y su orden entre las* 100 *características seleccionadas en el Interspeech 2009 Feature Sub-Challenge*

5.4.4.3. Resultados

A partir del análisis de los resultados, se estudió si entre las 100 características la VoQ fue elegida como información significante. La selección de los atributos, dada la elevada cantidad de información de la que se disponía, fue realizada siguiendo el criterio de Mínima-Redundancia Máxima-Relevancia —*Minimal-Redundancy Maximal-Relevance*— (mRMR) (Peng et al., 2005), que tiene como objetivo seleccionar aquellos atributos con menor información mutua cuya relevancia esté lo más próxima a su clase. En la Tabla 5.15 se muestra cada uno los parámetros de VoQ, y en el caso de haber sido uno de los 100 atributos elegidos se indica la estadística que fue seleccionada y el orden de relevancia usando el criterio mRMR. A pesar de que el parámetro HNR fue dado por los organizadores, no deja de ser un parámetro de VoQ y como tal se incluye en la misma tabla.

En la Tabla 5.15 se observa como la configuración de parámetro-estadística, considerada como de mayor utilidad para la detección de las emociones, resultó ser la mediana del *jitter*. Las estadísticas para los parámetros *jitter* y *shimmer* fueron

las características con mayor representación, entre todas las relacionadas con los parámetros de VoQ, que fueron seleccionadas usando el criterio mRMR. Se puede comprobar como todos los parámetros de VoQ, excepto SFM, y todas las estadísticas, excepto la asimetría, han sido seleccionadas entre las 100 características más significativas para la discriminación de habla expresiva espontánea.

Estos resultados son de utilidad a la hora de plantear la modificación de los parámetros de VoQ durante la transformación de estilos de habla expresivos (Sección 5.5), ya que permiten ver la importancia que tienen algunos parámetros (como el *jitter* y el *shimmer*) y relajar las exigencias de disponer de un modelo para todos y cada uno de ellos en aquellos casos en los que el parámetro no ha demostrado ser de tanta utilidad como otros (p. ej. el SFM).

5.5. Transformación de estilos de habla expresivos

Conocidos los parámetros de VoQ (Sección 5.2.1) y las posibilidades que ofrecen en el reconocimiento automático de estilos de habla expresivos, especialmente después de comprobar su capacidad para discriminarlos (Sección 5.4), el siguiente paso fue el de aplicar todo este conocimiento en la transformación de estilos de habla expresivos (Monzo et al., 2010).

5.5.1. Introducción

En esta sección, se propone la metodología que se llevó a cabo para la transformación del habla, desde un estilo neutro hacia los otros estilos definidos en el corpus del grupo de investigación GTM (Sección 4.3), empleando tanto parámetros de prosodia como de VoQ (Sección 5.5.2). El objetivo marcado fue doble:

1. Proponer un modelo de modificación de los parámetros de VoQ utilizando HNM para la SHE (Sección 5.5.3).

2. Mostrar la mejora de la calidad del habla y de la identificación del estilo de habla expresivo transmitido por el sistema de CTH para la SHE, mediante la combinación de parámetros de prosodia y de VoQ utilizando HNM con respecto al uso de diferentes configuraciones de parámetros implicados (prosodia y VoQ) y de algoritmos usados en la CTH (PSOLA y HNM) (Sección 5.5.4).

Para llevar a cabo las transformaciones se eligió la síntesis basada en HNM debido a su flexibilidad (Stylianou, 2001; Drioli et al., 2003; Erro, 2008), permitiendo el acceso directo a las variables que controlan a cada uno de los parámetros que caracterizan el habla (prosodia y VoQ). En experimentos previos se había utilizado el sistema de CTH, implementado por la herramienta Praat, basado en PSOLA (Monzo et al., 2008b), que presentaba el inconveniente de que no todos los parámetros podían ser modificados, debido al desconocimiento de la componente determinista y/o la estocástica. Además, las transformaciones de bandas espectrales usando PSOLA son más complejas, y más teniendo en cuenta que no se desea introducir distorsión

en el habla resultante. Por tanto, HNM permite un mayor control de los parámetros del que presenta PSOLA, facilitando su manipulación.

De este modo, para la modificación de la VoQ se empleó la implementación de Calzada (2008, 2010) de un sistema de CTH en el que se incorpora el HNM (Sección 3.3.2.5). Recientemente se ha incrementado el interés en el uso de HNM para la transformación de habla (Erro, 2008), ya que se consigue una gran calidad y a la vez flexibilidad de modificación y conversión. La parametrización del habla en una componente determinista (armónica) y otra estocástica (ruido) permite una flexible manipulación de la VoQ y de las escalas de tiempo y F_0, manteniendo la naturalidad del habla. En el caso de haber utilizado PSOLA en lugar de HNM, este tiene el inconveniente de la falta de control sobre los parámetros espectrales HNR, SFM, do1000, *HammI* y pe1000, necesitando estos de un análisis de la señal de voz que permitiera extraer la información espectral.

Para los experimentos de transformación se utilizó el corpus expresivo en castellano presentado en la Sección 4.3. A modo de recordatorio, este corpus tiene una duración de 5 horas y 27 minutos y consta de 5 estilos de habla expresivos: neutro, alegre, sensual , agresivo y triste. Para los experimentos aquí realizados se emplearon las frases y no las palabras portadoras, debido a que contenían mayor cantidad de información, especialmente para el correcto modelado prosódico al ser enunciados más largos que aportan mayor riqueza a la caracterización del estilo de habla expresivo.

La modificación de los parámetros se llevaron a cabo sobre toda la frase, considerando toda la señal de voz como información que se desea modificar. A partir de la parametrización de la señal de voz se aplicaron los modelos de los parámetros en el estilo de habla expresivo deseado, permitiendo así su modificación y por tanto la transformación del estilo neutro.

Finalmente, las transformaciones de estilos de habla expresivos se evaluaron mediante pruebas perceptivas (Sección 5.5.4.2), pudiendo comprobar si los oyentes preferían el resultado obtenido mediante la modificación de los parámetros de VoQ conjuntamente con la manipulación de la prosodia respecto al uso únicamente de prosodia. También se recogieron aspectos de la variación de la calidad y de la identificación de los estilos de habla expresivos con el fin de evaluar la calidad con la que el oyente captaba el enunciado y la intensidad del estilo expresivo percibido.

5.5.2. Parámetros implicados

Los parámetros implicados en los experimentos de transformación de los estilos de habla expresivos fueron los de prosodia y de VoQ. A partir de los modelos de cada uno de estos parámetros se modificaron los parámetros resultantes de la parametrización del habla basada en HNM, que permitió la manipulación de los parámetros anteriores y la síntesis de la señal de voz final (Sección 5.5.3).

5.5.2.1. Parámetros del modelo armónico más ruido (HNM)

La parametrización del habla basada en HNM, utilizada en los experimentos de transformación de estilos de habla expresivos, es la que encontramos en el Apéndice B.1. A partir de las componentes determinista y estocástica se ha llevado a cabo el proceso de transformación aplicando los modelos de prosodia y de VoQ.

A continuación se presentan los parámetros del HNM usados durante los experimentos y su relación con la nomenclatura definida en el Apéndice B.3:

- Ancho de banda armónico. Valor constante fijado a 5000 KHz (Erro, 2008). Este valor está relacionado con el valor de I_k, de manera que el número de armónicos escogidos para cada trama es tal que el ancho de banda es constante.
- Las marcas de análisis de la componente determinista son las marcas de *pitch* situadas en los instantes t_k^s (Ecuación B.4).
- El orden del filtro LPC del ruido. Corresponde con el valor de Q, un valor constante fijado a 30.
- La longitud de las tramas de ruido. Se ha fijado a un valor constante de 10 ms, asociado al tamaño de trama en muestras T^r mediante la frecuencia de muestreo.

5.5.2.2. Parámetros de prosodia

Los parámetros de prosodia fueron extraídos del corpus de habla expresiva y modelados usando el CBR, una herramienta de minería de datos que ha demostrado ser de utilidad en aplicaciones de SHE (Iriondo et al., 2007c). Para un texto de entrada al sistema de CTH, el CBR devuelve una predicción de los parámetros que mejor encajan con él para un estilo expresivo destino. Estos parámetros predichos son:

- Contorno de frecuencia fundamental (F_0) de cada grupo acentual[6].
- Contorno de energía (un valor de la energía por fonema).
- Duración segmental (duración por fonema).

5.5.2.3. Parámetros de cualidad de la voz

Los parámetros de VoQ utilizados durante los experimentos de transformación de habla expresiva son los siguientes, presentados en la Sección 5.2.1 con alguna excepción que se justifica a continuación:

- *Jitter*
- *Shimmer*

[6] Definido como una palabra acentuada y, si es el caso, las palabras átonas que la preceden.

- *Harmonic-to-Noise Ratio* (HNR)
- *Drop-off of Spectral Energy above 1000 Hz* (do1000)
- *Hammarberg Index* (HammI)
- *Relative Amount of Energy above 1000 Hz* (pe1000)

Como se observa en la selección propuesta, ni el parámetro GNE ni SFM se encuentran entre los elegidos. Las razones son, en primer lugar, los beneficios que presenta el uso del HNR junto con HNM frente al GNE (Sección 5.2.3), que aunque representando conceptos similares de la señal de voz, el HNR puede hacer uso de la flexibilidad que le aporta conocer la componente armónica y la de ruido para su estimación y su modificación. Por otro lado, respecto del parámetro SFM, debido a su papel en la discriminación de estilos de expresivos (Sección 5.4, y en especial a su papel en la Sección 5.4.4.3) y de las limitaciones encontradas para su modificación (Sección 5.2.3), se decidió que fuera descartado por el momento.

El proceso de modelado de los parámetros *jitter* y *shimmer* siguieron la propuesta metodológica presentada en la Sección 5.3. El análisis y la modificación de estos dos parámetros es independiente del sistema de síntesis empleado, dado que el primero de ellos solamente tiene en cuenta la información de F_0 y el segundo la señal de habla. En cuanto al caso del *jitter*, tal y como se muestra en las Secciones 5.3.2.1 y 5.3.3, utiliza la información de la F_0 media del corpus para cada estilo de habla expresivo, siendo una información tenida en cuenta como frecuencia de referencia (F_{ref}) durante el proceso de modificación: neutro (188 Hz), alegre (294 Hz), triste (184 Hz), sensual (154 Hz) y agresivo (278 Hz). Respecto al parámetro del modelo correspondiente al filtro de mediana, utilizado durante la búsqueda de tramos donde la tendencia de la prosodia se mantiene (F_0 y amplitudes pico a pico), se fijó un tamaño de 5 muestras tanto para el *jitter* como para el *shimmer*.

Respecto al resto de parámetros de VoQ, estos fueron adaptados para utilizar HNM (Secciones 5.2.2 y 5.2.3), esto quiere decir que se hizo uso de la información de la componente determinista de la señal del habla para extraer la información espectral (HNR, do1000, HammI y pe1000) y la componente estocástica para la información de ruido (HNR). La parametrización previa del habla mediante HNM hace que el análisis y posterior modificación de los parámetros de VoQ sea simple y flexible, hecho que no hubiera sido así en el caso de utilizar un método de CTH basado en PSOLA, que hubiera necesitado usar alguna técnica de análisis y síntesis alternativa que permitiera la manipulación espectral de la señal de voz sintética.

5.5.3. Propuesta metodológica para la transformación

En esta sección se presenta la propuesta metodológica para la transformación a estilos de habla expresivos partiendo de un estilo neutro. El beneficio de añadir VoQ al proceso de transformación, para mejorar la percepción del habla expresiva, fue analizado y comparado con la transformación empleando únicamente la predicción de la prosodia. Trabajos previos ya habían demostrado la utilidad de modelar la prosodia

en la transformación de estilos de habla expresivos, y asimismo se han identificando los problemas que aparecían durante su modelado (Iriondo et al., 2007c), como por ejemplo añadir efectos de temblor y de ruido en la voz o enfatizar ciertas zonas del espectro de frecuencia. Es a partir de estas limitaciones que el uso conjunto de la VoQ y de la prosodia ha sido propuesto en trabajos anteriores, y se ha demostrado que su uso combinado ayuda a mejorar la percepción de la expresividad (Cabral y Oliveira, 2005; Audibert et al., 2006; Monzo et al., 2008b; Iriondo et al., 2009), efecto que se demuestra en los siguientes estudios. De este modo, las limitaciones presentadas para la prosodia pueden resolverse aplicando los parámetros de VoQ, que aportan las características que la prosodia no puede manipular, como por ejemplo mediante el *jitter* y el *shimmer* para aportar un efecto de temblor a la voz, el HNR para variar el grado de ruido y el resto de parámetros para modificar la energía de las diferentes bandas de frecuencia.

En la Figura 5.9 se presenta el diagrama de bloques del sistema de transformación propuesto, que se divide principalmente en tres partes:

- El bloque de **análisis** del HNM que parametriza el habla original y el bloque de **síntesis** que la resintetiza una vez se ha realizado la transformación.
- **Modelado** de la **prosodia** y de la **VoQ**. La prosodia se predice mediante el sistema CBR (Iriondo et al., 2007c), obteniendo el contorno de F_0, el contorno de energía y la duración segmental para cada fonema en el estilo destino. Para el modelado de la VoQ, se extrajeron las relaciones entre estilos de habla expresivos y los parámetros que mejor los caracterizaban a partir de los estudios de discriminación y, mediante el análisis de los diferentes corpus expresivos, se obtuvieron las medidas estadísticas que permitieron disponer de los valores de los parámetros en los estilos origen y destino para así poder llevar a cabo la transformación de estilos (Secciones 5.3 y 5.4).
- **Transformación** del habla a partir de los resultados del análisis HNM y el modelado de la prosodia y de la VoQ del estilo destino a conseguir.

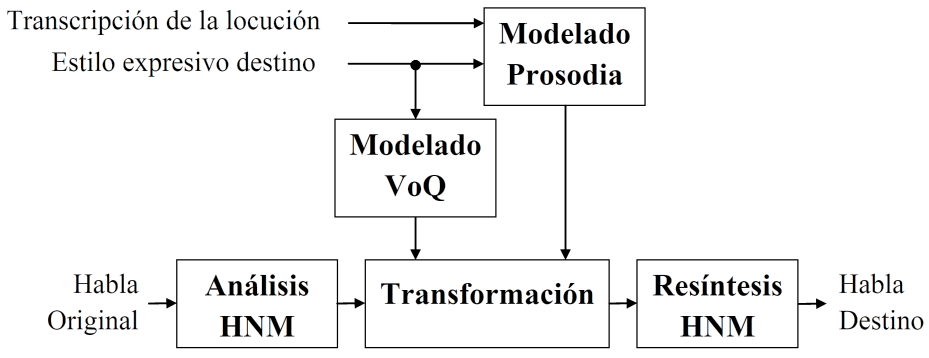


Figura 5.9: *Diagrama de bloques para la transformación de estilos de habla expresivos*

Analíticamente, el modelado de la VoQ (análisis y modificación) se realizó siguiendo las indicaciones presentadas en las Secciones 5.2.2 y 5.2.3. En ellas se presenta cada uno de los parámetros de VoQ y la representación analítica paso a paso para su modelado.

Varias consideraciones deben de ser tenidas en cuenta a la hora de determinar qué parámetros, y sus valores asociados, deberían de estar involucrados en cada una de las transformaciones de cada estilo de habla expresivo. Mientras que para las transformaciones de prosodia se vieron implicados los tres parámetros asociados (F_0, energía y duración segmental), para el caso de la VoQ la selección de parámetros se hizo a partir de los siguientes criterios:

1. Resultados de los estudios previos acerca del uso de parámetros de VoQ en la discriminación de estilos de habla expresivos (Secciones 5.3 y 5.4). Esto está ligado con la hipótesis formulada durante la presentación de los objetivos a alcanzar por esta tesis (Sección 1.2), con lo que a partir de estos estudios se conocen los factores, o parámetros, que rigen la capacidad de percibir correctamente el mensaje oral de un interlocutor, pudiendo ser utilizados durante la transformación de estilos de habla expresivos.

2. Estadística descriptiva calculada para todos los estilos de habla expresivos del corpus y para todos los parámetros de VoQ, comparando los resultados del estilo neutro con los obtenidos para el resto.

3. Pruebas heurísticas realizadas durante el proceso de diseño del sistema de transformación, con el fin de ajustar los parámetros involucrados en las distintas transformaciones.

Con respecto a los valores finales de los parámetros durante la transformación, tanto de prosodia como de VoQ, se tiene por un lado que los valores de destino para la prosodia fueron obtenidos a partir de la predicción, por medio del CBR, de un nuevo contorno de F_0, contorno de energía y duraciones segmentales. Por otra parte, para la modificación de la VoQ se calcularon las estadísticas sobre las distribuciones de los parámetros seleccionados sobre las vocales de cada uno de los estilos expresivos del corpus, tal y como se propuso durante los experimentos de la capacidad discriminatoria de la VoQ (Sección 5.4). De los posibles valores, la media de cada parámetro fue elegida para ser el valor de referencia para las transformaciones de cada estilo de habla expresivo, sobre la cual se calculó el factor de modificación lineal multiplicativo β (Apéndice B.4), aplicado sobre el total de la frase que se deseaba transformar.

El procedimiento que se siguió para la modificación de los parámetros de prosodia y de VoQ fue el de aplicar una transformación lineal sobre de los parámetros obtenidos del análisis del HNM, es decir sobre las amplitudes, frecuencias y fases de la componente determinista y sobre las energías de la componente estocástica. Se multiplicó el parámetro original para cada trama analizada (correspondiente a cada periodo de *pitch*), excepto para el *jitter* y *shimmer* en los que se realizó sobre el conjunto de la frase, por dicho factor de modificación β, convirtiendo previamente a lineal aquellos valores que estuvieran expresados en dB (Sección 5.2.3). Para

el caso concreto de la modificación prosódica del contorno de energía se realizó la transformación sobre las muestras de la señal de voz, mientras que para el parámetro de VoQ *shimmer* se empleó la señal de voz asociada a la componente determinista (Sección 5.3). De este modo, las modificaciones llevadas a cabo fueron de dos tipos:

- **Modificación relacionada con la prosodia.** Utilizando la información de prosodia original y la de destino, a partir de la predicción del CBR, se obtuvieron los factores de transformación para llevar a cabo la modificación de los parámetros. La modificación del contorno de F_0 y de las duraciones segmentales fue realizada sobre la información de frecuencias y de tramas de análisis del HNM de acuerdo al trabajo de Calzada (2010), con lo que se obtuvo la nueva información de frecuencias y la reestimación de las amplitudes y las fases asociadas del HNM. Sin embargo, el contorno de energía fue directamente modificado sobre las muestras de la señal de voz, multiplicando cada una de las muestras por el factor de transformación correspondiente (Iriondo et al., 2007c).
- **Modificación relacionada con la VoQ.** Por su parte, los parámetros de VoQ modifican de forma distinta la representación con HNM (Sección 5.2.3), debiendo de reestimar sus parámetros cuando sea necesario (Calzada, 2008, 2010):
  - El ***jitter*** necesita modificar el contorno de F_0 (Sección 5.2.3.1), de este modo las frecuencias se ven modificadas y como consecuencia se reestiman las amplitudes y las fases del HNM.
  - En el caso del ***shimmer***, todos los parámetros del HNM están implicados debido a que las modificaciones se realizan directamente en el dominio temporal sobre la componente determinista. Debido a que se modifica directamente la señal de voz contenida en la componente determinista (Sección 5.2.3.2), su modificación se realiza como paso previo a la resíntesis, evitando así tener que reestimar las amplitudes, frecuencias y fases del HNM.
  - Para la transformación del parámetro **HNR** tanto la componente armónica como la de ruido deben de ser modificadas, mediante la manipulación respectiva de las amplitudes y de la varianza (Sección 5.2.3.3).
  - El parámetro **do1000** (Sección 5.2.3.4) modifica las amplitudes de la componente determinista.
  - Debido a que los parámetros **HammI** (Sección 5.2.3.5) y **pe1000** (Sección 5.2.3.6) están relacionados con la energía en diferentes bandas de frecuencia de la componente armónica, las amplitudes deben de ser modificadas.
  - Durante la transformación de los parámetros **HNR**, **HammI** y **pe1000** se distribuye el factor de modificación β aplicado entre las bandas de frecuencia, asegurando que la energía total del habla resultante se mantiene constante durante la transformación, utilizando para ello el factor correctivo de la energía α (Sección 5.2.3).

La elección de qué parámetros se modificarían en cada una de las transformaciones, desde el estilo neutro hacia alguno del resto de los estilos de habla expresivos, se realizó a partir del estudio de discriminación (Sección 5.4), eligiendo aquellos parámetros que mostraban las características que mejor podían influir en la percepción del estilo destino. En la Tabla 5.16 se presenta la selección de parámetros implicados en cada una de las transformaciones y el factor β asociado, calculado como el porcentaje de variación entre el estilo destino y el neutro (origen). Esta selección se puede relacionar con los efectos acústicos aparecidos en la voz, presentando la utilidad de cada uno de ellos en la caracterización del estilo de habla expresivo concreto:

- Para todas las transformaciones se han empleado los parámetros HammI y pe1000, controlando el efecto de tensión en la voz, manifestándose como un esfuerzo o relajación en la fonación. Por ejemplo, los estilos alegre y agresivo manifiestan unos valores de estos parámetros que recalcan una elevada energía de la banda alta de frecuencia, apareciendo un efecto de esfuerzo en el habla resultante. Por el contrario, para los estilos sensual y triste ocurre lo contrario, siendo dos estilos con un habla más relajada.
- Los parámetros *jitter* y *shimmer* permiten controlar el efecto de temblor en la voz, de ahí que su uso sea más notable en los estilos sensual y triste.
- El control de la cantidad de ruido aparecido en el habla se lleva a cabo con el parámetro HNR, de especial utilidad durante la generación de habla en estilo sensual.

(%)	***Jitter***	***Shimmer***	**HNR**	**do1000**	**HammI**	**pe1000**
Alegre	- -	- -	- -	- -	-60	110
Sensual	175	85	-50	- -	155	-50
Agresivo	-20	-45	- -	- -	-70	220
Triste	-45	90	- -	- -	655	-75

Tabla 5.16: *Conjunto de parámetros de cualidad de la voz seleccionados para la transformación de estilos de habla expresivos neutro-destino y el % de factor original a ser aplicado para cada parámetro (con "- -" se indican los parámetros que no fueron utilizados en la transformación)*

Debido a que en la transformación de la VoQ se vio implicada toda la señal de voz de la frase y dado que el modelo de VoQ ha usado el promedio de los valores analizados sobre las vocales, sin tener en cuenta, entre otras, la información fonética o la posición dentro de la frase, algunos de los parámetros necesitaron de un ajuste de este factor, presentado en la Tabla 5.17. Estos ajustes se realizaron en base a pruebas heurísticas realizadas durante el proceso de diseño del sistema de transformación, mejorando la percepción del estilo generado y minimizando la distorsión debida a las modificaciones practicadas. Dichas pruebas heurísticas partieron de la configuración de parámetros y sus factores β esperados del proceso de modelado, con lo que se generaron ejemplos de síntesis para aquellos casos a mejorar (en torno a 5 ejemplos por configuración) empleando la configuración de HNM y el modelado de prosodia

a utilizar en el proceso de síntesis definitivo. Para cada estilo expresivo destino, en función de la percepción de la señal de voz generada por parte de los diseñadores del módulo de transformación (entre 1 y 4 miembros del equipo), los parámetros asociados al efecto acústico detectado fueron ajustados (p. ej. a partir del grado de ruido en el habla se ajustó el parámetro HNR).

(%)	***Jitter***	***Shimmer***	**HNR**	**do1000**	**HammI**	**pe1000**
Alegre	- -	- -	- -	- -	-20^*	300^*
Sensual	-70^*	-60^*	-95^*	- -	155	-95^*
Agresivo	-90^*	-90^*	- -	- -	-70	540^*
Triste	-70^*	-5^*	- -	- -	90^*	-75

Tabla 5.17: *Conjunto de parámetros de cualidad de la voz seleccionados para la transformación de estilos de habla expresivos neutro-destino y el $\%$ de factor finalmente aplicado para cada parámetro (con "$*$" se indica que se aplicó un ajuste sobre el factor y con "- -" se señalan los parámetros que no fueron utilizados en la transformación)*

En la Tabla 5.17, el símbolo "$*$" indica que el factor β fue ajustado respecto del original para maximizar su percepción y evitar la degradación de la calidad del habla, mientras que el símbolo "- -" señala que esos parámetros no están involucrados en la transformación. Hay que destacar los casos del *jitter* y del *shimmer*, que han sufrido fuertes modificaciones, incluso cambiando la tendencia de las mismas (descenso del valor del parámetro cuando el modelo inicial lo aumentaba respecto del estilo neutro) con respecto del factor β calculado, debido principalmente a que las distribuciones de los parámetros extraídas durante el análisis del corpus presentaban valores atípicos que sesgaban las estadísticas obtenidas (Sección 5.3.3). Este sesgo ha provocado que el modelo de VoQ requiriera valores de los parámetros mayores de los que realmente necesitaba, de modo que el efecto generado ha provocado una excesiva modificación de estos parámetros y con ella un descenso de la calidad. En cambio, para el resto de los parámetros implicados en las transformaciones, la necesidad de un ajuste ha estado ligado a la percepción final del habla destino sintetizada, modificando los factores a aplicar maximizando la percepción del estilo expresivo sin verse afectada la calidad final. Por tanto, del ajuste heurístico se desprende que una mayor calidad puede exigir una relajación en los factores aplicados sobre los parámetros durante la transformación.

Nótese que el parámetro do1000 no aparece en la selección de parámetros (Tabla 5.17). Esto se debe a que las transformaciones llevadas a cabo mediante la modificación de este parámetro no obtuvieron los resultados esperados, ya que la aparición de artefactos indeseados en el habla final dieron lugar a una voz desagradable y poco natural.

5.5.4. Evaluación de la transformación de estilos de habla expresivos

La transformación de estilos de habla expresivos durante la SHE fue evaluada, utilizando para ello la plataforma web TRUE (Planet et al., 2008), mediante el

planteamiento de dos enfoques. El primero, destinado a la evaluación del efecto de introducir VoQ en el proceso de transformación expresiva basado en prosodia, con el que se analizó la eficacia de la transformación en cuanto a si la VoQ incrementaba la percepción del estilo expresivo destino respecto de la solución que sólo utiliza la prosodia (Sección 5.5.4.1).

Debido a que el primer enfoque sólo estaba destinado a la evaluación de la preferencia del uso de la VoQ, sin aportar información acerca de la calidad del habla generada ni de la identificación real de los estilos de habla expresivos que ella transmite, se optó por realizar un segundo test, enfocado por tanto a la evaluación de la calidad y a la identificación del estilo expresivo generado (Sección 5.5.4.2). En este caso no se evaluaron únicamente los resultados para la metodología de transformación de la prosodia y de la VoQ usando HNM, sino que se tuvieron en cuenta configuraciones de voz natural, de resíntesis y de modificación de prosodia y de VoQ, usando los algoritmos PSOLA y HNM, de forma que se pudiera evaluar la variación de la calidad y de la tasa de identificación subjetiva de los estilos expresivos para cada una de ellas.

5.5.4.1. Evaluación del efecto de la cualidad de la voz en la transformación

Para evaluar el efecto de introducir VoQ en la transformación de estilos de habla expresivos basada en prosodia, se llevó a cabo un test perceptivo del tipo CMOS (ITU-P.800, 1996), comparando la modificación de únicamente la prosodia con la modificación conjunta de la prosodia y de la VoQ.

En esta evaluación, se eligieron 8 enunciados neutros del corpus (Apéndice C.2). Para cada uno se realizó un análisis mediante HNM, y tanto la prosodia como la VoQ fueron modelados y modificados a partir del estilo de habla destino deseado. La transformación para cada estilo (alegre, sensual, agresivo y triste) resultó en 32 nuevos enunciados, por un lado con la transformación de la prosodia y por el otro con la transformación conjunta de la prosodia y de la VoQ, resultando en un total de 64 nuevos enunciados diferentes.

Para cada par de nuevos enunciados (con las configuraciones de prosodia y de prosodia más VoQ), se realizó la siguiente pregunta al evaluador: *"¿Cuál de los enunciados presentados cree que representa con mayor intensidad el estilo expresivo indicado?"*. A los evaluadores se les presentaba el estímulo sonoro acompañado de la información sobre el estilo de habla expresivo reproducido, forzando así la comparativa entre ambas configuraciones sin errores debidos a una posible confusión del estilo percibido. Cada evaluador contestó de acuerdo a sus preferencias a la pregunta formulada, pudiendo elegir entre 7 posibles respuestas basadas en el test CMOS: prefería un enunciado "mucho más", "más", "ligeramente más" o "igual" que el otro, con puntuaciones de 3, 2, 1, 0, -1, -2 y -3. Los valores positivos fueron asignados a aquellos en los que la VoQ incrementaba la percepción del estilo de habla expresivo y los negativos para el caso contrario. Un total de 15 voluntarios participaron en el test, 5 mujeres y 10 hombres de edades comprendidas entre los 24 y los 50 años, de los que el $40\,\%$ eran expertos en tecnologías del habla. Una vez finalizado el test se permitió

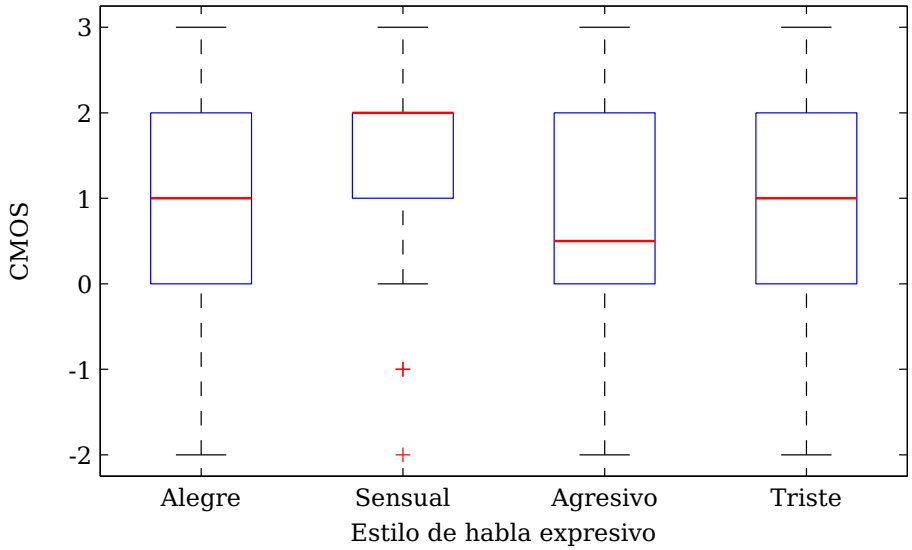


Figura 5.10: *Resultados de la prueba CMOS para la transformación de estilos de habla expresivos usando prosodia y prosodia más cualidad de la voz*

al evaluador escribir comentarios sobre el mismo, recopilando así información sobre el criterio que cada evaluador había tomado y opiniones acerca de la dificultad y de la calidad de los enunciados.

Los resultados obtenidos, presentados en los *boxplots* de la Figura 5.10, demuestran que la preferencia es la de añadir la parametrización de VoQ a la de la prosodia durante el proceso de transformación. A pesar de que la VoQ mejora la percepción de todos los estilos expresivos, existen dificultades en algunos casos, la cual se presume que es debido principalmente a las razones que se exponen a continuación:

1. Cada evaluador tiene un criterio distinto a la hora de decidir la intensidad con la que percibe un estilo expresivo.

2. El contenido semántico del enunciado afecta inconscientemente a su percepción.

3. La modificación de los parámetros del habla puede producir su degradación, de forma que los evaluadores tienden a preferir una voz en la que no existe degradación, a pesar de que el estilo expresivo no sea percibido correctamente.

Una vez se dispuso de los resultados del test se llevó a cabo un análisis de la mediana y de los intervalos de confianza (Tabla 5.18) utilizando el test de Wilcoxon (Apéndice F.6), con el que se obtuvieron los valores de mediana y el margen en el que esta se mueve, pudiendo comprobar los beneficios del uso de VoQ junto con la prosodia. Estos resultados indican que el uso de VoQ junto a la prosodia se prefiere significativamente a usar solamente prosodia. Esto es interesante ya que se da tanto

en aquellos casos donde la prosodia ya obtenía buenos resultados (triste) como en aquellos donde el uso únicamente de prosodia presentaba problemas para la identificación del estilo expresivo (agresivo). Además, para el caso sensual, se dio el mejor resultado con una mediana igual a 2, con una clara preferencia por el uso de VoQ, mientras que el resto de estilos de habla expresivos mostraron un valor de la mediana de la distribución entre $0,5$ (agresivo) y 1 (alegre y triste), decantando la preferencia en todos los casos hacia el uso conjunto de prosodia y VoQ. Finalmente, para todos los estilos de habla expresivos, el intervalo de confianza muestra una dispersión (para un nivel de confianza del $95\,\%$) que también refuerza la conclusión que los mejores resultados fueron obtenidos usando VoQ, mostrando la preferencia por el uso conjunto de prosodia y VoQ (Tabla 5.18).

		Alegre	**Sensual**	**Agresivo**	**Triste**
Mediana		1	2	$0,5$	1
Intervalo de confianza	**Mínimo**	1	$1,5$	$0,5$	$0,5$
	Máximo	$1,5$	2	1	$1,5$

Tabla 5.18: *Mediana e intervalo de confianza para la evaluación de la transformación de estilos de habla expresivos (nivel de confianza del $95\,\%$)*

Finalmente, para entender mejor los resultados obtenidos, se puede relacionar el uso de la VoQ con el efecto que produce sobre el habla:

- El estilo sensual se ha visto mejorado por usar un efecto de susurro sobre la voz.
- Para los estilos alegre y agresivo se consiguieron mejores resultados cuando se hizo presente en la voz un efecto de tensión.
- El estilo triste mejoró sus resultados mediante la introducción de un efecto de voz temblorosa.

5.5.4.2. Evaluación de la calidad y de la eficacia de la transformación

Vista la preferencia en el uso de la VoQ junto a la prosodia durante la transformación de estilos de habla expresivos (Sección 5.5.4.1), el siguiente paso fue preguntarse cómo afectaba la transformación de los parámetros sobre la calidad del habla generada, y si la identificación de los estilos destino justificaba estas transformaciones.

Con el objetivo de analizar la eficacia de la metodología de transformación propuesta (Sección 5.5.3), se evaluaron la calidad del habla generada y la tasa de identificación de los estilos expresivos destino. Los resultados fueron comparados con una serie de configuraciones, tanto de parámetros a ser transformados (solamente prosodia, de prosodia más *jitter* y *shimmer* y, por último, prosodia más la mejor combinación de parámetros de VoQ seleccionada por transformación) como de algoritmos de generación del habla (PSOLA y HNM). Estos algoritmos de generación del habla siguieron la implementación realizada por la herramienta Praat (Boersma, 2001) para el PSOLA y la de Calzada (2008, 2010) para el HNM. A continuación se describe cada

una de las configuraciones bajo evaluación, donde el modelado de la prosodia y de la VoQ se ha realizado a partir de la metodología presentada en la Sección 5.5.3 e indicando el nombre que las identifican en los experimentos:

1. **'Natural'**: habla natural. Ejemplos directamente extraídos del corpus.

2. **'Resint_PSOLA'** y **'Resint_HNM'**: resíntesis basada en PSOLA y en HNM respectivamente. A partir de ejemplos del corpus se ha realizado el proceso de análisis y de síntesis sin aplicar ninguna modificación sobre los parámetros del habla.

3. **'PSOLA_Pros'** y **'HNM_Pros'**: transformación de la prosodia basada en PSOLA y en HNM respectivamente. Enunciados expresados originariamente en estilo neutro fueron transformados a partir de los modelos de prosodia.

4. **'PSOLA_Pros_Ji_Sh'** y **'HNM_Pros_Ji_Sh'**: transformación de la prosodia, del *jitter* y del *shimmer*, basada en PSOLA y en HNM respectivamente. Partiendo de modelos de prosodia y con los valores destino conocidos para el *jitter* y el *shimmer*, se transforman estos parámetros de enunciados expresados en estilo neutro.

5. **'HNM_Pros_VoQ'**: modificación de la prosodia y de la VoQ basada en HNM. En base a modelos de prosodia y de configuraciones de VoQ que caracterizan a los estilos de habla expresivos bajo estudio, se transforman enunciados que originariamente estaban expresados en estilo neutro. La configuración de parámetros de VoQ dependerá en cada caso del estilo expresivo destino.

En primer lugar, se evaluó la calidad del habla generada mediante un corpus en estilo neutro, usando transformación hacia alegre, sensual, agresivo y triste. La evaluación se realizó mediante un test MOS (ITU-P.800, 1996) con cinco posibles respuestas, asociadas a una puntuación entre 1 y 5, siendo 5 la máxima calidad y 1 la mínima: "Excelente" (5), "Buena" (4), "Regular" (3), "Mediocre" (2) y "Mala" (1). Las configuraciones con las que se comparan los resultados son de dos tipos:

1. Habla natural (caso 1) y resintetizada (caso 2), indicando los valores máximos de referencia por ser casos reales (natural) y los mejores resultados posibles que un sistema de CTH empleando los algoritmos PSOLA o HNM pueden conseguir (resíntesis).

2. También se comparó la metodología de interés (caso 'HNM_Pros_VoQ') con el resto de configuraciones donde existe la transformación de prosodia y, de forma común a todos los estilos expresivos, los parámetros *jitter* y *shimmer*, observándose los cambios de la calidad entre los algoritmos PSOLA y HNM bajo las mismas condiciones de transformación.

En segundo lugar se realizó la evaluación de la identificación del estilo de habla expresivo destino, mediante un test con 5 posibles respuestas, 4 de ellas correspondientes a los estilos destino: alegre, sensual, agresivo y triste, y una quinta categoría correspondiente a "Otros". Con la categoría de "Otros", sin asignarse a ningún estilo en particular, se evitaba desviar la opinión hacia el resto de opciones cuando no estaba clara la respuesta o cuando no se percibía ninguno de los estilos evaluados. De forma análoga a la calidad, la configuración de interés se comparó con el resto de opciones posibles, pudiendo comprobar así las diferencias entre el uso de PSOLA y HNM, y de unos u otros parámetros del habla (prosodia y VoQ). Asimismo, el caso de la transformación de prosodia, ya sea aisladamente o junto a *jitter* y *shimmer*, permitió comparar la eficacia de los algoritmos PSOLA y HNM bajo las mismas condiciones.

El hecho de comparar los parámetros de prosodia y de VoQ es de interés en cuanto a que se pueden conocer las variaciones de la calidad para una determinada tasa de identificación de los estilos de habla expresivos, pudiéndose dar principalmente dos situaciones sobre las que se puede tomar medidas. En el caso de una alta calidad y baja tasa de identificación, nos encontraremos en una situación donde es necesario mejorar el modelado de los estilos expresivos (más parámetros o bien diferentes), y en el caso de baja calidad y alta tasa de identificación debe de plantearse qué algoritmo es preferible (PSOLA o HNM), a partir de analizar si este soporta la transformación de parámetros deseada y si esta transformación debe ser tan exigente que pueda provocar una degradación de la calidad (pudiendo trabajar en su perfeccionamiento para adecuarlo a las necesidades de la transformación).

El test de evaluación de la calidad y de la identificación de los estilos de habla expresivos, se llevó a cabo realizando dos consultas para el mismo enunciado presentado: "*Valore la calidad global del audio*" para la evaluación de la calidad del habla e "*Indique qué estilo expresivo le transmite el audio*" para la identificación del estilo expresivo transmitido. Se generaron 160 enunciados para su evaluación (Apéndice C.2), que contenían el mismo número de casos para cada una de las configuraciones y de los estilos de habla expresivos, de los cuales se seleccionó un grupo de 10 de ellos que fueron mostrados a los evaluadores al inicio de la prueba para que se familiarizaran con los audios que iban a escuchar, sin que estos tuvieran conocimiento de esta finalidad. El número total de evaluadores que participaron en el test fue de 17, de edades comprendidas entre 24 y 50 años, entre los que había 13 hombres y 4 mujeres, siendo un $47\,\%$ de ellos expertos en tecnologías del habla.

Llegados a este punto pasemos a ver los resultados obtenidos y las conclusiones extraídas. Se empieza mostrando, para cada una de las configuraciones y estilos de habla expresivos, los resultados de la calidad para posteriormente pasar a presentar los de identificación subjetiva.

En la Figura 5.11 se muestran los resultados, en forma de *boxplot*, de calidad de las configuraciones que han servido de referencia: 'Natural' (Figura 5.11a), 'Resint_PSOLA' (Figura 5.11b) y 'Resint_HNM' (Figura 5.11c). Se observa la elevada calidad de estos enunciados ("Excelente"), donde únicamente el estilo triste, con resíntesis basada en HNM, tiene cierta dispersión en los resultados. Estos resultados son coherentes con las configuraciones empleadas, dado que, por un lado, en la resíntesis

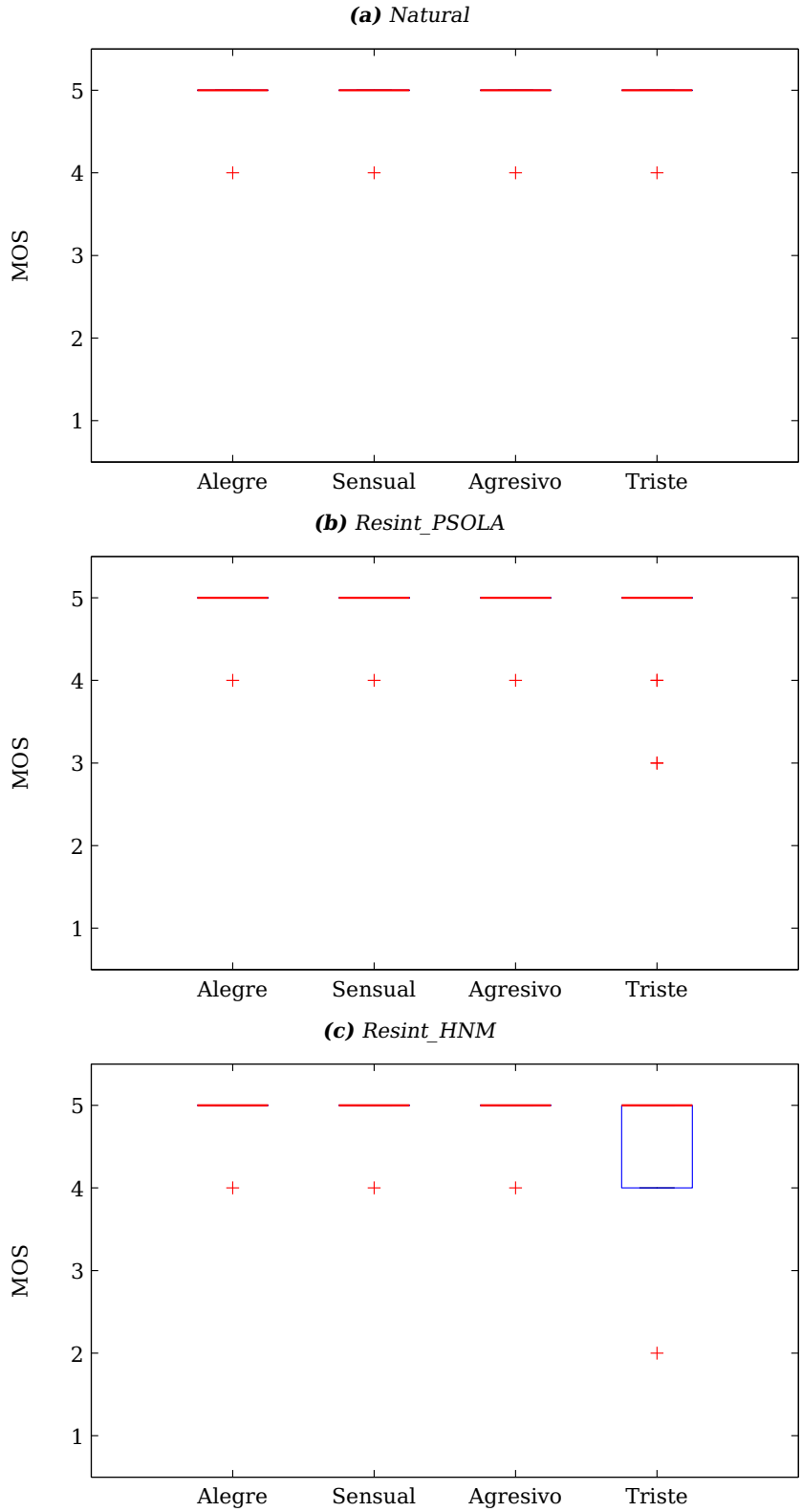


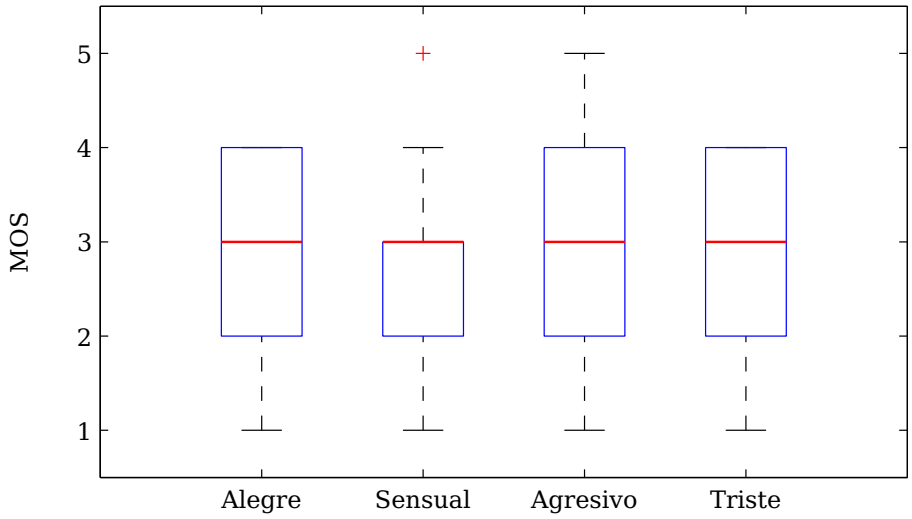
Figura 5.11: *Resultado del test MOS de la calidad para las configuraciones de habla natural, resíntesis usando PSOLA y resíntesis usando HNM*

usando PSOLA el algoritmo genera prácticamente la misma señal que si se tratara de voz natural. Por tanto, los resultados de la calidad para habla natural y resíntesis con PSOLA deberían de ser similares, tal y como se observa en los resultados (Figuras 5.11a y 5.11b). Por otra parte, en términos generales la síntesis usando HNM depende de la parametrización del habla, por lo que le afectan errores graves de las marcas de *pitch* más que al PSOLA, así como también los valores de parámetros del HNM asociados a la componente determinista y estocástica. Es por esta razón, que la resíntesis basada en HNM puede dar peores resultados, aunque próximos al natural y de gran calidad, ya que dependiendo del estilo expresivo podrían existir mayores errores en su parametrización. Este podría ser el caso del estilo triste (Figura 5.11c), que dadas las características acústicas que presenta, como por ejemplo temblor en la voz, pudo dar pie a mayores dificultades de análisis y por consiguiente de síntesis.

A partir de los valores de referencia para la calidad, se analizan los resultados para cada configuración evaluada hasta llegar a la de interés, siendo esta la metodología de transformación utilizando HNM junto con modificación de los parámetros de prosodia y de VoQ ('HNM_Pros_VoQ'). Primero se presentan los valores de la calidad para las configuraciones 'PSOLA_Pros' (Figura 5.12a) y 'HNM_Pros' (Figura 5.12b), en los que se observa una cierta degradación de la calidad por haber aplicado un modelo de prosodia expresiva sobre enunciados neutros. Se puede comprobar como el valor de la calidad se centra en "Regular" para todos los estilos, presentando una mejor respuesta para el caso de usar HNM. Esto se debe a que HNM muestra una mayor flexibilidad y estabilidad a la hora de realizar la síntesis de la señal, ya que no la reconstruye a partir de información acústica (como PSOLA) sino a partir de un modelo paramétrico a partir de esta, con una mayor capacidad de control de los parámetros transformados y evitando así la aparición de artefactos u otros elementos no deseados en el habla final generada.

Vistos los resultados cuando se ve implicada únicamente la transformación de prosodia, veamos qué ocurre cuando dos parámetros de VoQ, el *jitter* y el *shimmer*, se transformaron empleando tanto PSOLA como HNM. Para ello se analizan los valores resultantes de la evaluación de las configuraciones 'PSOLA_Pros_Ji_Sh' (Figura 5.13a) y 'HNM_Pros_Ji_Sh' (Figura 5.13b), de forma que se pueden extraer conclusiones sobre el mejor enfoque (PSOLA o HNM) para realizar las transformaciones. Lo primero que se observa es que la calidad decae respecto de transformar solamente prosodia, especialmente para el caso de PSOLA, ya que se ha aplicado la transformación de más parámetros del habla (VoQ) aumentando así la aparición de elementos no deseados (PSOLA) y exigiendo más modificaciones a los parámetros del HNM. A pesar de ello, el valor de la calidad para HNM se mantiene prácticamente estable en mediana ("Regular"), excepto para el caso de agresivo ("Mediocre"). Asimismo, aunque en comparación el HNM dio mejores resultados que PSOLA, el descenso general de la calidad también puede asociarse a que, en la modificación del *jitter* y del *shimmer*, no se ha tenido en cuenta el papel que estos juegan en la transformación de estilos expresivos de forma independiente al resto de parámetros de VoQ. Aun así, HNM se perfila como buena opción en la transformación de los estilos de habla expresivos, ofreciendo una buena calidad en el habla resultante.

(a) *PSOLA_Pros*



(b) *HNM_Pros*

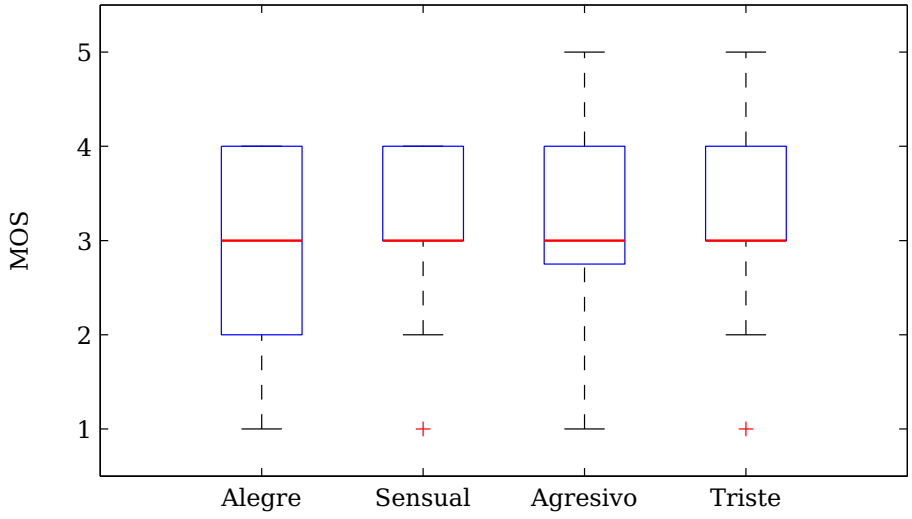
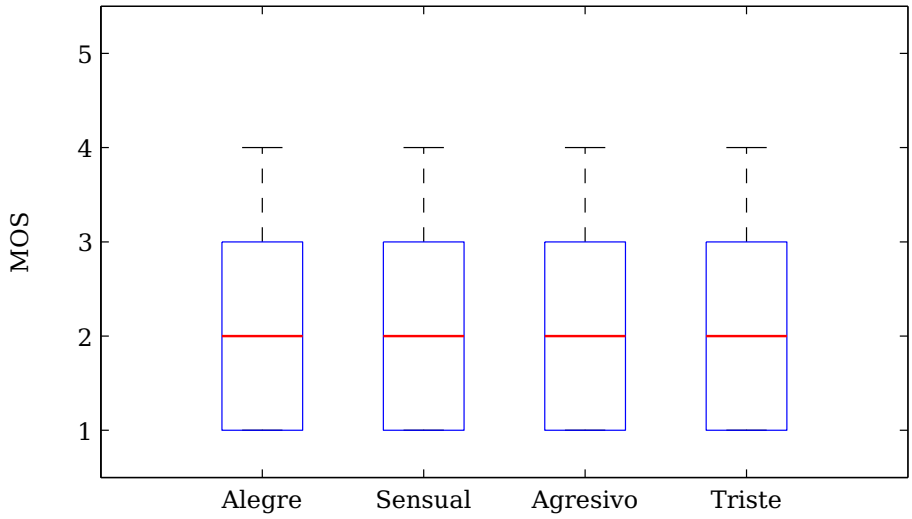


Figura 5.12: *Resultado del test MOS de la calidad para las configuraciones de transformación de prosodia usando PSOLA y HNM*

(a) *PSOLA_Pros_Ji_Sh*



(b) *HNM_Pros_Ji_Sh*

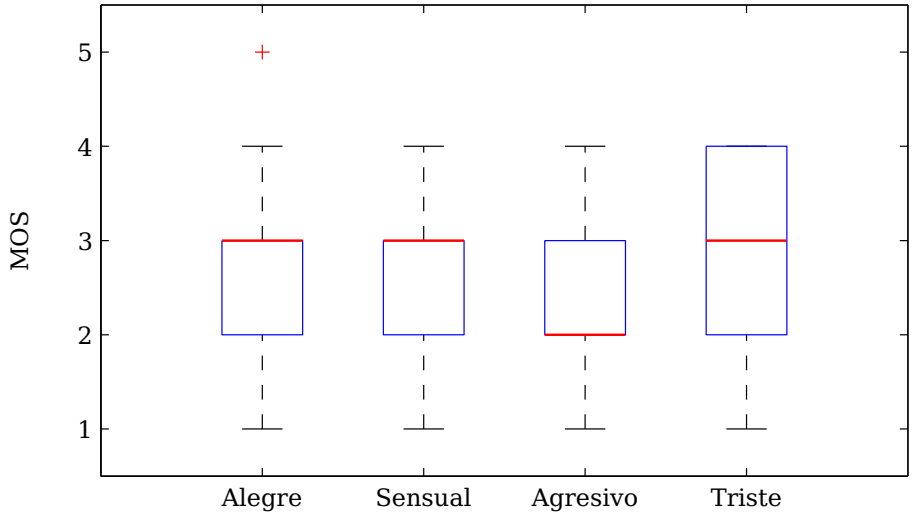


Figura 5.13: *Resultado del test MOS de la calidad para las configuraciones de transformación de prosodia, jitter y shimmer usando PSOLA y HNM*

Por último, solamente queda evaluar la calidad que obtuvo la configuración 'HNM_Pros_VoQ' (Figura 5.14), donde la transformación de parámetros se ajustaba a las necesidades de cada uno de los estilos de habla expresivos destino (Tabla 5.17). En este caso no se ha considerado la posibilidad de aplicar la transformación usando PSOLA, ya que como se comentó en la Sección 5.5.2.3 sería necesario añadir alguna técnica que permitiera la manipulación espectral y el desglose de la señal en componentes armónica y de ruido (p. ej. usando STRAIGHT (Kawahara et al., 1999)).

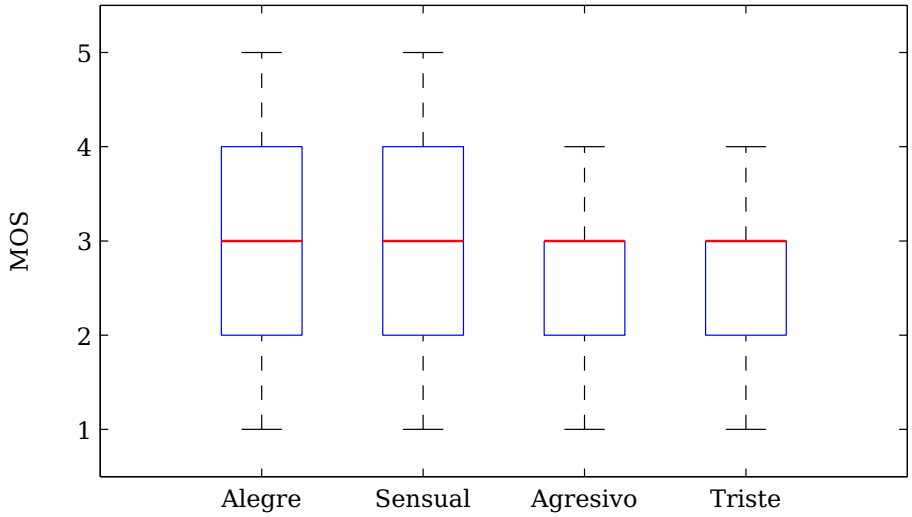


Figura 5.14: *Resultado del test MOS de la calidad para la configuración de transformación de prosodia y cualidad de la voz usando HNM ('HNM-ProVoQ')*

La Figura 5.14 muestra como la calidad se mantiene en límites aceptables ("Regular") tal y como había venido pasando para la configuración donde se había visto envuelto el HNM únicamente con la modificación de la prosodia (Figura 5.12b), hecho que demuestra que HNM permite realizar las transformaciones del habla sin reducir la calidad. Asimismo, se observa como en este caso, la calidad para los estilos alegre, sensual y agresivo ha aumentado respecto de la transformación de la prosodia junto con únicamente el *jitter* y *shimmer* (Figura 5.13b), apreciándose un pequeño descenso en el estilo triste, hecho que ya ocurría en los casos de referencia vistos en la Figura 5.11. Con esto, se observa como el uso de la parametrización y transformación de VoQ permite mejorar la percepción y la calidad de la síntesis expresiva. No obstante, un exceso de manipulación de la señal puede conllevar también efectos negativos, dando pie a una disminución en la calidad, como por ejemplo es el estilo triste, que como se muestra en Tabla 5.22 presentó la mejor identificación del estilo destino aunque no obtuvo la mejor calidad.

Una vez analizados los resultados obtenidos para la evaluación de la calidad, se presentan los equivalentes al análisis de la identificación subjetiva de cada uno de los estilos de habla expresivos destino. El objetivo en este caso fue similar al de la calidad, ya que se deseaba comprobar si la configuración 'HNM_Pros_VoQ' obtenía las mejores o peores tasas de identificación. Recordemos que la transformación se realizó desde

el estilo neutro hacia otro que denote expresividad (alegre, sensual, agresivo y triste) y, dado que podrían existir dudas en la asignación del estilo percibido, también se incluyó en el test la opción de "Otros", evitando así desviar la medida hacia ninguno de ellos.

Natural	(%)	**Alegre**	**Sensual**	**Agresivo**	**Triste**	**Otros**
	Alegre	$94,1$	$0,0$	$5,9$	$0,0$	$0,0$
	Sensual	$0,0$	$100,0$	$0,0$	$0,0$	$0,0$
	Agresivo	$0,0$	$0,0$	$100,0$	$0,0$	$0,0$
	Triste	$0,0$	$9,4$	$0,0$	$90,6$	$0,0$
	(***F*1**)	*0, 97*	*0, 96*	*0, 97*	*0, 95*	- -
Resint_PSOLA	(%)	**Alegre**	**Sensual**	**Agresivo**	**Triste**	**Otros**
	Alegre	$96,5$	$0,0$	$3,5$	$0,0$	$0,0$
	Sensual	$0,0$	$97,6$	$0,0$	$2,4$	$0,0$
	Agresivo	$1,2$	$0,0$	$98,8$	$0,0$	$0,0$
	Triste	$0,0$	$11,8$	$0,0$	$88,2$	$0,0$
	(***F*1**)	*0, 98*	*0, 93*	*0, 98*	*0, 93*	- -
Resint_HNM	(%)	**Alegre**	**Sensual**	**Agresivo**	**Triste**	**Otros**
	Alegre	$87,1$	$0,0$	$11,8$	$0,0$	$1,2$
	Sensual	$1,2$	$95,3$	$1,2$	$2,4$	$0,0$
	Agresivo	$1,2$	$0,0$	$98,8$	$0,0$	$0,0$
	Triste	$0,0$	$11,8$	$0,0$	$88,2$	$0,0$
	(***F*1**)	*0, 92*	*0, 92*	*0, 93*	*0, 93*	- -

Tabla 5.19: *Matrices de confusión (%) y medidas $F1$, en la identificación de estilos de habla expresivos, para las configuraciones de referencia 'Natural', 'Resint_PSOLA' y 'Resint_HNM'*

Los resultados de identificación subjetiva se presentan mediante una matriz de confusión (%) y la medida $F1$ (Apéndice F.7). Primero, la matriz de confusión da una idea de lo buena que ha sido la identificación y, sobre todo, permite detectar confusiones entre estilos. Segundo, la medida $F1$ ha sido la de interés, por dar una visión más real y compacta de lo buena que ha sido la identificación, teniendo en cuenta tanto los casos bien identificados como la confusión existente entre ellos, ya sea por no identificar ese estilo cuando realmente lo era (relacionado con la cobertura) o por identificarlo cuando en realidad se trataba de otro (relacionado con la precisión).

Tal y como se hizo para la evaluación de la calidad, primero se muestra la referencia para los mejores casos posibles, que son los de habla natural y los de resíntesis usando PSOLA y HNM (Tabla 5.19). Se observa como las tres configuraciones dieron resultados similares tal y como era de esperar, existiendo confusión entre los estilos alegre-agresivo y sensual-triste, siendo más acusada para el caso de ambas resíntesis.

El siguiente paso es analizar la transformación del estilo neutro hacia los estilos expresivos destino, viendo la identificación conseguida para cada una de las configuraciones implicadas, las limitaciones de cada una de ellas y la mejora en la percepción de los estilos expresivos debida al uso de parámetros de VoQ respecto a usar únicamente prosodia. Para ello, en primer lugar se presentan los resultados de

la identificación para las configuraciones 'PSOLA_Pros' y 'HNM_Pros' (Tabla 5.20), en las que se hizo la primera tentativa de transformación a partir de parámetros de prosodia. En segundo lugar se presentan los resultados para los casos de 'PSOLA_Pros_Ji_Sh' y 'HNM_Pros_Ji_Sh' (Tabla 5.21), donde se compara la transformación mediante PSOLA y HNM utilizando parámetros de prosodia y los mismos parámetros de VoQ para todas las transformaciones (*jitter* y *shimmer*), pudiendo así hacer una doble comparativa bajo las mismas condiciones, por un lado de método de transformación y síntesis y por el otro el beneficio de haber incluido más parametrización del habla junto con la prosodia. Finalmente, se muestran los resultados para la configuración de interés, la 'HNM_Pros_VoQ' (Tabla 5.22), observando la identificación final obtenida por la transformación de prosodia y de las configuraciones de VoQ que caracterizaban a los estilos expresivos destino, así como los porcentajes de mejora de $F1$ respecto de todas las demás configuraciones (Tabla 5.23).

PSOLA_Pros	(%)	**Alegre**	**Sensual**	**Agresivo**	**Triste**	**Otros**
	Alegre	$38,8$	$0,0$	$8,2$	$7,1$	$45,9$
	Sensual	$4,7$	$1,2$	$9,4$	$27,1$	$57,6$
	Agresivo	$43,5$	$1,2$	$12,9$	$18,8$	$23,5$
	Triste	$4,7$	$1,2$	$2,4$	$49,4$	$42,4$
	(F1)	$0,40$	$0,02$	$0,19$	$0,49$	- -
HNM_Pros	(%)	**Alegre**	**Sensual**	**Agresivo**	**Triste**	**Otros**
	Alegre	$30,6$	$3,5$	$17,6$	$28,2$	$20,0$
	Sensual	$1,2$	$31,8$	$9,4$	$40,0$	$17,6$
	Agresivo	$35,3$	$1,2$	$21,2$	$23,5$	$18,8$
	Triste	$2,4$	$18,8$	$10,6$	$41,2$	$27,1$
	(F1)	$0,36$	$0,41$	$0,27$	$0,35$	- -

Tabla 5.20: *Matrices de confusión (%) y medidas $F1$, en la identificación de estilos de habla expresivos, para las configuraciones 'PSOLA_Pros' y 'HNM_Pros'*

En la Tabla 5.20 se presenta la primera configuración de transformación, usando solamente prosodia ('PSOLA_Pros' y 'HNM_Pros'). Analizando los valores de $F1$ se observa como la identificación, para ambos casos, decae drásticamente para todos los estilos respecto los de referencia (Tabla 5.19). Hay que hacer una mención especial al caso del estilo sensual utilizando PSOLA, donde se puede ver una identificación prácticamente inexistente ($F1 = 0,02$), pero en cambio para HNM se consiguió un resultado mejor al del resto de los estilos ($F1 = 0,41$). En cuanto a la comparativa entre PSOLA y HNM, mientras PSOLA presentaba ligeros mejores resultados que HNM para los estilos alegre y triste, lo contrario pasa para sensual y agresivo. Los mejores resultados para HNM (mejora del $1950\,\%$ para sensual y del $42,1\,\%$ para agresivo) han podido ser debidos a la menor degradación que esta metodología provoca durante el proceso de generación del habla tras la transformación, obteniendo por tanto una mejor calidad (Figura 5.12b) que repercutiría en una mejor percepción de los estilos.

Mostrados los resultados para la primera configuración de la transformación, pasemos a ver qué ocurrió al añadir dos parámetros de VoQ: el *jitter* y el *shimmer*

PSOLA_Pros_Ji_Sh	(%)	**Alegre**	**Sensual**	**Agresivo**	**Triste**	**Otros**
	Alegre	$42,4$	$0,0$	$15,3$	$14,1$	$28,2$
	Sensual	$9,4$	$9,4$	$8,2$	$27,1$	$45,9$
	Agresivo	$41,2$	$0,0$	$18,8$	$14,1$	$25,9$
	Triste	$10,6$	$3,5$	$9,4$	$32,9$	$43,5$
	$(\boldsymbol{F1})$	$\mathit{0,42}$	$\mathit{0,17}$	$\mathit{0,25}$	$\mathit{0,35}$	- -
HNM_Pros_Ji_Sh	(%)	**Alegre**	**Sensual**	**Agresivo**	**Triste**	**Otros**
	Alegre	$30,6$	$3,5$	$16,5$	$27,1$	$22,4$
	Sensual	$3,5$	$30,6$	$4,7$	$36,5$	$24,7$
	Agresivo	$32,9$	$2,4$	$18,8$	$22,4$	$23,5$
	Triste	$4,7$	$17,6$	$1,2$	$58,8$	$17,6$
	$(\boldsymbol{F1})$	$\mathit{0,36}$	$\mathit{0,40}$	$\mathit{0,27}$	$\mathit{0,48}$	- -

Tabla 5.21: *Matrices de confusión (%) y medidas $F1$, en la identificación de estilos de habla expresivos, para las configuraciones 'PSOLA_Pros_Ji_Sh' y 'HNM_Pros_Ji_Sh'*

('PSOLA_Pros_Ji_Sh' y 'HNM_Pros_Ji_Sh'). En este caso, tal y como se explicó para el test de la calidad, ambos parámetros fueron transformados sin tener en cuenta su comportamiento en la transformación y su dependencia con otros parámetros de VoQ. Simplemente se analizó si el hecho de añadir nuevas parametrizaciones, siendo las mismas para PSOLA y HNM, ayudaban en el proceso de identificación y si este dependía del algoritmo de generación de habla empleado. En la Tabla 5.21 se presentan los resultados obtenidos, observándose como las medidas de identificación han mejorado, respecto al uso de sólo prosodia, en el caso PSOLA (sensual y agresivo) y mantenido estables para HNM (excepto para triste ($F1 = 0,48$) en que ha mejorado). Las mejoras de PSOLA, aunque se mantienen en niveles bajos de identificación, han sido debidas a la mejor caracterización de los estilos destino, mientras que para el caso triste (donde ha habido un descenso de la identificación) ha podido ser debido a un exceso de modificación de los parámetros, provocando mayores artefactos que conllevan a la confusión entre estilos (especialmente se ve un aumento de la clasificación de la categoría "Otros" y, en general, una mayor confusión con el resto de estilos). En cuanto al HNM, ocurre que mejora precisamente en el estilo triste ($37,1\,\%$), donde coincide con el máximo de la calidad para esta configuración (Figura 5.13b), posiblemente debido a la estabilidad que el HNM presenta ante la transformación de parámetros, pudiendo caracterizar cada estilo sin añadir tanta distorsión como PSOLA.

A partir de estos resultados ya sólo queda el análisis de la última configuración ('HNM_Pros_VoQ'), donde tanto los parámetros de prosodia como diferentes configuraciones de VoQ se vieron envueltas en la transformación de estilos de habla expresivos (Tabla 5.22). La primera observación a destacar es la obtención de los mejores resultados respecto a todas las configuraciones (Tabla 5.23) en las que se vieron envueltas transformaciones de parámetros ('PSOLA_Pros', 'HNM_Pros', 'PSOLA_Pros_Ji_Sh' y 'HNM_Pros_Ji_Sh'). El segundo punto a destacar es el buen resultado obtenido para

HNM_Pros_VoQ	(%)	Alegre	Sensual	Agresivo	Triste	Otros
	Alegre	$34,1$	$3,5$	$25,9$	$14,1$	$22,4$
	Sensual	$0,0$	$40,0$	$5,9$	$34,1$	$20,0$
	Agresivo	$23,5$	$2,4$	$31,8$	$16,5$	$25,9$
	Triste	$3,5$	$20,0$	$1,2$	$64,7$	$10,6$
	$(\boldsymbol{F1})$	$\boldsymbol{0,42}$	$\boldsymbol{0,48}$	$\boldsymbol{0,39}$	$\boldsymbol{0,56}$	- -

Tabla 5.22: *Matrices de confusión (%) y medidas $F1$, en la identificación de estilos de habla expresivos, para la configuración 'HNM_Pros_VoQ'*

el estilo triste ($F1 = 0,56$), seguido por el del sensual ($F1 = 0,48$), de especial importancia teniendo en cuenta las confusiones que existían ya en la referencia entre ambos casos. El valor para el estilo alegre ($F1 = 0,42$) también ha resultado interesante, sobre todo viendo la progresión respecto del uso de solamente prosodia y del *jitter* y *shimmer*. Por último, el estilo agresivo es el que ha conseguido los peores resultados ($F1 = 0,39$), aunque como punto a favor hay que decir que ha sido el que ha obtenido el mayor incremento en su identificación (tal y como muestra la Tabla 5.23 ha sido una mejora del $44,4\,\%$). Una conclusión que se desprende de los resultados, tanto a partir de la calidad como de la identificación, es que cuanto mayor ha sido la mejora en la caracterización (agresivo y triste) respecto del uso de prosodia, peores han sido los niveles de la calidad (Figura 5.14), siguiendo el planteamiento de que habrá que llegar a un compromiso entre la calidad esperada y la cantidad de modificación de los parámetros de VoQ que exija la percepción del estilo expresivo deseado. En cambio, donde los parámetros no han implicado una distorsión tan elevada de la señal y, por tanto, la calidad se ha mantenido más constante (p. ej. alegre y sensual), el incremento de la percepción no ha sido tan importante.

(%)	Alegre	Sensual	Agresivo	Triste
PSOLA_Pros	$5,0$	$2300,0$	$105,3$	$14,3$
HNM_Pros	$16,7$	$17,1$	$44,4$	$60,0$
PSOLA_Pros_Ji_Sh	$0,0$	$182,4$	$56,0$	$60,0$
HNM_Pros_Ji_Sh	$16,7$	$20,0$	$44,4$	$16,7$

Tabla 5.23: *Porcentaje de la mejora de $F1$ (%) usando 'HNM_Pros_VoQ' respecto al resto de configuraciones de transformación*

El motivo principal del error en la identificación es la confusión existente entre los estilos alegre-agresivo y sensual-triste, que ya se daba en los valores de referencia, y en la confusión de los estilos hacia triste y la categoría de "Otros". A partir de los comentarios que los evaluadores podían realizar al terminar el test se vio como "Otros" correspondía en general a la detección del estilo neutro (estilo de origen), cosa que hace plantear la necesidad de una mayor modificación de los valores de parámetros durante la transformación. Gracias a la estabilidad que ha demostrado el uso de HNM, tanto en la calidad (Figuras 5.12b, 5.13b y 5.14) como en la identificación (Tablas 5.20, 5.21 y 5.22), se pueden plantear estas transformaciones con

mayores garantías de éxito. Asimismo, en base a las observaciones de los evaluadores, también se ha detectado que el oyente tiene la tendencia a identificar el estilo transmitido por el enunciado en función de su contenido semántico, desviando de este modo la medida hacia estilos incorrectos, especialmente en aquellos casos en los que sus características acústicas no los identifique claramente (como por ejemplo es el caso de una voz susurrante en el estilo sensual o de una voz temblorosa en el triste).

A modo de resumen, se pueden extraer las siguientes conclusiones sobre la viabilidad de la transformación de estilos de habla expresivos, tanto desde el punto de vista de su calidad como de su identificación:

1. El uso de análisis y de síntesis del habla mediante HNM permite obtener buena calidad y control de las transformaciones.
2. La combinación de parámetros de prosodia y de VoQ obtiene mejoras importantes respecto del uso de únicamente prosodia.
3. Se hace patente la necesidad de un buen modelado tanto de prosodia como de VoQ, especialmente para conseguir la desambiguación de parejas de estilos expresivos en los que existen dificultades de identificación incluso en habla natural.
4. El factor de modificación en cada caso es un parámetro más con el que trabajar, ya que se debe de llegar a un compromiso entre la calidad y la identificación del estilo de habla expresivo.
5. El contenido semántico del enunciado puede ser un condicionante que limite la identificación del estilo de habla expresivo.

Finalmente, los resultados obtenidos, respecto a la calidad y la identificación del estilo expresivo, se consideran un buen inicio para continuar con el modelado de la prosodia y de la VoQ utilizando el análisis y la síntesis del habla mediante HNM. Es especialmente interesante haber comprobado las posibilidades reales de la metodología de transformación presentada, ya que a pesar de ser una metodología simple ha demostrado que puede llegar a conseguir grandes mejoras en sistemas de SHE.

5.6. Otras aplicaciones de la cualidad de la voz

El conocimiento adquirido sobre el comportamiento de la VoQ en el reconocimiento y síntesis de estilos de habla expresivos, ha permitido abrir líneas de trabajo y de colaboración con el fin de ampliar las aplicaciones de la VoQ. En particular, se ha colaborado con otros investigadores en las siguientes líneas de trabajo:

1. Validación automática de corpus de habla expresiva.
2. Evaluación objetiva de sistemas de CTH.

Como puede observarse, el factor común es el análisis objetivo de la voz, en especial para el caso de SHE, ya sea natural como en la validación de un corpus de habla expresiva o sintética en la evaluación de sistemas de CTH.

5.6.1. Validación automática de corpus de habla expresiva

La contribución a la validación automática de corpus de habla expresiva se llevó a cabo mediante la colaboración en los trabajos de Iriondo et al. (2007a, 2009).

Se participó en la creación de un sistema automático capaz de mejorar la expresividad de un corpus de voz grabado a partir de habla actuada o estimulada. El sistema se entrenó con los resultados de una evaluación subjetiva realizada sobre un reducido conjunto del corpus original. Una vez el sistema fue entrenado se pudo pasar a comprobar el corpus completo, y realizar una poda (*pruning* en inglés) de aquellos enunciados con un estilo expresivo diferente del pretendido por el corpus, manteniendo en este el contenido más cercano a los resultados obtenidos por la clasificación subjetiva. El corpus empleado para realizar los experimentos fue el corpus en castellano presentado en la Sección 4.3.

A lo largo de esta sección se presenta una visión general del trabajo desarrollado, para poder así introducir al lector en el contexto donde la VoQ ha sido aplicada. Los diferentes puntos que se tratan son el de la evaluación subjetiva (Sección 5.6.1.1) y la identificación automática del estilo de habla expresivo del locutor (Sección 5.6.1.2).

5.6.1.1. Evaluación subjetiva

La evaluación subjetiva del corpus partió de la selección de aproximadamente el $10\,\%$ de los enunciados (recordemos que el corpus tiene una duración mayor de 5 horas) para ser usados en el test perceptivo. Se eligieron aleatoriamente 96 enunciados para cada estilo de habla expresivo, dando lugar a un total de 480. Los evaluadores fueron 25 voluntarios, aunque debido a que la evaluación de 480 enunciados por persona sería demasiado trabajoso, el conjunto de test fue dividido en 4 subconjuntos de 120 enunciados cada uno. Se asignaron a cada sujeto un par ordenado de subconjuntos (con la posibilidad de que a dos sujetos les correspondiera el mismo par pero presentados en orden inverso), generando 12 combinaciones diferentes, para impedir que las pruebas realizadas en segunda ronda fueran más sencillas debido al proceso de entrenamiento previo.

El test estaba diseñado de manera que se forzaba dar respuesta a la pregunta *"¿Qué emoción reconoces en el habla del locutor en este enunciado?"*. Las posibles respuestas fueron los 5 estilos de habla expresivos (neutro (NEU), alegre (ALE), sensual (SEN), agresivo (AGR) y triste (TRI)), además de una sexta opción de "No sé/Otra" (Ns/O) para evitar desviar los resultados en caso de confusión o dudas entre opciones. Los resultados obtenidos se presentan en la Tabla 5.24, donde se muestra la matriz de confusión resultante, indicando los porcentajes de clasificación correcta y los errores cometidos por la confusión entre opciones. Como resultado general, el test subjetivo muestra como todos los estilos de habla expresivos de los que consta el corpus obtienen un elevado porcentaje de identificación ($87,1\,\%$ en promedio). El estilo triste presenta el mayor porcentaje de clasificación correcta ($98\,\%$), seguido por sensual ($86,8\,\%$), neutro ($86,4\,\%$), agresivo ($82,7\,\%$) y, finalmente, alegre ($81,0\,\%$). La matriz de confusión revela que los principales errores se producen entre agresivo y

alegre, identificando como alegre ($14,2\,\%$) y agresivo ($15,6\,\%$) respectivamente. Además, neutro se confunde ligeramente con todas las opciones y existe un cierto nivel de confusión de sensual con triste ($5,7\,\%$) y neutro ($4,7\,\%$). La opción "Ns/O", aunque apenas utilizada, estuvo más presente en los estilos neutro y sensual que en el resto de estilos de habla expresivos.

(%)	NEU	ALE	SEN	AGR	TRI	Ns/O
NEU	$\mathbf{86,4}$	$1,3$	$3,6$	$5,3$	$0,7$	$2,7$
ALE	$1,9$	$\mathbf{81,0}$	$0,2$	$15,6$	$0,1$	$1,2$
SEN	$4,7$	$0,1$	$\mathbf{86,8}$	$0,0$	$5,7$	$2,6$
AGR	$1,8$	$14,2$	$0,1$	$\mathbf{82,7}$	$0,1$	$1,1$
TRI	$0,5$	$0,0$	$0,6$	$0,0$	$\mathbf{98,8}$	$0,1$

Tabla 5.24: *Matriz de confusión promedio (%) para el test subjetivo del corpus con 5 estilos de habla expresivos (en* ***negrita*** *el porcentaje máximo de clasificación correcta)*

5.6.1.2. Identificación automática del estilo de habla expresivo del locutor

Como se ha mostrado en la Sección 5.6.1.1, algunos de los enunciados existentes en el corpus tienen limitaciones a la hora de expresar el estilo expresivo deseado. Si estos enunciados se mantienen en el corpus podrían acarrear problemas, tanto en el modelado acústico como durante el proceso de síntesis. Una revisión manual exhaustiva del corpus requeriría mucho tiempo dado el tamaño del mismo, de modo que la solución pasaría por desarrollar un sistema automático capaz de validar todos los enunciados grabados en cuanto a expresividad. Los experimentos, llevados a cabo con tal fin, partieron del análisis acústico de los enunciados del corpus, parametrizaciones de las que se extrajeron estadísticas usadas por algoritmos de Aprendizaje Automático —*Machine Learning*— (ML) para realizar el reconocimiento automático de los diferentes estilos de habla expresivos del corpus.

El análisis acústico implicó la parametrización de la señal de voz del corpus. Para ello se tuvieron en cuenta tanto parámetros de prosodia como de VoQ, cuya detallada descripción se encuentra en el trabajo de Iriondo et al. (2009) y la Sección 5.2.1 respectivamente.

- Prosodia:
  - Parámetros relacionados con la frecuencia fundamental (F_0): escalas lineal y logarítmica.
  - Parámetros relacionados con la energía: escalas lineal y logarítmica.
  - Parámetros relacionados con el ritmo: modelado de la duración basada en la medida $z-$*score* (Ecuación 5.65).

$$z - score \quad = \quad \frac{\text{duración}(ms) - \mu}{\sigma} \tag{5.65}$$

Donde la duración se expresa en milisegundos, μ es su valor medio y σ su desviación estándar.

- Cualidad de la Voz —*Voice Quality*— (VoQ):
  - *Jitter*
  - *Shimmer*
  - *Glottal-to-Noise Excitation Ratio* (GNE)
  - *Hammarberg Index* (HammI)
  - *Drop-off of Spectral Energy above 1000 Hz* (do1000)

Los parámetros de VoQ se eligieron a partir de los resultados obtenidos en Monzo et al. (2007), decantándonos por ellos debido a las posibilidades que presentaban para la caracterización de cada estilo de habla expresivo.

Las pruebas preliminares estaban centradas en la aplicación de las técnicas de ML únicamente sobre los parámetros de prosodia de cada enunciado (Iriondo et al., 2007b). Sobre los parámetros se calcularon la primera y segunda derivada discreta y se extrajeron las siguientes estadísticas: media, varianza, valor máximo, valor mínimo, rango, asimetría (*skewness*), curtosis (*kurtosis*), cuartiles y rango intercuartílico. A partir de los experimentos efectuados se obtuvo la matriz de confusión presentada en la Tabla 5.25, donde se observa una cierta confusión entre los estilos alegre-agresivo y, en menor grado, neutro-alegre. Cualitativamente, las mismas parejas fueron confundidas en el test subjetivo (Tabla 5.24), aunque sin embargo, a diferencia del sistema automático donde la principal confusión se daba en la pareja sensual-neutro, los evaluadores eran propensos a confundir el estilo alegre con el agresivo.

(%)	**NEU**	**ALE**	**SEN**	**AGR**	**TRI**
NEU	$\mathbf{93,9}$	$0,9$	$4,5$	$0,2$	$0,4$
ALE	$1,2$	$\mathbf{97,1}$	$0,2$	$1,6$	$0,0$
SEN	$4,9$	$0,1$	$\mathbf{94,8}$	$0,0$	$0,2$
AGR	$0,0$	$0,8$	$0,0$	$\mathbf{99,1}$	$0,1$
TRI	$0,4$	$0,1$	$0,1$	$0,2$	$\mathbf{99,3}$

Tabla 5.25: *Matriz de confusión promedio (%) para la identificación automática preliminar de estilos de habla expresivos (en* ***negrita*** *el porcentaje máximo de clasificación correcta)*

Considerando únicamente el reconocimiento automático de emociones, los resultados obtenidos podrían ser considerados como excelentes. Sin embargo, recordemos que el objetivo que se pretendía conseguir era el de validar la autenticidad del corpus grabado, y el test subjetivo reveló como un pequeño porcentaje de enunciados erróneos, desde el punto de vista de la expresividad, no fueron descartados por el sistema automático.

Se pretendía ajustar el clasificador automático para que emulase el criterio subjetivo, con lo que la selección de atributos fue un aspecto importante. La necesidad de definir una medida de coincidencia entre el resultado automático y subjetivo se

hizo patente, con lo que se utilizó la medida $F1$ para ello (Iriondo et al., 2009). Con el objetivo de mejorar los resultados, se siguieron las siguientes pautas:

- Inclusión de parámetros de VoQ: un estudio basado únicamente en parámetros de prosodia muestra la dificultad en la discriminación entre algunos estilos de habla expresiva como triste o sensual (Iriondo et al., 2009). El estudio de estos enunciados indicó que los parámetros de VoQ podrían ser útiles para esta tarea.

- Nueva estrategia de selección de atributos: las estrategias de selección de atributos usadas *Forward Selection* (FW)[7] y *Backward Elimination* (BW)[8] no pueden deshacer decisiones ya tomadas. Una decisión combinada puede evitar este problema y evitar caer en un mal máximo local.

- Combinación de clasificadores: dado que los clasificadores presentan mejores y peores resultados en cuanto a precisión y cobertura (directamente relacionado con la medida $F1$), una combinación de estos podría mejorar el resultado final.

Algoritmo	**FW sin VoQ**	**FW + VoQ**	**3FW-1BW + VoQ**
SMO	$0,49$	$0,59$	$0,61$
J48	$0,43$	$0,52$	$0,56$
NB	$0,42$	$0,48$	$0,58$

Tabla 5.26: *Análisis de las mejoras introducidas, en términos de medida $F1$, por usar parámetros de cualidad de la voz y diferentes estrategias de selección de atributos*

Por tanto, según lo que nos interesa en esta sección, que es la aportación que la VoQ ha producido en la aplicación de validar automáticamente un corpus de habla expresiva, vamos a ver la comparativa de resultados conseguidos por su uso sobre los diferentes experimentos y configuraciones probadas. Los clasificadores empleados fueron los tres que demostraron mejores resultados: SMO (Apéndice F.9), J48 (Apéndice F.10) y Naïve Bayes (NB) (Apéndice F.11), mientras que la selección de atributos era FW, BW y decisión combinada. Los resultados de las mejoras propuestas se muestran en la Tabla 5.26, para los tres algoritmos de clasificación y dos estrategias de selección (FW, 3FW-1BW). Se puede comprobar como las mejoras introducidas, por utilizar parámetros de VoQ y una selección de atributos bidireccional, implican un incremento relativo superior al $20\,\%$ en términos de medida $F1$.

[7] *Forward Selection* (FW) es una técnica de selección de atributos que empieza por un conjunto de datos vacío, el atributo que mejora el rendimiento del clasificador se elige para ser añadido en la siguiente iteración.

[8] *Backward Elimination* (BW) es una técnica de selección de atributos que empieza con el conjunto completo de datos. El atributo que mejora la medida de comparación cuando no ha sido considerado se elimina en la próxima iteración.

5.6.2. Evaluación objetiva de la síntesis del habla expresiva

5.6.2.1. Descripción

La evaluación objetiva de sistemas de CTH se planteó tanto como una necesidad de sustituir a las pruebas subjetivas como una manera más de evaluar la calidad del habla generada. En primer lugar, la necesidad vendría dada por el hecho de que las evaluaciones subjetivas suponen un proceso lento y costoso de evaluación, dado que se necesita de una etapa de diseño de las pruebas, voluntarios que las realicen, tiempo de realización y el procesamiento de toda la información recuperada. Además, disponer de una metodología que permitiera un análisis objetivo de la calidad facilitaría la evaluación de los sistemas de CTH durante el proceso de desarrollo. En segundo lugar, complementa a las pruebas subjetivas, ya que permite detectar la efectividad del enfoque realizado por el sistema de CTH. En el caso concreto en el que nos hallamos, la evaluación objetiva adquiere mayor notoriedad por estar trabajando con diferentes estilos de habla expresivos. Las diferencias existentes entre ellos, a nivel de prosodia y de VoQ, hacen que en la decisión del mejor enfoque, para abordar el problema de generar habla con el estilo expresivo deseado, sea necesario detectar los puntos fuertes y las limitaciones del sistema de CTH.

La contribución realizada en el área de la evaluación objetiva de sistemas de CTH se llevó a cabo a partir de la colaboración con otros investigadores. En un primer momento nació una línea de colaboración dentro de los sistemas de CTH basados en HMM (Sección 3.3.2.4) dando pie a diferentes publicaciones, donde se propuso el sistema de conversión para castellano que sería la base para futuros experimentos (Gonzalvo et al., 2007c). También se colaboró en trabajos en los que se analizó la síntesis basada en HMM y el modelado prosódico (Gonzalvo et al., 2007a,b), trabajos en los que se probaron técnicas desarrolladas por otros miembros del grupo de investigación, como es el CBR para el modelado prosódico (Iriondo et al., 2007c). Una conclusión interesante, directamente ligada con los estudios de VoQ en los que se centra esta tesis, fue la de conocer las posibilidades que podía ofrecer la VoQ en la evaluación objetiva de la síntesis del habla, especialmente para valorar las técnicas de transformación y conversión de voz implicadas (Gonzalvo et al., 2007c). Finalmente, se llegó a los experimentos donde se usó la evaluación objetiva, para demostrar la efectividad del sistema de CTH propuesto, con diferentes estilos de habla expresivos (Gonzalvo et al., 2009, 2010).

En el trabajo de Gonzalvo et al. (2009, 2010) se describe un sistema de CTH de alta calidad para estilos de habla expresivos. La calidad del habla generada se vio mejorada por la extracción de espectro STRAIGHT y excitación mixta (Zen et al., 2007), y se evaluaron dos técnicas para modelar los estilos de habla expresivos. Primero, un método que modelaba simultáneamente todos los estilos expresivos con un único modelo acústico. En segundo lugar, se aplicó una técnica de adaptación que convertía un estilo expresivo neutro en otro. La información utilizada durante los experimentos fue el corpus expresivo de frases en castellano presentado en la Sección 4.3, del cual se hizo uso de los siguientes estilos de habla expresivos: neutro (NEU), alegre (ALE) y triste (TRI). Se llevó a cabo una evaluación subjetiva que mostró

la calidad del sistema y la intensidad del estilo expresivo, mientras que se introdujo la evaluación objetiva basada en parámetros de VoQ para medir la efectividad de las propuestas.

5.6.2.2. Evaluación subjetiva

Tal y como se presenta en la Sección 5.6.2.1, la evaluación del sistema se dividió en dos partes: subjetiva y objetiva. Antes de pasar a describir la aportación en la evaluación objetiva, se considera necesario presentar los resultados de la subjetiva, para poder así relacionar ambos más adelante.

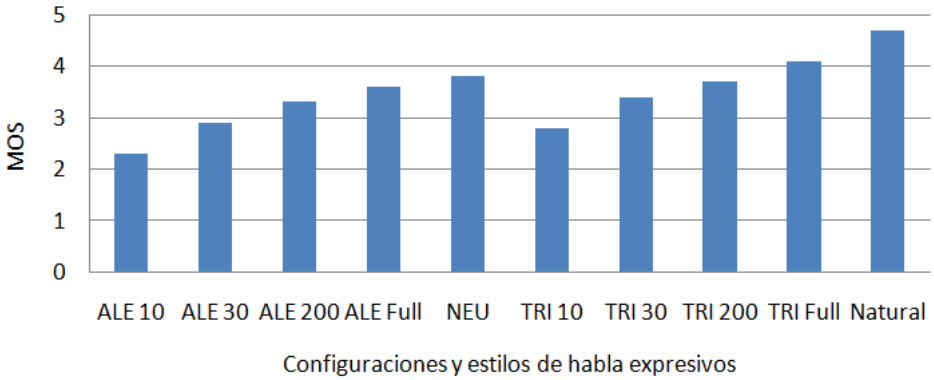


Figura 5.15: *MOS para la evaluación de la naturalidad por configuración y estilo de habla expresivo. Por ejemplo, ALE200 indica que se trata del estilo expresivo destino alegre, utilizando* 200 *para la adaptación de su modelo, mientras que NEU indica que se trata de síntesis en estilo neutro y Natural que se trata de ejemplos del corpus.)*

La evaluación subjetiva, usada para evaluar el comportamiento de las configuraciones del sistema, se llevó a cabo sobre 20 enunciados de test. Los estilos alegre y triste fueron usados como estilos destino, adaptados a partir del estilo neutro con 3 configuraciones: 10, 30 y 200 enunciados. El identificador "Full" se reservó para el estilo mixto, donde se modeló un corpus construido por completo a partir de los estilos neutro, alegre y triste (véase la Figura 5.15 que muestra la evaluación de la calidad de los 3 estilos y la Figura 5.16 que muestra la de la intensidad de alegre y triste). En la Figura 5.15 se presenta el resultado de un test MOS (ITU-P.800, 1996) para evaluar la naturalidad del habla sintetizada. Para el estilo neutro se consiguió un MOS de $3,8$, obteniéndose valores por encima y por debajo en los otros dos estilos de habla expresivos. El estilo alegre se ve afectado por altos valores de F_0 que distorsionan la calidad de la señal, mientras que el estilo triste presenta una mejor calidad, próxima al natural, con un MOS cercano al 4 para el modelo de estilo mixto. Es un hecho conocido que los parámetros generados por un sistema de CTH basado en HMM sufren de un efecto de suavizado (Toda y Tokuda, 2005), efecto que ayuda al estilo triste a obtener un buen resultado debido a su baja variación de F_0. Tanto la Figura 5.15 como la Figura 5.16 dan una medida de cómo de bien reproduce el sistema de CTH cada

estilo de habla expresivo. A partir de una puntuación superior a 3 se consideró que el estilo de habla expresivo se reproducía adecuadamente.

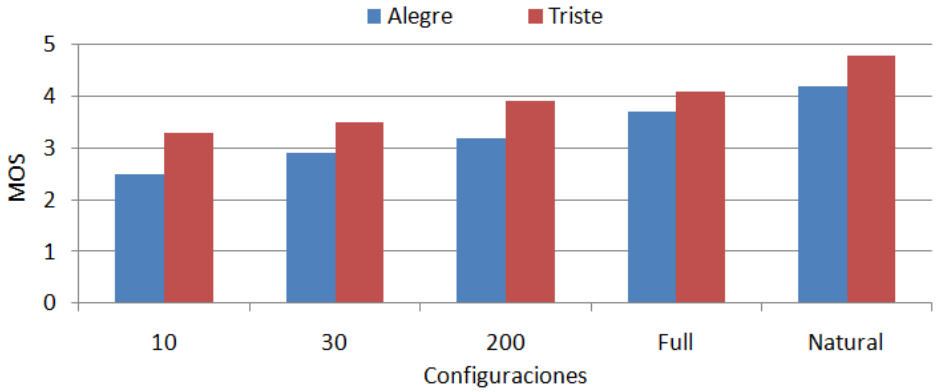


Figura 5.16: *MOS para la evaluación de la intensidad de los estilos de habla expresivos por configuración*

5.6.2.3. Evaluación objetiva

Una vez realizada la evaluación subjetiva (Sección 5.6.2.2), se pasa a mostrar la evaluación objetiva. El experimento llevado a cabo tenía como objetivo confirmar el efecto de la producción del estilo de habla expresivo. En algunos trabajos, la distorsión del *mel-cepstrum* se calcula para determinar la distancia acústica entre los estilos neutro y destino (Kawanami et al., 2003), mientras que otros presentan la raíz cuadrada del error cuadrático medio de la F_0 (Yamagishi et al., 2006). Debido a que la adaptación de los estilos de habla expresivos afecta tanto al tracto vocal como a la prosodia, sería deseable disponer de una evaluación que midiera implícitamente su comportamiento. Los parámetros de VoQ cumplen con estas condiciones, y ya fueron presentados como buenos elementos para la discriminación de estilos de habla expresivos (Monzo et al., 2007).

Los parámetros con los que se ha trabajado y su capacidad discriminatoria se presenta en las Secciones 5.2.1 y 5.4 respectivamente. Los parámetros fueron calculados sobre las vocales de 200 enunciados de test sintetizados, ya que representan zonas sonoras estables en las que calcular los parámetros de VoQ (Sección 5.4.1). A continuación se listan los parámetros utilizados:

- *Jitter*
- *Shimmer*
- *Glottal-to-Noise Excitation Ratio* (GNE)
- *Spectral Flatness Measure* (SFM)
- *Drop-off of Spectral Energy above 1000 Hz* (do1000)

- *Hammarberg Index* (HammI)
- *Relative Amount of Energy above 1000 Hz* (pe1000)

El *jitter* y *shimmer* fueron calculados siguiendo la metodología presentada en la Sección 5.3. En cuanto al parámetro GNE, fue seleccionado en lugar de HNR porque ambos describen un ruido aditivo, pero el primero es prácticamente independiente del *jitter* y *shimmer* que pueda aparecer en el habla. El parámetro do1000, aunque inicialmente se tuvo en cuenta para los experimentos, fue finalmente descartado debido a la poca dependencia que demostró tener con los estilos de habla expresivos bajo estudio (Sección 5.4).

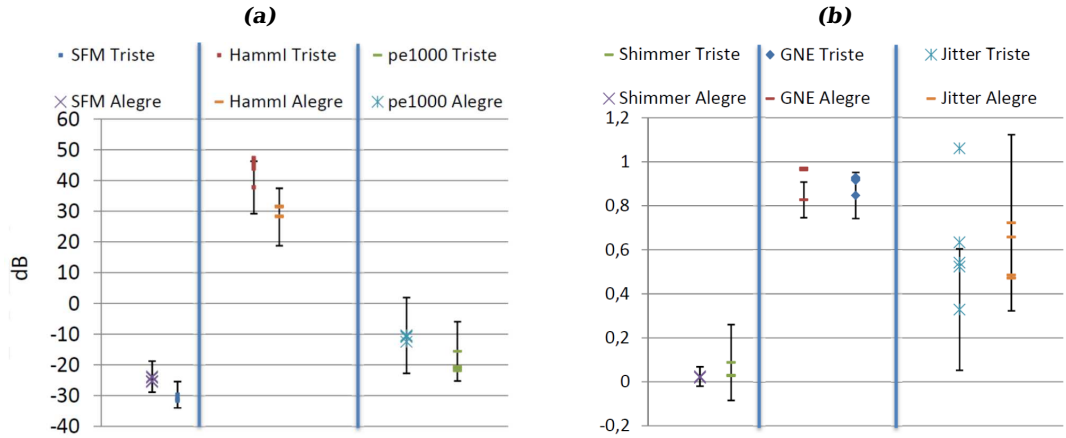


Figura 5.17: *Parámetros de cualidad de la voz en la evaluación objetiva de la conversión de texto en habla*

No es el objetivo de esta sección la discriminación de los estilos de habla expresivos, sino mostrar las limitaciones que existen cuando estos son sintetizados. De esta manera, la Figura 5.17 muestra las estadísticas descriptivas de los parámetros de VoQ para los estilos alegre y triste. El área dentro de las líneas verticales, para cada parámetro y estilo, representa la desviación estándar del habla natural, pudiéndose ver como la mayoría de los parámetros de VoQ se encuentran dentro de los límites marcados por esta. Sin embargo, puede apreciarse que el parámetro HammI, para el estilo triste, tiende hacia valores altos, lo cual significa que el caso sintético es de tono más grave que el natural. También es importante subrayar que el papel del parámetro GNE para el estilo alegre, en el que se pueden observar valores elevados que se manifiestan como un ruido aditivo, coincidiendo con la medida subjetiva en la que este estilo se ve afectado por una distorsión, haciendo que su MOS para la naturalidad sea inferior al resto de estilos. Finalmente, los resultados del *jitter* para el estilo triste indican que el habla sintetizada sufre de un efecto de habla temblorosa y, en concreto, cuantos más enunciados fueron usados en la adaptación mayor fue el *jitter*.

CAPÍTULO 6

Conclusiones y líneas de futuro

Este capítulo se organiza en cuatro secciones, revisando el trabajado desarrollado en esta tesis, presentando las conclusiones alcanzadas y proponiendo las líneas de futuro a seguir:

- Para empezar se presentan unas conclusiones y líneas de futuro generales (Sección 6.1), dando una visión de conjunto de los aspectos más relevantes que se han tratado y de los resultados obtenidos.
- Con la visión global del punto anterior se entra en detalle para cada uno de los tres grandes bloques en los que el modelado de la Cualidad de la Voz —*Voice Quality*— (VoQ) ha sido el foco de interés:
  - La parametrización (Sección 6.2).
  - La discriminación de estilos de habla expresivos (Sección 6.3).
  - La Síntesis del Habla Expresiva (SHE) (Sección 6.4).

6.1. Conclusiones y líneas de futuro generales

Esta tesis nació con el objetivo de mejorar la naturalidad de la interacción hombre-máquina, con lo que se ha trabajado en aumentar la capacidad de transmitir y percibir estilos de habla expresivos de forma automática. Para ello, se ha realizado una exhaustiva búsqueda de cuál era la problemática a resolver, cómo se estaba afrontando hasta el momento el problema y dónde se podría actuar de forma que se obtuvieran mejoras.

Es a partir de estas cuestiones que se llega a la conclusión de la necesidad de proponer nuevas parametrizaciones que sean capaces de caracterizar el habla, pudiendo de este modo trabajar directamente sobre aquellos factores que determinan

su naturalidad: la VoQ. En base a esta premisa, se llevó a cabo una importante tarea de documentación, averiguando más sobre esta parametrización, conociendo sus usos, viendo sus limitaciones y descubriendo sus posibilidades. La visión adoptada ha sido la de intentar abrir lo máximo posible el foco de atención sobre aplicaciones donde históricamente ha sido utilizada, buscando similitudes con la situación en la que nos encontramos y analizando cuál de los enfoques existentes encaja mejor con nuestro caso. Todo este proceso de documentación ha formado una parte importante dentro del trabajo desarrollado en la presente tesis y, es por eso, que se ha querido plasmar toda esta información de forma que pueda ser utilizada como guía práctica tanto para quien se inicie en la temática de la parametrización del habla como para quien desee disponer de una guía de consulta.

Del proceso de documentación, finalmente se seleccionaron aquellas configuraciones de parámetros que mejor encajan con la problemática que nos vamos a encontrar. Debido a que la interacción hombre-máquina busca cada vez más maximizar la naturalidad, facilitando al máximo el acceso a la información, se decidió centrar la atención en aquellos parámetros medidos directamente sobre la señal acústica, siendo calculados sin tener que acceder de forma invasiva al interlocutor (p. ej. sólo se necesitará un micrófono). Entre las opciones de medida de los parámetros, se vio la utilidad de realizarla a partir de la señal acústica, haciéndola independiente del uso de modelos de señal glótica que podrían añadir errores por el propio modelo, siendo idónea en la transformación de estilos de habla expresivos al no necesitar de un proceso que invierta este modelo.

Un factor importante fue el de disponer de material de habla para realizar los experimentos, siendo este el corpus expresivo en castellano del grupo de investigación GTM, además de todo el trabajo previo en modelado de prosodia. A pesar de esto, no quedó ahí la colaboración entre miembros del grupo, ya que se cooperó con líneas de investigación paralelas, como son el reconocimiento de emociones, o estilos de habla expresivos en general, y los sistemas de CTH basados en HMM y HNM. Por esta razón, una característica importante de esta tesis, considerada fundamental tratándose de un trabajo de investigación, es la colaboración establecida con otros investigadores, participando en multitud de aplicaciones en las que era necesaria una mejora en las prestaciones obtenidas hasta el momento. De este modo, este trabajo nace a partir de las limitaciones que los basados en la parametrización de prosodia habían detectado, tanto en aplicaciones de reconocimiento de emociones como de SHE. Es así como, en base a estas limitaciones, se iniciaron las primeras experimentaciones con los parámetros de VoQ propuestos, analizando la capacidad discriminatoria que poseían y de ahí siendo aplicados en reconocimiento y síntesis de estilos de habla expresivos.

Además de la línea de trabajo principal orientada a la SHE y, muy ligada a ella, el reconocimiento de emociones, se ha participado en otras aplicaciones donde el uso de VoQ ha demostrado su utilidad. Estas son la validación automática de corpus de habla expresiva y la evaluación objetiva de la calidad del habla en sistemas de CTH. Se ha mostrado como utilizando la parametrización y las técnicas adecuadas, una tarea tan costosa como es la creación de un corpus y la evaluación subjetiva del habla

puede ser llevada a cabo mediante análisis objetivos, pudiendo mejorar la calidad del corpus por haber sido validado y la del habla sintetizada por poder probar el sistema de síntesis durante su diseño.

Una vez presentada la motivación y el escenario donde se enmarca la tesis, se pasan a resumir los aspectos más relevantes de las aportaciones llevadas a cabo, en las que la VoQ ha sido el centro de atención. Son tres los objetivos que se deseaba conseguir, de los que se dará una visión global a continuación y detallada a lo largo del capítulo, cubriendo la parametrización (parámetros que serán seleccionados) (Sección 6.2), la capacidad discriminatoria (relaciones entre parámetros y estilos de habla expresivos) (Sección 6.3) y la SHE (metodología de transformación de estilos de habla expresivos) (Sección 6.4):

Parametrización de la VoQ. Partiendo de la selección de los parámetros de VoQ, se presenta el proceso de modelado de los mismos. El proceso de análisis se plantea bajo dos enfoques distintos, dependiendo de si se utiliza la herramienta Praat o Matlab. La ventaja de Praat es que permite un rápido desarrollo de los primeros experimentos y es una herramienta ampliamente conocida, con lo que facilita la presentación de los resultados a la comunidad científica. En cuanto a Matlab, presenta mejores características para el modelado, ya que pueden ser desarrolladas nuevas metodologías de análisis y de modificación, y permite mayor facilidad de integración con trabajos de otros miembros del grupo de investigación, como es el caso del análisis mediante la previa parametrización del habla basada en HNM. Esta última opción liga directamente con el proceso de síntesis, donde el procedimiento a seguir utiliza la parametrización del habla basada en HNM, con lo cual el proceso de análisis y de síntesis siguen criterios idénticos, posibilitando la modificación de parámetros de VoQ durante la transformación de estilos de habla expresivos. Complementariamente a los parámetros ya conocidos, se plantean nuevas metodologías de análisis y de modificación para el *jitter* y *shimmer*, adecuando su modelado a las necesidades aparecidas por trabajar con habla expresiva.

Capacidad discriminatoria de estilos de habla expresivos usando VoQ. Con la propuesta de parámetros para ser utilizados en aplicaciones relacionadas con habla expresiva, se estudia la capacidad que estos tienen a la hora de discriminar los diferentes estilos expresivos. Demostrada la posibilidad de ser capaces de caracterizar el habla con parámetros de VoQ, se buscan aquellas configuraciones que permitan la máxima identificación. Asimismo, utilizando diferente material de voz (corpus expresivos en castellano de frases, de palabras alineadas y en alemán) se extrae la dependencia entre estilos y qué parámetros son los que tienen mayor presencia en su modelado, pudiendo aplicar este conocimiento posteriormente en el reconocimiento de emociones y la SHE.

Transformación de estilos de habla expresivos en SHE. Conocidos los parámetros de VoQ y cómo estos pueden modelar a los estilos de habla expresivos, se afronta el siguiente reto, la transformación de estilos de habla expresivos por medio de la modificación de los parámetros de VoQ junto a los de prosodia. Para ello, el habla original sufre una parametrización inicial basada en HNM, cuyos parámetros resultantes son los que permitirán el modelado de la VoQ. El experimento de transformación se inicia desde enunciados en estilo neutro hacia el resto de estilos de los que consta el corpus de frases en castellano (alegre, sensual, agresivo y triste). La elección final de qué parámetros, junto a los de prosodia, y qué valores de transformación se aplican, dependerá de la calidad final del habla generada y de la intensidad con la que ese estilo se percibe, con lo que el trabajo desarrollado en discriminación se convierte en una herramienta de consulta que permite guiar a la transformación.

Viendo en conjunto el trabajo realizado, se plantean una serie de conclusiones y de líneas de trabajo futuro asociadas. Se ha demostrado la utilidad que los parámetros de VoQ presentan en la caracterización de estilos de habla expresivos y, juntamente en colaboración con los de prosodia, pueden ser empleados en reconocimiento y síntesis de estilos de habla expresivos. Asimismo, se ha mostrado como pueden ser empleados en otras aplicaciones como la validación automática de corpus de habla expresiva y la evaluación objetiva del habla generada por un sistema de CTH. A partir de los primeros experimentos y los resultados prometedores conseguidos, se plantean las limitaciones detectadas que deben de ser abordadas en el futuro:

- Aparece la necesidad de un modelado más específico para los diferentes estilos y enunciados, tal y como sucede para la prosodia. No es suficiente analizar la señal de voz del corpus en conjunto, ya que características como la posición en la frase, las características del fonema, la modalidad oracional, entre otras, pueden ser condicionantes de los valores adoptados por los parámetros.

- Mejoras en el proceso de modificación de los parámetros que permitan la modificación de todos ellos, minimizando la degradación de la señal del habla resultante y maximizando la percepción del estilo transmitido.

- La búsqueda de la mejora de la calidad del habla obtenida, sin ir en detrimento de la percepción del estilo de habla expresivo transmitido, implica la mejora continua de todas las herramientas implicadas: el modelado de la VoQ, el modelado de la prosodia y la parametrización del habla basada en HNM.

- Vista la utilidad de la VoQ y todas las variantes que tiene en su definición (extraída de la señal glótica, modelo de la señal glótica a partir de la señal acústica, o bien directamente de la señal acústica), puede plantearse su utilización en futuras aplicaciones de análisis y de síntesis.

- El modelado de la VoQ ha demostrado su utilidad y deberá plantearse la integración en software que pueda hacer uso de sus beneficios (p. ej. un sistema de CTH).

Las líneas de futuro aquí presentadas quedan ampliadas y matizadas en las Secciones 6.2, 6.3 y 6.4, donde se entra en detalle en el trabajo llevado a cabo para cada uno de los objetivos marcados en esta tesis.

Para terminar, el presente trabajo de tesis ha propiciado diversas contribuciones a congresos internacionales y nacionales, y se ha visto influenciado por la participación de su autor en diferentes proyectos de investigación y desarrollo de ámbito europeo y nacional. En el Apéndice A se presentan las principales aportaciones realizadas.

6.2. Parametrización de la cualidad de la voz

El uso de la parametrización de la VoQ aparece como respuesta a la necesidad de hallar parámetros capaces de caracterizar el habla, de tal forma que estilos de habla expresivos puedan ser reconocidos y generados por medio de un sistema automático. Hasta el momento se había trabajado generalmente con parámetros de prosodia, que aunque dando buenos resultados para algunos estilos expresivos presentaban grandes limitaciones para otros.

La VoQ ha sido utilizada por la comunidad científica en multitud de ámbitos, ya sea estudios sociales, estudios clínicos de patologías de la voz o en las mismas tecnologías del habla. Una problemática detectada en este último punto, las tecnologías del habla, es que no ha existido un consenso entre qué parámetros utilizar, ya que la VoQ tiene multitud de variantes en los parámetros que la definen: a partir de la extracción de la señal glótica, del modelado de la señal glótica a partir de la acústica y directamente desde la señal acústica. En el escenario en el que se enmarca esta tesis, la mejora de la naturalidad de la interacción entre hombre-máquina, se considera que la forma más universal de acceso a la información será la señal de habla, con lo que los parámetros se analizan directamente a partir de su captura con un micrófono, de forma no invasiva, sin la necesidad de hardware extra ni de modelos de señal glótica que puedan introducir errores.

Dentro del conjunto de parámetros de interés se ha seleccionado aquellos que representaran diferentes aspectos del habla sin introducir redundancias ni interferencias sobre otros: *jitter*, *shimmer*, HNR, GNE, SFM, do1000, HammI, pe1000. A pesar de esto, cada uno de ellos ha sido elegido en cada aplicación dependiendo del objetivo de su uso, ya que a esta puede serle de mayor utilidad uno u otro. Como ejemplo se tiene el caso del NNE, HNR y GNE, donde finalmente se descartaba NNE por tener similitudes con los otros dos y se proponía HNR o GNE en función de su aplicación final (p. ej. HNR junto a la parametrización del habla basada en HNM).

Establecido el conjunto de parámetros de interés, su modelado se define para las herramientas Praat y Matlab, disponiendo así de un *toolbox* de investigación y desarrollo de prototipos. Praat permitió utilizar una herramienta que implementaba el análisis de algunos de los parámetros así como que permitía el cálculo de otros, pudiendo realizar las primeras pruebas de VoQ y presentarlas a la comunidad científica. Con la migración de Praat a Matlab se pudo avanzar en el desarrollo de nuevas definiciones para los parámetros *jitter* y *shimmer*, adecuándolos a las necesidades

propias del habla expresiva. El caso de Matlab también tiene a su favor el hecho de que facilita la realización de prototipos que se comunicaran con los de otros investigadores, pudiendo de este modo integrar distintas líneas de trabajo y extraer resultados conjuntos. Además, para la metodología de análisis se desarrolló una variación para ser empleada con habla espontánea, demostrando tanto la utilidad de los parámetros como la de la metodología en situaciones reales.

De los procesos de análisis y de modificación de los parámetros, se concluye que existen ciertas limitaciones a la hora de crear un modelo de cada parámetro ligado a un estilo de habla expresivo concreto. Esto se aprecia especialmente en el hecho que, en ocasiones, las distribuciones de parámetros tienen una elevada dispersión, dificultando la tarea de discriminación o de modificación en aplicaciones de reconocimiento y síntesis de habla expresiva. Debido al contexto del habla expresiva, aparecen problemáticas no planteadas en otras áreas (p. ej. la variabilidad de los parámetros debidos a la entonación de la frase), en las que la información de voz que se puede utilizar está acotada (p. ej. examen clínico). Es por tanto, que a partir de la parametrización y del modelado de la VoQ obtenido, se plantean una serie de propuestas de líneas de trabajo futuras:

- Una vez se ha trabajado con los parámetros seleccionados, se plantea el cálculo de parámetros propios de la señal glótica a partir de la señal acústica. Las primeras aproximaciones al análisis de estos parámetros ya ha sido planteada bajo las mismas metodologías, empleando Praat y Matlab, observando la viabilidad de ser incluidos en el conjunto de parámetros de interés.

- Del mismo modo que se han planteado mejoras en el modelado de los parámetros *jitter* y *shimmer*, en el contexto del habla expresiva, se propone estudiar la necesidad de adaptación del resto de parámetros. Para las nuevas propuestas metodológicas de análisis y modificación, así como para las descritas en la presente tesis (*jitter* y *shimmer*), se considera de utilidad la realización de estudios comparativos con las tradicionales para cuantificar las mejoras obtenidas.

- A lo largo de los experimentos se ha observado que sería un factor interesante a tener en cuenta la dependencia existente entre el parámetro y otras informaciones relativas al fonema bajo análisis. Por tanto, se propone modelar la VoQ de una forma similar a como se lleva a cabo el modelado de la prosodia, siendo esta una línea de trabajo que podría dar lugar a mejoras importantes en aplicaciones basadas en el modelado de estos parámetros.

6.3. Discriminación de estilos de habla expresivos

A pesar de la documentación existente sobre el uso de la VoQ en el modelado del habla expresiva, no existen trabajos donde se vean implicados todos los parámetros que se han seleccionado, que sea una cantidad tan elevada de parámetros y, fundamentalmente, con un objetivo tan claro como es el de la mejora de la naturalidad en

sistemas de interacción hombre-máquina complementando a la prosodia. La implicación que esto tiene, tanto en su aplicación en reconocimiento de emociones como en sistemas de SHE, es que ha sido necesario un estudio previo acerca de cómo estos parámetros son capaces de discriminar entre estilos expresivos. De este modo, se ha podido conocer en qué parámetro o configuraciones de parámetros se debe de prestar atención dependiendo de la aplicación a la que van dirigidos (p. ej. SHE con un número de estilos expresivos destino acotado y conocido, donde se pueden emplear aquellos parámetros que mejor caractericen a los estilos en cuestión).

La capacidad discriminatoria, que los parámetros de VoQ presentan ante estilos de habla expresivos, se llevó a cabo bajo dos enfoques: utilizando corpus expresivos grabados bajo unas determinadas condiciones de control y, a partir de estos resultados, utilizando habla espontánea que mostrara realmente la utilidad de los parámetros en un entorno no controlado. El procedimiento a seguir fue el de analizar todos los parámetros, extraer estadística descriptiva sobre los valores para observar las tendencias de cada uno de ellos y aplicar una clasificación automática para extraer relaciones entre parámetros y estilos de habla expresivos. Este procedimiento se realizó para diferentes configuraciones de parámetros y de estilos expresivos, obteniendo los siguientes resultados:

1. La capacidad de discriminación de cada uno de los parámetros, que de forma individual se vieron implicados en la discriminación de todos los estilos del corpus.

2. Configuración de parámetros que dan los máximos valores de discriminación, tanto para el caso de verse implicados todos los estilos a la vez como en el caso de analizar la relación entre parejas de estilos, siendo el primer paso para la obtención de modelos de VoQ relacionados con los estilos de habla expresivos.

3. Debido a que los resultados de discriminación podían ser similares variando el número de parámetros utilizados, se estudió la relevancia de cada uno de ellos en la obtención de los mejores resultados de discriminación, es decir, cuántas veces aparecía un parámetro en las configuraciones que estuvieran a una cierta distancia del valor máximo de referencia. De este modo, se dispone de tablas que muestran aquellos parámetros más relevantes en la discriminación de parejas de estilos expresivos, de utilidad en el modelado de la VoQ para la transformación de estilos de habla expresivos durante la SHE.

4. A partir de los resultados obtenidos sobre corpus orales grabados bajo determinadas condiciones de control, se demostró la utilidad de los parámetros en un contexto de habla espontánea, donde las condiciones de grabación y la disponibilidad de información extra (p. ej. transcripción fonética) estaba fuera de nuestro alcance.

Concluidos todos los experimentos acerca de la capacidad discriminatoria de la VoQ, es cuando se observa la utilidad y las posibilidades de la VoQ, pudiendo caracterizar estilos de habla expresivos mediante la combinación de los diferentes parámetros

y disponiendo de la información necesaria como para poder aplicarlos en reconocimiento y síntesis del habla expresiva. Aún así, también aparecen las limitaciones de los parámetros de VoQ propuestos ante los estilos de habla expresivos, observando la dificultad en la construcción de un modelo de aplicación "universal", ya que son varias las configuraciones de parámetros que pueden obtener buenos resultados. No obstante, gracias a esta variedad de buenas configuraciones de parámetros y, dependiendo de la aplicación final en la que se esté trabajando, se podrá utilizar el modelo que mejor se adapte a las necesidades (p. ej. qué parámetros a utilizar en SHE para mantener una cierta calidad con un determinado nivel de percepción del estilo generado).

A partir de los resultados obtenidos y de las conclusiones presentadas, se plantea un siguiente paso como líneas de futuro, ya sea para obtener mejores resultados como para iniciar o continuar con líneas de investigación paralelas:

- El conocimiento de la dependencia entre los parámetros de VoQ y la posición dentro del enunciado analizado podría permitir un mejor modelado del estilo bajo análisis, permitiendo mayores niveles de discriminación.
- Tal y como se plantea en la Sección 6.2, la ampliación de nuevos parámetros, junto con la propuesta de un mejor modelado, podría mejorar los resultados especialmente entre aquellos estilos que han demostrado mayores dificultades en su discriminación (p. ej. alegre y agresivo o triste y sensual).
- Se inicia el camino para que otros estudios, asociados al reconocimiento de emociones, puedan incrementar sus niveles de clasificación gracias a estos parámetros, tal y como ya se ha desprendido de las primeras pruebas que se han realizado al respecto.
- Con los resultados obtenidos para diferentes corpus e incluso para habla espontánea, se plantea trabajar con más estilos de habla expresivos, ampliando así la capacidad de extraer más modelos. El uso de corpus con varios locutores y lenguas, tal y como ya se ha introducido en esta tesis, es beneficioso para poder comparar y ratificar dichos resultados.

6.4. Síntesis del habla expresiva

La Síntesis del Habla Expresiva (SHE), el último gran bloque tratado en la presente tesis, muestra el trabajo realizado en la transformación de estilos de habla expresivos mediante la modificación de los parámetros de VoQ junto con los de prosodia. Para llevar a cabo la transformación de estilos de habla expresivos se tienen en cuenta tanto los experimentos propios realizados en VoQ, como el trabajo del resto de miembros del GTM en el modelado de la prosodia y la parametrización del habla basada en HNM.

A partir de los experimentos de discriminación usando los parámetros de VoQ, se diseña la metodología de transformación de estilos de habla expresivos. Esta metodología parte de un sistema de análisis y de síntesis del habla basado en HNM, cuyos

parámetros serán los utilizados en el modelado de la VoQ. El proceso de transformación parte de un enunciado originalmente expresado en estilo neutro, generando los estilos de habla expresivos contenidos en el corpus oral expresivo de referencia de frases en castellano: alegre, sensual, agresivo y triste. Para cada par de transformaciones se propone una configuración de parámetros y factores de modificación, obtenidas respectivamente del estudio del corpus mediante la capacidad discriminatoria de cada uno de ellos y del cálculo de estadística descriptiva sobre los valores.

Una vez propuesta la metodología y disponibles los resultados de la transformación en SHE es necesaria su evaluación. Su eficacia y calidad se evalúa mediante una serie de pruebas perceptivas que se centran en comprobar la preferencia de usar VoQ junto con prosodia, en evaluar la calidad global del habla generada y en la identificación de los estilos destino resultantes de la transformación por parte del evaluador. Del análisis de los resultados se extraen las siguientes conclusiones:

1. Preferencia entre el uso de únicamente prosodia o de una combinación de esta y de VoQ. Los resultados mostraron como el uso de VoQ era claramente preferido para todas las transformaciones.

2. Además de la preferencia de usar VoQ junto a la prosodia, se consideró de importancia la evaluación de la calidad global del habla generada. Este test se realizó para diferentes configuraciones metodológicas que incluían habla natural y síntesis usando los algoritmos PSOLA y HNM: resíntesis, modificación de prosodia, modificación de prosodia y de *jitter* y de *shimmer* (permitiendo la comparación entre ambos algoritmos de síntesis) y la metodología de interés, es decir, modificación de prosodia y diferentes configuraciones de parámetros de VoQ usando HNM. Con estas pruebas ser observan los mejores resultados de HNM sobre PSOLA, llegando a la metodología de interés donde se consiguen niveles de calidad aceptables (regular dentro de una escala MOS) durante las transformaciones y se muestra como una acertada transformación puede incrementar la percepción de la calidad de la señal de voz.

3. La última de las pruebas perceptivas es la de la identificación del estilo expresivo resultante de la transformación. Está ligada al test de la calidad, ya que se realizó al mismo tiempo y bajo la comparación de las mismas configuraciones metodológicas, permitiendo observar como HNM mantiene mayor estabilidad en los resultados de identificación respecto de PSOLA, demostrándose por parte de la metodología de interés una mejora significativa de los resultados respecto del resto de transformaciones.

De los resultados de la evaluación queda demostrado como la metodología de transformación propuesta, usando la transformación de prosodia y de VoQ de forma conjunta mediante el proceso de análisis y de síntesis del habla basado en HNM, mantiene la calidad durante las transformaciones sin introducir graves degradaciones en el habla generada, ofreciendo mejoras considerables en la identificación de los estilos expresivos destino respecto de usar solamente prosodia y el algoritmo PSOLA.

Algunas propuestas de líneas de futuro, en las que se considera interesante dedicar esfuerzos, son:

- A partir de los resultados obtenidos, se plantea la realización de un modelo de transformación de VoQ más complejo, en la línea en la que se viene trabajando para la prosodia. Por tanto, el proceso de modificación de los parámetros de VoQ deberá de ser adaptado a este modelo.
- En la línea de la mejora del modelo de transformación de la VoQ, sería interesante profundizar en el estudio de la dependencia existente entre los parámetros de VoQ y el orden en el que estos son modificados, pudiendo ser necesario un proceso de ajuste de los valores de aquellos parámetros implicados durante la transformación.
- La realización de experimentos donde se evalúe la degradación de la calidad en función de la magnitud de la transformación llevada a cabo. Conocer el compromiso entre cuánto pueden ser modificados los parámetros (aumentando la capacidad de identificación del estilo destino) sin alterar la calidad del habla generada daría un mayor control al sistema de CTH durante la SHE.
- La mejora continua que ya se está efectuando en el grupo de investigación, tanto en el modelado de la prosodia como en el análisis y la síntesis basada en HNM, participará en la mejora de los resultados de la calidad global y de la identificación de los estilos de habla expresivos.
- El modelado de VoQ y su transformación podría ser aplicado a otros sistemas de CTH como por ejemplo HMM, donde habría que estudiar la mejor forma de ponerlos en práctica.

APÉNDICE A

Aportaciones

A.1. Publicaciones científicas

El trabajo de investigación llevado a cabo durante la realización de la presente tesis ha dado como resultado la participación en diferentes publicaciones científicas, conferencias y revistas, tanto nacionales como internacionales.

Dada la temática abordada en esta tesis, el trabajo desarrollado ha tenido como punto fuerte la elevada cooperación que ha existido con otros miembros del grupo de investigación, ya que con el trabajo llevado a cabo en Cualidad de la Voz —*Voice Quality*— (VoQ) se han explorado los beneficios que podrían darse en distintas aplicaciones (reconocimiento y síntesis del habla expresiva), metodologías de Conversión de Texto en Habla (CTH) (PSOLA, HNM y HMM) y en la evaluación objetiva del habla (en diseño de corpus y sistemas de CTH).

Por tanto, en este apéndice se presentan las publicaciones científicas que ha generado el desarrollo de esta tesis, pudiendo dividirlas en tres bloques principales: diseño y etiquetado de corpus orales, parametrización y modelado de estilos de habla expresivos utilizando VoQ y trabajo realizado en CTH, incluyendo Síntesis del Habla Expresiva (SHE) y su evaluación objetiva. Adicionalmente, se muestran los trabajos llevados a cabo durante los inicios de la etapa investigadora, cuya temática no fue finalmente la asociada al tema principal de esta tesis.

A.1.1. Diseño y etiquetado de corpus orales

Debido a la necesidad de disponer de información para los experimentos con estilos de habla expresivos y el sistema de CTH, uno de los trabajos realizados fue el relativo al diseño y etiquetado de corpus orales. A continuación se presentan, en orden cronológico, las publicaciones que se derivaron.

Conferencias

1. Alías, F., Monzo, C. y Socoró, J. C. (2006). "A Pitch Marks Filtering Algorithm based on Restricted Dynamic Programming". En: *The 9th International Conference on Spoken Language Processing (Interspeech'2006)*, pp. 1698–1701. Pittsburgh, USA.

2. Iriondo, I., Planet, S., Socoró, J. C., Alías, F., Monzo, C. y Martínez, E. (2007). "Expressive speech corpus validation by mapping subjective perception to automatic classification based on prosody and voice quality". En: *The 16th International Congress of Phonetic Sciences (ICPhS'2007)*, pp. 2125–2128. Saarbrücken, Germany.

Revistas

1. Monzo, C., Alías, F., Morán, J. A. y Gonzalvo, X. (2006). "Transcripción fonética de acrónimos en castellano utilizando el algoritmo C4.5". *Procesamiento del Lenguaje Natural*, **37**, pp. 275–282.

A.1.2. Parametrización y modelado de estilos de habla expresivos

A caballo entre el diseño de corpus orales y la SHE se trabajó en la parametrización de la voz utilizando VoQ. Se realizaron experimentos de clasificación de estilos de habla expresivos, útiles en sí mismos y con el objetivo de tanto el diseño de corpus orales como de aplicaciones de CTH. A continuación se presentan, en orden cronológico, las publicaciones que se derivaron.

Conferencias

1. Monzo, C., Alías, F., Iriondo, I., Gonzalvo, X. y Planet, S. (2007). "Discriminating Expressive Speech Styles by Voice Quality Parameterization". En: *The 16th International Congress of Phonetic Sciences (ICPhS'2007)*, pp. 2081–2084. Saarbrücken, Germany.

2. Monzo, C., Iriondo, I. y Martínez, E. (2008). "Procedimiento para la medida y la modificación del jitter y del shimmer aplicado a la síntesis del habla expresiva". En: *V Jornadas en Tecnología del Habla*, pp. 58–61. Bilbao, España.

3. Planet, S., Iriondo, I., Socoró, J. C., Monzo, C. y Adell, J. (2009). "GTM-URL Contribution to the INTERSPEECH 2009 Emotion Challenge". En: *The 10th Annual Conference of the International Speech Communication Association (Interspeech'2009). Special session: Emotion Challenge*, pp. 316–319. Brighton, United Kingdom. ISSN 1990-9772.

Revistas

1. Iriondo, I., Planet, S., Socoró, J. C., Martínez, E., Alías, F. y Monzo, C. (2009). "Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification". *Speech Communication*, **51**, pp. 744–758.

Hay que destacar que, por una parte, la publicación de conferencia número 3, quedó en primera posición en el *Feature Sub-Challenge* y en segunda posición en el *Classifier Sub-Challenge* del *INTERSPEECH 2009 Emotion Challenge*[1]. Por otro lado, la publicación en revista número 1 fue tercera finalista del premio al mejor artículo, en el año 2009, de la Red Temática en Tecnologías del Habla[2].

A.1.3. Conversión de texto en habla

El contexto donde esta tesis ha centrado su objetivo principal, juntamente con el trabajo mostrado en las Secciones A.1.1 y A.1.2, ha sido la CTH. En CTH se incluye la investigación en el sistema de CTH del grupo, nuevos enfoques para su mejora, la SHE y la evaluación objetiva de la misma. A continuación se presentan, en orden cronológico, las publicaciones que se derivaron.

Conferencias

1. Alías, F., Iriondo, I., Formiga, L., Gonzalvo, X., Monzo, C. y Sevillano, X. (2005). "High quality Spanish restricted-domain TTS oriented to a weather forecast application". En: *The 9th European Conference on Speech Communication and Technology (Interspeech'2005)*, pp. 2573–2576. Lisboa, Portugal.

2. Gonzalvo, X., Iriondo, I., Socoró, J. C., Alías, F. y Monzo, C. (2007). "HMM-based Spanish speech synthesis using CBR as F0 estimator". En: *ISCA Tutorial and Research Workshop on Non Linear Speech Processing (NOLISP'2007)*, pp. 7–10. Paris, France.

3. Gonzalvo, X., Socoró, J. C., Iriondo, I.,Monzo, C. y Martínez, E. (2007). "Linguistic and Mixed Excitation Improvements on a HMM-based speech synthesis for Castilian Spanish". En: *The 6th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis (SSW6)*, pp. 362–367. Bonn, Germany.

4. Monzo, C., Formiga, L., Adell, J., Iriondo, I., Alías, F. y Socoró, J. C. (2008). "Adaptación del CTH-URL para la competición Albayzin 2008". En: *V Jornadas en Tecnología del Habla*, pp. 87–90. Bilbao, España.

5. Gonzalvo, X., Taylor, P.,Monzo, C., Iriondo, I. y Socoró, J. C. (2009). "High Quality Emotional HMM-Based Synthesis In Spanish". En: *ISCA Tutorial and Research Workshop on Non Linear Speech Processing (NOLISP'2009)*, Vic, Spain.

[1] http://emotion-research.net/sigs/speech-sig/emotion-challenge/
[2] http://lorien.die.upm.es/~lapiz/rtth/

6. Monzo, C., Calzada, À., Iriondo, I. y Socoró, J. C. (2010). “Expressive Speech Style Transformation: Voice Quality and Prosody Modification Using a Harmonic plus Noise Model”. En: *Speech Prosody*, Chicago, USA.

Revistas

1. Gonzalvo, X., Iriondo, I., Socoró, J. C., Alías, F. y Monzo, C. (2007). “Mixing HMM-Based Spanish Speech Synthesis with a CBR for Prosody Estimation”. En: *Advances in Nonlinear Speech Processing*, volumen 4885/2007 de *Lecture Notes in Computer Science (LNCS)*, pp. 78–85. Springer, Heidelberg.

2. Gonzalvo, X., Taylor, P., Monzo, C., Iriondo, I. y Socoró, J. C. (2010). “High Quality Emotional HMM-Based Synthesis in Spanish”. En: *Advances in Nonlinear Speech Processing*, volumen 5933/2010 de *Lecture Notes in Computer Science (LNCS)*, pp. 26–34. Springer, Heidelberg.

A.1.4. Otras publicaciones

En los inicios de la etapa investigadora se llevaron a cabo trabajos que finalmente no formaron parte de la línea de investigación principal, pero que permitieron tener una primera toma de contacto con la metodología científica necesaria para la realización de la tesis. A continuación se presentan, en orden cronológico, las publicaciones que se derivaron.

Conferencias

1. Monzo, C., Morán, J. A., Planet, S. y Gonzalvo, J. (2004). “Protocolo DS-CDMA orientado a la gestión de un auditorio”. En: *XIX Simposium Nacional de la Unión Científica de Radio (URSI’2004)*, Barcelona, Spain.

2. Planet, S., Morán, J. A., Monzo, C. y Gonzalvo, X. (2004). “Asistencia al seguimiento docente basada en técnicas avanzadas de análisis”. En: *XIX Simposium Nacional de la Unión Científica de Radio (URSI’2004)*, Barcelona, Spain.

3. Gonzalvo, J., Morán, J. A., Monzo, C. y Planet, S. (2005). “Entorno para el aprendizaje automático de estrategias de diálogo”. En: *XX Simposium Nacional de la Unión Científica de Radio (URSI’2005)*, Gandia, Spain.

A.2. Comités técnicos

Además de las aportaciones realizadas como autor, se ha formado parte del programa del comité técnico de las siguientes publicaciones científicas:

1. SAMT 2009, 4th International Conference on Semantic and Digital Media Technologies

2. WMSCI 2008, 12th Worl Multi-Conference on Systematics, Cybernetics and Informatics

A.3. Proyectos de investigación y desarrollo

En esta sección se describen los proyectos de investigación y desarrollo en los que esta tesis se vio enmarcada. Estos proyectos, de financiación pública y privada, se realizaron como miembro del grupo de investigación GTM.

A.3.1. Financiación pública

SALERO: Semantic AudivisuaL Entertaiment Reusable Objects (FP6/2004/IST/4)

Proyecto financiado por el VI Programa Marco de la Unión Europea (2006-2009)[3]. Participaron trece socios entre empresas y centros de investigación. Su objetivo era facilitar la creación de nuevos productos multimedia como juegos, películas o programas de televisión, haciéndola mejor, más rápida y con menos costes gracias a la combinación de gráficos por ordenador, tecnología del habla y el lenguaje, web semántica y búsquedas basadas en contenido.

IntegraTV4all (FIT-350301-2004-2)

Proyecto financiado por el Ministerio de Ciencia y Tecnología (2004-2005)[4]. Fue un proyecto de I+D, desarrollado por la ingeniería de software TMT Factory junto con la fundación ONCE, las universidades Carlos III y Politécnica de Madrid, que tenía como objetivo el desarrollo de servicios adaptados de ocio, información y tele-trabajo, a través de la televisión, para hoteles. A estos servicios se les dotó de funcionalidades avanzadas de visión y habla asistida para facilitar la estancia a huéspedes con discapacidades sensoriales.

A.3.2. Financiación privada

Módulo sintetizador de voz para aplicación a meteorología

Proyecto financiado por la *Corporació Catalana de Ràdio i Televisió* (CCRTV) (2004-2005)[5]. Se desarrolló un módulo de síntesis de habla de alta calidad, en un dominio restringido al sistema de generación automática de previsiones meteorológicas, sincronizado con el personaje virtual.

A.4. Participación en otros eventos

Como parte de la actividad investigadora se llevó a cabo la participación en diferentes eventos, no asociados ni a publicaciones científicas ni a la participación en proyectos.

[3] http://www.salero.info
[4] http://research.tmtfactory.com/index.php/tmtresearch/projects/12/
[5] http://www.meteosam.com/

European Masters in Language and Speech (Utrecht, 10-14 de julio de 2006)

Enmarcado en el curso de verano *7th summer school of the European Masters in Language and Speech*[6], organizado por la Universiteit Utrecht, se presentó un póster donde se mostraba la línea de investigación que el autor de esta tesis estaba desarrollando.

SALERO *Open House* (Helsinki, 13 de junio de 2007)

Participación como ponente en este evento internacional[7], donde se presentaron los resultados de la investigación de nuestro grupo de investigación dentro del proyecto SALERO (Apéndice A.3.1).

Dia de la Ciència a les Escoles (2008-2009)

Se trata de una actividad realizada durante la *Setmana de la Ciència*[8], en la que de forma simultánea, científicos transmiten su experiencia investigadora en escuelas de bachillerato y de formación profesional repartidas por toda Catalunya. En dos ocasiones: Badia del Vallès (2008) y Santa Coloma de Gramenet (2009), el autor de la presente tesis ha tenido la oportunidad de participar presentando temas relacionados con la síntesis de personajes virtuales con emociones.

[6] http://tstmaster.let.uu.nl/EMaster-SS2006/
[7] http://www.salero.info/6165caec3e36f8be1800d98d7cbe33f9/en/events/salero_open_house.html
[8] http://www.setmanaciencia.org/

APÉNDICE B

HNM para el modelado de la cualidad de la voz

Este apéndice se trata de un complemento a las explicaciones llevadas a cabo sobre los experimentos presentados en el Capítulo 5. Los temas que se tratan en las siguientes secciones son:

- Introducir la parametrización del habla basada en HNM empleada durante los experimentos (Apéndice B.1).
- Presentar las consideraciones que se deben de tener en cuenta durante la parametrización de la VoQ a partir del HNM utilizado (Apéndice B.2), que es fundamental para los procesos de análisis y de modificación de la VoQ.
- Establecer la notación que se ha seguido, para el uso de HNM, en el modelado de la VoQ (Apéndice B.3).
- Profundizar en los factores aplicados durante la transformación de estilos de habla expresivos (Sección 5.2.3). Estos factores de son:
  - Factor de modificación de la VoQ (β) (Apéndice B.4), mostrando la metodología de aplicación.
  - Factor correctivo de la energía (α) (Apéndice B.5), del que se explica su obtención y su desarrollo matemático.

B.1. Parametrización HNM utilizada

En la parametrización del habla basada en HNM, la señal de voz ($x(n)$) se puede expresar, tal y como se muestra en la Ecuación B.1, como la suma de una componente determinista o armónica ($s(n)$) y de otra estocástica o de ruido ($r(n)$) (Laroche et al., 1993) para cada muestra n. La parametrización empleada en los experimentos del Capítulo 5, se basa en la implementación *pitch* sincrónica realizada por uno de los

miembros del grupo de investigación del GTM (Calzada, 2008, 2010) (Ecuaciones B.1 a B.6).

$$x(n) = s(n) + r(n) \tag{B.1}$$

La banda espectral más baja se modela, principalmente, con una suma de sinusoides ($s(n)$) relacionadas armónicamente (componente determinista o armónica) que caracterizan la parte sonora de la señal de voz. La Ecuación B.3, siguiendo la nomenclatura definida en el Apéndice B.3, muestra este proceso, en el que la componente determinista queda completamente caracterizada mediante la variación en el tiempo de las **amplitudes** (A_{ik}), las **frecuencias** (F_{ik}) y las **fases** (ϕ_{ik}) de estas sinusoides.

$$F_{ik} = F_{0k} \cdot i \tag{B.2}$$

$$s_k(n) = \sum_{i=1}^{I_k} A_{ik} \cdot \cos(2\pi \cdot F_{ik} \cdot n + \phi_{ik}) \ , \ -T_{k-1}^s \leq n < T_k^s \ , \ T_k^s = t_{k+1}^s - t_k^s \tag{B.3}$$

La $s(n)$ generada en el periodo de trama que va de $n \in \left[t_k^s, t_{k+1}^s\right]$, a partir de la interpolación lineal de las medidas realizadas en el centro de las dos tramas adyacentes, se expresa finalmente como se muestra en la Ecuación B.4.

$$s(t_k^s + m) = \left(\frac{T_k^s - m}{T_k^s}\right) \cdot s_k(m) + \left(\frac{m}{T_k^s}\right) \cdot s_{k+1}(m - T_k^s) \ , \ 0 \leq m < T_k^s \tag{B.4}$$

Los sonidos sordos y todos los eventos no-periódicos del habla quedan modelados por la componente estocástica o de ruido. Esto se realiza mediante un modelo Autorregresivo (AR) (Apéndice F.1) donde tanto las fluctuaciones espectrales como temporales quedan representadas por los **coeficientes LPC** y la **varianza**. En las Ecuaciones B.5 y B.6 se muestran las ecuaciones que caracterizan al ruido, siendo presentada la nomenclatura empleada en el Apéndice B.3.

$$r(n) = \sum_{j=1}^{J} \sigma_j \cdot r_j \left(n + (j-1) \cdot T^r\right) \ , \ 0 \leq n < J \cdot T^r \tag{B.5}$$

$$r_j(n) = \begin{cases} b(n) - \sum_{q=1}^{Q} c_{qj} \cdot r_j(n-Q) \ , \ 0 \leq n < T^r \\ r_{j-1}(n + T^r) \ , \ -Q \leq n \leq -1 \end{cases} \tag{B.6}$$

La implementación del análisis está basado en el algoritmo en el dominio de la frecuencia de Depalle y Helie (1997). Aunque el HNM implica que la banda más baja del espectro de la señal es armónica, este algorimo no garantiza la armonicidad de las frecuencias estimadas, de forma que la solución ha pasado por su modificación utilizando el procedimiento de optimización de los multiplicadores de Lagrange (Moon y Stirling, 2000).

B.2. Parámetros de VoQ a partir de parámetros de HNM

En esta sección se muestran las consideraciones a tener en cuenta en la parametrización de la VoQ a partir de los parámetros de HNM:

- Consideraciones previas (Apéndice B.2.1).
- Cálculo de la energía (Apéndice B.2.2).
- Subbandas (Apéndice B.2.3).

B.2.1. Consideraciones previas

Para el cálculo de los parámetros de VoQ mediante el uso de la parametrización de la señal con el modelo HNM se ha realizado las siguientes consideraciones:

- Los periodos de integración para el cálculo de energías en diferentes bandas y componentes de la señal se ha basado en los periodos de análisis usados en la representación con el propio modelo de HNM (ver Apéndice B.1, donde se indica el modelo usado así como aspectos de detalle de las tramas de análisis). Se asume que la duración de las tramas (periodo de integración) es mayor que el periodo de *pitch*.
- Se ha realizado un cálculo de valores promedio para un segmento de voz dado, dentro del cual la parametrización con HNM posee distintas duraciones de trama para las componentes determinista y estocástica. Para llevar a cabo el proceso de promediado se ha aplicado en algunos casos (p. ej. HNR), una aproximación que conlleva un ahorro computacional en la medida. En cada caso se indica la aproximación realizada. Exceptuando el caso de la medida HNR, en el resto de parámetros de VoQ se ha hecho únicamente uso de la parte armónica del modelo HNM y se ha prescindido de la información aportada por la componente estocástica. Cabe puntualizar que la componente armónica del modelo utilizado ha sido analizada en la banda de frecuencias de $f \in [0Hz, 5000Hz]$, con lo que se estiman las amplitudes, las fases y las frecuencias de todos los armónicos que caben dentro de ella para cada trama analizada. Esta decisión se ha tomado para simplificar tanto el análisis como la síntesis, asumiendo que la repercusión de la componente ruidosa sobre estos parámetros no sea significativa perceptivamente.

B.2.2. Cálculo de energías

Para el estudio de los parámetros de VoQ a partir del modelo HNM de la señal de voz, se requiere del cálculo de las energías de las componentes armónicas y de ruido. Asumiendo un modelo HNM donde la parte armónica se analiza con tramas de longitud T_k^s muestras, donde k es el índice de trama, y K es el número total de tramas analizadas, y la parte ruidosa se analiza con tramas de longitud T^r muestras y J es el número total de tramas analizadas, veamos cómo se expresará la energía de una componente aislada.

B.2.2.1. Energía de la componente determinista

Para una de las componentes de la componente determinista y una trama dada (el armónico o sinusoide i en la trama $k-$ésima, con amplitud A_{ik}) estimamos el valor E^s_{ik}, su energía durante un periodo de integración igual al periodo de la trama T^s_k, como:

$$E^s_{ik} = \int_{T^s_k} A^2_{ik} \cdot \cos^2(2\pi \cdot F_{ik} \cdot t + \phi_{ik})dt \approx \frac{A^2_{ik}}{2} \cdot T^s_k \tag{B.7}$$

Donde en el término de energía E^s_{ik} se omite el periodo de muestreo, ya que sería un factor constante a lo largo de todos los análisis, para simplificar cálculos.

Para obtener la energía de todos los armónicos de la componente determinista del segmento de señal de voz a lo largo de la trama $k-$ésima, realizamos el sumatorio para todos los armónicos presentes en la trama:

$$E^s_k = \sum_{i=1}^{I_k} E^s_{ik} = \sum_{i=1}^{I_k} \frac{A^2_{ik}}{2} \cdot T^s_k = T^s_k \cdot \sum_{i=1}^{I_k} \frac{A^2_{ik}}{2} \tag{B.8}$$

B.2.2.2. Energía de la componente estocástica

Para la componente estocástica estimamos el valor E^r_j, su energía durante un periodo de integración igual al periodo de la trama T^r, como:

$$E^r_j = \sigma^2_j \cdot T^r \tag{B.9}$$

Siendo σ^2_j la varianza de la trama j de la parte ruidosa.

B.2.3. Subbandas

En el caso de los parámetros de VoQ en los que debe de realizarse una distinción entre subbandas (do1000, HammI y pe1000), se define el índice de armónico i^f_k como el asociado al armónico con frecuencia F_{ik} mínima mayor o igual a la frecuencia de corte f.

$$i^f_k = \text{mín}\,(i \mid F_{ik} \geq f) \tag{B.10}$$

De este modo, dada una determinada frecuencia de corte $f \in \{1000, 2000\}$ Hz y una determinada trama de análisis, podemos definir el conjunto de armónicos que están por debajo de f (es decir, $\{1, 2, \ldots, i^f_k\}$), así como el conjunto de armónicos que está por encima de esta (es decir, $\{i^f_k+1, i^f_k+2, \ldots, I_k\}$) de la componente determinista de la señal de voz.

B.3. Notación

La notación seguida, para el uso de HNM en el modelado de la cualidad de la voz, es la siguiente:

- Señal de voz

- $x(n)$: señal de voz que contiene una componente determinista y otra estocástica.
- $s(n)$: señal de voz correspondiente a la componente determinista.
- $r(n)$: señal de voz correspondiente a la componente estocástica.
- n: índice de la muestra temporal.

- Componente determinista (armónica):
  - A_{ik}: amplitud del armónico i en la trama $k-$ésima.
  - F_{ik}: frecuencia del armónico i en la trama $k-$ésima.
  - ϕ_{ik}: fase del armónico i en la trama $k-$ésima.
  - k: índice de trama correspondiente a un periodo de análisis del HNM.
  - K: número de tramas en las que la señal de voz original ha quedado dividida durante el análisis.
  - i: índice del armónico de interés en la trama $k-$ésima.
  - I_k: número máximo de armónicos i en la trama $k-$ésima.
  - i_k^f: índice de armónico asociado al armónico con frecuencia F_{ik} mínima mayor o igual a la frecuencia de corte f (Apéndice B.2.3).
  - t_k^s: instante de análisis (expresado en muestras) de la trama $k-$ésima de la componente determinista (en el caso que nos ocupa corresponde con la marca de *pitch*).
  - T_k^s: diferencia entre instantes de análisis consecutivos (en nuestro caso es igual al periodo de *pitch*) de la componente determinista.
  - T_{seg}: duración total (expresada en número de muestras) correspondiente al segmento de señal de voz.
  - m: índice de la componente determinista para $n \in \left[t_k^s, t_{k+1}^s\right]$.
  - E_k^s: energía de la trama k de la parte armónica (o sonora) de la voz.
  - E_{ik}^s: energía durante un periodo de integración igual al periodo de la trama T_k^s para una de las componentes de la parte armónica y una trama k dada (el armónico o sinusoide i en la trama k, con amplitud A_{ik}).
  - $(E_{ik}^s)'$: energías resultantes de la modificación de E_{ik}^s.
  - A'_{ik}: amplitudes resultantes de la modificación de A_{ik}.

- Componente estocástica (ruido):
  - σ_j^2: varianza de la trama $j-$ésima.
  - E_j^r: energía de la trama $j-$ésima.
  - j: índice de la trama.
  - J: número de ventanas de análisis en las que la componente estocástica ha sido dividida.

- T^r: tamaño de la trama de ruido expresada en número de muestras (en el caso que nos ocupa es un parámetro constante del modelo).
- $b(n)$: ruido blanco Gaussiano de potencia 1.
- q: índice del coeficiente del filtro LPC de la trama de ruido $j-$ésima.
- Q: número de coeficientes del filtro LPC (en el caso que nos ocupa es un parámetro constante del modelo).
- c_{qj}: coeficiente q del filtro LPC para la trama de ruido $j-$ésima.

Para el caso de la componente determinista, se tiene la información de las amplitudes, las frecuencias y las fases, ya que para cada periodo de *pitch*, correspondiente a cada trama de la componente determinista, se hace un análisis de amplitudes por armónico. En cuanto a la componente estocástica, el número de tramas en las que se divide, así como su tamaño, depende de un parámetro de configuración del modelo, fijando el tamaño de la ventana en el que se calcula la varianza del ruido.

B.4. Factor de modificación de la cualidad de la voz (β)

En las Secciones 5.2.3 y 5.5 se muestra la modificación de los parámetros del HNM para la transformación de estilos de habla expresivos. Este proceso se realiza mediante la aplicación del modelado de los parámetros de VoQ, identificando tres pasos a seguir:

- Conocimiento del valor en el estilo de habla expresivo destino.
- Análisis de los parámetros que indica el modelo de transformación.
- Factor de modificación del parámetro de interés en el habla original (β).

En primer lugar, el conocimiento de los valores a asociar en cada transformación se explica en la Sección 5.5, con lo que no se entra a discutir en esta sección. En segundo lugar, en referencia al análisis de los parámetros de VoQ, este se lleva a cabo para todos los parámetros que se vean implicados en la transformación, siguiendo el procedimiento descrito en Sección 5.2.2.

El factor de modificación β se aplica como un factor de energía multiplicativo sobre los parámetros del HNM para cada trama de análisis k. Puede ser directamente indicado a la función de transformación como un valor lineal, aunque existe la posibilidad de ser calculado si se desea utilizar un valor destino conocido del parámetro, por medio del cociente entre el valor destino deseado (indicado como parámetro de entrada a la función de transformación) y el valor analizado, ambos expresados en lineal. El *jitter* y el *shimmer*, tal y como se presenta en la Sección 5.3.3, difieren del resto de parámetros en el uso de este factor de modificación β, ya que trabajan con un valor destino de potencia de ruido esperado, con lo que la explicación encontrada en esta sección no se les aplica.

Analíticamente, considerando un factor indicado como parámetro de la transformación, el valor de β será constante para todas las tramas k analizadas, tal y como

ocurre en los experimentos presentados en la Sección 5.5.3 ($\beta_k = \beta$). No obstante, cabría la posibilidad de poder indicar un valor destino deseado para el parámetro de VoQ en cuestión, siendo calculado el factor β para cada una de las tramas como el cociente entre los valores del parámetro deseado y el resultante del proceso de análisis.

$$\beta_k = \frac{VoQ^d}{VoQ^o_k} \tag{B.11}$$

B.5. Factor correctivo de la energía (α)

Tal y como se presenta en la Sección 5.2.3, el factor α es un factor correctivo de la energía, aplicado durante la modificación de parámetros de VoQ que implican la manipulación de la energía de diferentes bandas de frecuencia.

El objetivo es que la energía global de la señal de voz resultante se mantenga invariante a lo largo de todas las modificaciones que pudiera sufrir. Para ello, partiendo de la energía de cada una de las bandas de frecuencia ($\hat{E}_k$ y $\check{E}_k$) y el factor de modificación β_k (Apéndice B.4) por trama k de análisis de la componente determinista (en el caso que este fuera cambiando a lo largo de la transformación), se calcula el factor de corrección de la energía en la trama $k-$ésima (α_k).

Se considera que cada banda de frecuencias se ve afectada en un factor tal que la relación entre $\hat{E}_k$ y $\check{E}_k$ es β_k, siendo estos valores $\sqrt{\beta_k}$ y $\frac{1}{\sqrt{\beta_k}}$ respectivamente. En las Ecuaciones B.13 a B.19 se muestra el desarrollo de la expresión analítica de este factor correctivo de la energía α_k en la trama $k-$ésima.

Se parte de la premisa de que se desea que la energía final se mantenga después de aplicar el factor de modificación β_k (Ecuación B.12).

$$E'_k = E_k \tag{B.12}$$
$$E_k = \check{E}_k + \hat{E}_k \tag{B.13}$$
$$E'_k = \check{E}'_k + \hat{E}'_k \tag{B.14}$$
$$\check{E}'_k = \sqrt{\beta_k} \cdot \check{E}_k \tag{B.15}$$
$$\hat{E}'_k = \sqrt{\frac{1}{\beta_k}} \cdot \hat{E}_k \tag{B.16}$$

A partir de la premisa de mantener la energía antes y después de modificar cada una de las bandas de frecuencia (Ecuación B.12), se plantea el cálculo del factor de corrección α_k (Ecuaciones B.17 y B.18).

$$\alpha_k \cdot (\check{E}'_k + \hat{E}'_k) = \check{E}_k + \hat{E}_k \tag{B.17}$$
$$\alpha_k \cdot \left(\sqrt{\beta_k} \cdot \check{E}_k + \sqrt{\frac{1}{\beta_k}} \cdot \hat{E}_k \right) = \check{E}_k + \hat{E}_k \tag{B.18}$$

Finalmente, el factor de corrección α_k se expresa tal y como se muestra en la Ecuación B.19:

$$\alpha_k = \frac{\check{E}_k + \hat{E}_k}{\sqrt{\beta_k} \cdot \check{E}_k + \frac{1}{\sqrt{\beta_k}} \cdot \hat{E}_k} \tag{B.19}$$

Donde:

- E_k: energía en la trama $k-$ésima, calculado como la suma de la energía de las dos bandas frecuencias modificadas.
- E'_k: nuevo valor de energía en la trama $k-$ésima.
- $\check{E}_k$, $\hat{E}_k$: energía asociada a cada banda modificada en la trama $k-$ésima.
- $\check{E}'_k$, $\hat{E}'_k$: nueva energía, en la trama $k-$ésima, de $\check{E}$ y $\hat{E}$ respectivamente.

APÉNDICE C

Frases utilizadas en las pruebas subjetivas

En este apéndice se presentan las frases empleadas en las pruebas subjetivas realizadas durante la evaluación de los experimentos presentados en esta tesis.

C.1. Análisis y modificación del *jitter* y del *shimmer*

En la evaluación de la metodología de análisis y de síntesis propuesta para los parámetros *jitter* y *shimmer* (Sección 5.3.4), los enunciados utilizados, seleccionados del corpus de frases neutras del grupo de investigación GTM y sintetizados hacia los otros cuatro estilos de habla expresiva de los que está formado (Sección 4.3), fueron los siguientes:

1. ¿Te has fijado en los móviles de los demás?
2. A veces la clave del éxito está en los pequeños detalles.
3. A trece mil metros de altura, cuando sus ideas surgen como un torrente, su pluma debe plasmarlas, con fluidez, sin que gotee.
4. ¿Su sistema de comunicación, podrá adaptarse al cambio que le exija el futuro?
5. ¡Papa he suspendido todas!

C.2. Transformación de estilos de habla expresivos

En la evaluación de la metodología de transformación de estilos de habla expresivos, utilizando prosodia y VoQ junto a la parametrización del habla basada en HNM (Sección 5.5.4), los enunciados utilizados dependieron del tipo de test realizado y de la configuración aplicada. Fueron dos las pruebas llevadas a cabo, ambas empleando el corpus de frases del grupo de investigación GTM (Sección 4.3).

En el primer test los enunciados utilizados fueron elegidos del corpus neutro (Sección 5.5.4.1):

1. ¿Se imagina un teléfono con trescientos metros de cable?
2. ¿Su sistema de comunicación, podrá adaptarse al cambio que le exija el futuro?
3. ¿Te has fijado en los móviles de los demás?
4. Cien por cien garantizado en todos los equipos de la oficina.
5. Mil novecientos noventa y seis va a ser el año más feliz de tu vida.
6. Doscientas cincuenta mil pesetas diarias te cambiarán la vida.
7. A trece mil metros de altura, cuando sus ideas surgen como un torrente, su pluma debe plasmarlas, con fluidez, sin que gotee.
8. Ponemos la tecnología, a su alcance.

La evaluación del segundo test se realizó para diferentes configuraciones (Sección 5.5.4.2), con lo que los enunciados no fueron siempre los mismos (p. ej. habla natural necesitó de los audios del corpus en cada uno de los estilos expresivos).

- Habla natural y resíntesis con PSOLA y HNM:
  - Neutro
    1. ¿Se imagina un teléfono con trescientos metros de cable?
    2. ¿Su sistema de comunicación, podrá adaptarse al cambio que le exija el futuro?
    3. ¿Te has fijado en los móviles de los demás?
    4. Cien por cien garantizado en todos los equipos de la oficina.
    5. Mil novecientos noventa y seis va a ser el año más feliz de tu vida.
  - Alegre
    1. La condición, indispensable.
    2. Esta colección, es la joya de mi biblioteca.
    3. ¡La diversión, es nuestra historia!
    4. Veintiocho premios, de ciento veinticinco millones.
    5. Alrededor del arte, siempre hay gente muy especial.
  - Sensual
    1. ¡Lo tiene, todo!
    2. ¡Fuera el estrés!
    3. ¿Y si se pudiese retrasar el tiempo?
    4. Al servicio de la belleza.
    5. Hidratación absoluta, nuevas sensaciones para una piel suave.

- Agresivo
  1. Actuaciones responsables.
  2. Quien conduce, lo sabe.
  3. Antes de elegir destino, elige compañero.
  4. Tú eres tú, y tus circunstancias.
  5. Lo demás es completamente distinto.
- Triste
  1. La mejor diversión, de tu vida.
  2. ¿Cuántas estrellas, tiene tu hotel?
  3. El verde más intenso, está cerca de ti.
  4. La historia, al alcance de tu mano.
  5. Una pequeña diferencia, puede significarlo todo.

- Modificación de prosodia y de VoQ:
  1. ¿Se imagina un teléfono con trescientos metros de cable?
  2. ¿Su sistema de comunicación, podrá adaptarse al cambio que le exija el futuro?
  3. ¿Te has fijado en los móviles de los demás?
  4. Cien por cien garantizado en todos los equipos de la oficina.
  5. Mil novecientos noventa y seis va a ser el año más feliz de tu vida.

APÉNDICE D

Diseño de corpus

En este apéndice se presenta el trabajo desarrollado en relación al diseño de corpus, mostrando la selección de textos (Apéndice D.1), la herramienta para llevar a cabo la segmentación (Apéndice D.2), la evaluación de sistemas de marcado de *pitch* (Apéndice D.3) y, por último, la herramienta creada para el análisis del etiquetado (Apéndice D.4).

D.1. Selección de textos

D.1.1. Gestión de excepciones en la selección de textos

Los textos utilizados para crear corpus de voz pueden provenir de multitud de fuentes (Torruella y Llisterri, 1999), de forma que no se tiene un control total sobre su contenido. Es por esta razón que se pueden encontrar desde una Palabra no Estándar —*Non-Standard Word*— (NSW), como son números (incluyendo romanos), abreviaturas o acrónimos, a palabras que no pertenecen a la lengua de interés. Esto es un problema desde el punto de vista del transcriptor fonético, que puede interpretar incorrectamente estas palabras y añadir errores al corpus creado, produciendo a su vez una posible pérdida de calidad en el sistema posterior que lo utilice. Lo que se pretende conseguir con este trabajo es minimizar la aparición de NSW y extranjerismos, de manera que se facilite el proceso de transcripción fonética por no necesitarse la creación de un diccionario de excepciones para controlar estos casos ni una revisión manual, incrementando así la calidad del corpus creado por no contener unidades mal etiquetadas. Debido al gran volumen de datos del que se disponía, en lugar de etiquetar estos casos y corregirlos (Sproat et al., 1999), lo que se plantea es eliminarlos de los textos candidatos a formar parte del corpus.

El gestor de excepciones que se ha implementado es un sistema basado en reglas. Se considera que un usuario experimentado conoce las posibles excepciones

que se pueden encontrar en la lengua de interés o bien las limitaciones del sistema de transcripción, de manera que se agrupan las posibles excepciones por categorías, indicándose en un archivo de configuración pensado con este fin. El sistema puede ser empleado para cualquier lengua, donde una serie de reglas identifiquen los casos que se desea que sean omitidos de los posibles textos que formarán el corpus final. Este software se ejecuta mediante un *script* desarrollado en Matlab, donde como entrada se tiene el archivo conteniendo todos los textos, el de salida donde se guardarán los textos que superen el filtrado y el archivo de configuración con las reglas que serán aplicadas. A continuación se detallan los diferentes grupos de excepciones que pueden ser controlados por el archivo de configuración:

- **punct_mark**: signos de puntuación que pueden aparecer en los textos y que servirán para tenerlos en cuenta durante la búsqueda de las distintas excepciones.
- **vowel_group**: aquellos caracteres considerados vocales que servirán para tenerlos en cuenta durante la búsqueda de las distintas excepciones.
- **del_only_cont**: se eliminará de dentro de la frase aquel texto que esté delimitado por las parejas de caracteres indicados.
- **del_num**: se eliminarán todas aquellas frases que contengan números.
- **del_char**: se eliminarán todas aquellas frases que contengan alguno de los caracteres señalados.
- **del_rep**: se eliminarán todas aquellas frases que tengan un carácter de manera consecutiva tantas veces como se indique.
- **del_start**: se eliminarán todas aquellas frases que contengan palabras comenzadas por el carácter indicado. Adicionalmente se puede indicar aquellos caracteres que siguiendo a este anulan la excepción.
- **del_cons_group**: se eliminarán aquellas frases formadas únicamente por consonantes, de modo que serán probablemente abreviaturas o acrónimos. El parámetro indicado señala la medida máxima de palabra analizada.
- **del_roman_number**: se eliminarán aquellas frases donde aparecen números romanos.

D.1.2. Eliminación de frases similares

Se han desarrollado *scripts* Matlab que nos facilitarán el proceso de selección de frases para la grabación de un corpus oral. A partir de un conjunto de textos no se seleccionarán aquellos que sean similares al resto en el factor de similitud expresado como un tanto por ciento deseado.

El funcionamiento está basado en la descomposición de los textos en n-gramas, esto es utilizar ventanas de n caracteres, donde n es un parámetro configurable (p.

ej. igual a 3), de forma que se contabilizará por frase su aparición para la posterior selección de aquellas que cumplan con el requisito de similitud deseada. La posición del n-grama dentro de la frase no se tendrá en cuenta, y en caso de haber dos frases similares se escogerá la más larga por cuestiones de diseño, ya que contendrá un mayor volumen de información. Los pasos para llevar a cabo esta selección son:

- Reordenación de las frases de la más a la menos larga.
- Creación de la tabla de n-gramas por frase.
- Búsqueda de la frase de interés dentro de la tabla de n-gramas.

Para la medición de la similitud se usarán todos los caracteres presentes en la frase pasándose previamente a minúsculas. Cuanto mayor sea el factor de similitud mayor probabilidad de hacer que la frase bajo estudio se añada a la lista de textos seleccionados.

D.2. Herramienta para la segmentación

La herramienta desarrollada para llevar a cabo la segmentación del corpus se basa en una serie de *scripts* y de interfaces gráficas, escritos en Matlab, que se comunican con diferentes módulos de la herramienta *Hidden Markov Model Toolkit* (HTK).

Para cada voz distinta que se desee segmentar es necesario realizar un entrenamiento de los modelos, de forma que esta quede caracterizada. En el caso de ser una misma voz, pero que ha cambiado por ejemplo de lengua o de expresividad, será necesario un nuevo entrenamiento que la modele correctamente.

En cuanto a los modelos de duración y límites de cada una de las unidades, se utilizará:

i. Información de segmentación etiquetada manualmente de manera que sirva de referencia para extraer los modelos.

ii. Sin ninguna información de referencia de modo que los modelos se extraerán en base al análisis de la señal de voz del corpus.

Los parámetros que modificarán los distintos modelos son los que se muestran a continuación (Young et al., 2006):

- Tipo de codificación de la voz (p. ej. MFCC con primera y segunda derivada).
- Tamaño y paso de ventana (p. ej. 15 y 5 ms respectivamente).
- Tipo de ventana a usar (p. ej. Hamming).
- Coeficiente del filtro de preénfasis.
- Número de canales del banco de filtros.
- Número de coeficientes.

- Número de estados en el HMM.
- Inicialización de las estadísticas de los HMM.
- Tipo de HMM, es decir, si deseamos que el modelo solamente permita saltos de izquierda a derecha, de izquierda a derecha con saltos paralelos, o saltos entre cualquier estado (modelo ergódico).

La ventaja de realizar el entrenamiento utilizando la interfaz gráfica desarrollada, es que permite variar los diferentes parámetros involucrados de una forma cómoda y sencilla. A continuación se presenta el caso del entrenamiento más genérico y automático, que se dará cuando el usuario no disponga de archivos previos de segmentación para la voz que está entrenando (Figuras D.1 y D.2). En la Figura D.1 se carga el archivo de configuración total del entrenamiento, mientras que en la Figura D.2 podremos elegir, mediante el menú desplegable, el archivo de configuración del HTK que se desea visualizar y, si fuera necesario, modificar. La Figura D.2 se carga cuando desde el menú principal se desea editar el archivo de configuración. Así pues, además de implementar el software que llevará a cabo el entrenamiento, se dispone de una interfaz de usuario que permite fácilmente controlar los parámetros que se usarán.

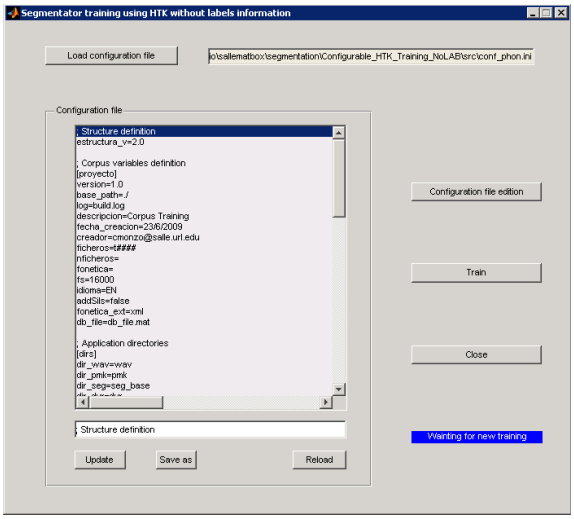


Figura D.1: *Manipulación del archivo de configuración general para el entrenamiento del segmentador*

Una vez concluido el entrenamiento de los modelos, se pasa a realizar el etiquetado de los límites de cada una de las unidades o segmentación. La segmentación se realizará sobre todos los archivos que forman el corpus, de modo que el resultado formará parte de su etiquetado. Existen dos posibilidades en cuanto a la segmentación, según lo que se señale en el archivo de configuración utilizado para indicar los distintos parámetros. Podemos obtener únicamente el inicio y final de cada una de las unidades entrenadas, las marcas límite de fonema y otras equidistantes dentro de este margen (usado por el sistema de CTH durante la síntesis), o bien las marcas para cada

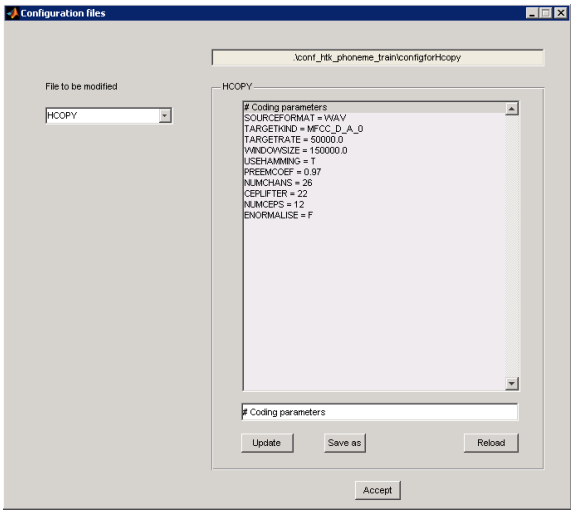


Figura D.2: *Manipulación de los archivos de configuración de parámetros para el entrenamiento del segmentador*

uno de los estados internos por los que el modelo de Markov ha pasado, obteniendo de esta manera las zonas de transición (zonas de estabilidad) de los fonemas.

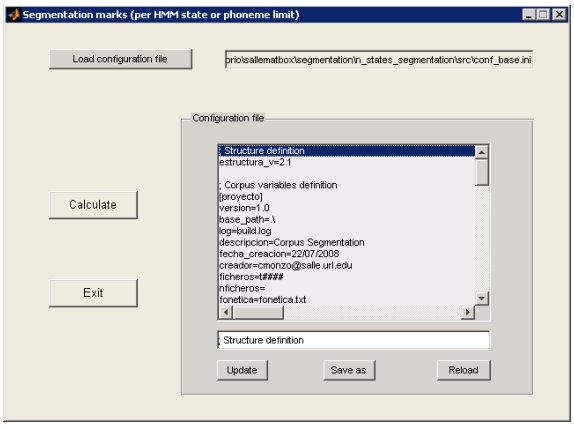


Figura D.3: *Interfaz de usuario para la segmentación de un corpus de voz*

De igual forma que para el entrenamiento, se dispone de una interfaz de usuario desarrollada en Matlab que permite configurar todo lo necesario para llevar a cabo la segmentación de los archivos de audio Figura D.3.

D.3. Evaluación de sistemas de marcado de *pitch*

Por lo que hace referencia a la evaluación de sistemas de marcado de *pitch*, existen medidas de evaluación de PDA bien conocidas, como por ejemplo el conocido como Tasa de Errores Grandes —*Gross Error Rate*— (GER), que computa como errores las estimaciones de F_0 con valores un $20\,\%$ por encima o debajo de los valores de

referencia (de Cheveigné y Kawahara, 2002; Sun, 2002). Por el contrario, para el caso de PMA no hay medidas de test estándar para su evaluación. Todas las propuestas existentes solamente pueden ser consideradas si las marcas evaluadas y las de referencia siguen el mismo criterio de colocación de marcas. Para tratar este problema, Harbeck et al. (1995) y Kounoudes et al. (2002) proponen alinear las marcas antes de llevar a cabo la comparación. Sin embargo, los errores debidos a la desalineación pueden conducir a resultados de la evaluación poco fiables.

A raíz de la colaboración con Alías et al. (2006), se ha trabajado en una nueva medida de evaluación con el objetivo de validar y comparar las prestaciones de diferentes PMA con distinto criterio de ubicación de marcas. Para ello, se consideran las diferencias relativas de periodicidad (p_r) de marcas de *pitch* consecutivas en lugar de su localización específica. Si esta diferencia es mayor que un umbral γ predefinido, esa marca de *pitch* se considera como errónea. A esta medida se la ha llamado *Gross Pitch Marks Error Rate* (GPMER) (Ecuación D.1) y puede ser definido como un GER fino, ya que el GER solamente compara el comportamiento del PDA a nivel de ventana. Las marcas de *pitch* de referencia guían el proceso de comparación a través de la señal, por tanto, las inserciones y omisiones son fácilmente detectadas. Para evitar predisponer la medida de evaluación, tanto las inserciones como las omisiones son solamente computadas como un único error.

$$GPMER(\%) = \frac{\#\left(\frac{|p'_r - p_r|}{p_r}\right) > \gamma}{\#p_r} \cdot 100 \tag{D.1}$$

Donde $\#$ indica "número de", p'_r es la periodicidad local evaluada estimada por el PMA y p_r es el valor de referencia correspondiente, cuya posición global guía el proceso de comparación para conseguir comparar vectores de marcas de *pitch* de longitudes distintas. En cuanto al valor umbral (γ), se utilizará un valor de $\gamma = 0,2$ del mismo modo que en el GER clásico.

Esta nueva medida nos sirvió para evaluar el comportamiento del Algoritmo de Filtrado de Marcas de *Pitch —Pitch Marks Filtering Algorithm—* (PMFA) (Sección 4.3.4), incluyéndose los errores de sonoridad también en los resultados. El análisis se llevó a cabo sobre dos bases de datos. La primera, un corpus de voz en castellano (DB1) grabado por una locutora profesional, muestreado a 16 KHz con 16-bit de resolución, usando tres estilos de habla expresivos distintos (Sección 4.3): neutro, alegre y sensual; por tanto con diferentes márgenes de F_0. La segunda, la base de datos Keele (Plante et al., 1995) (DB2), que contiene voces de 10 locutores (5 hombres y 5 mujeres), muestreadas a 20 KHz con 16-bit de resolución (con lo que se remuestreó para adaptarla a los 16 KHz de DB1), es una bien conocida base de datos de referencia para la evaluación de PDA (de Cheveigné y Kawahara, 2002; Sun, 2002; Dikshit et al., 2005). DB1 provee marcas de *pitch* manualmente supervisadas y validadas, mientras que DB2 provee valores de *pitch* con una velocidad de ventaneo de 10 ms. La duración de DB1 ($2,5$ horas) permite una evaluación del PMA mucho más fiable (≈ 900K marcas son comparadas) con respecto a otros trabajos con bases de datos de duración que oscila entre 1 (Harbeck et al., 1995) y $8,5$ (Lin y Jang, 2004) minutos. Además, las prestaciones del PMFA se comparan con el algoritmo RAPT (Talkin, 1995), YIN (de

Cheveigné y Kawahara, 2002) y SHRp (Sun, 2002) como entrada a un PMA y PDA. El sPMA de Goncharoff y Gries (1998) se emplea para obtener la marcas de *pitch* correspondientes al YIN y al SHRp. Finalmente, el margen de F_0 que se consideró en los experimentos fue de $[50, 550]$ Hz (Sun, 2002), ventaneando las marcas de entrada cada 5 ó 10 ms.

Medida	**Neutro**		**Alegre**		**Sensual**	
Ventana	**5 ms**	**10 ms**	**5 ms**	**10 ms**	**5 ms**	**10 ms**
RAPT	$10,91$		$10,88$		$31,07$	
RAPT + PMFAs13	*11*, *01*	*22*, *33*	*16*, *20*	*28*, *63*	$24,05$	*33*, *27*
RAPT + PMFAs24	$7,28$	$10,51$	$8,88$	*15*, *37*	$21,06$	$23,48$
RAPT + PMFAs34	$6,61$	$7,98$	$7,61$	$10,42$	$20,64$	$21,65$
RAPT + PMFAs37	**6,08**	$8,01$	**7,54**	$10,83$	$19,93$	$21,25$
RAPT + PMFAs48	$6,59$	$7,43$	$7,88$	$9,17$	$20,21$	$20,79$
RAPT + PMFAs68	$6,21$	$6,76$	$7,76$	**8,22**	**19,90**	**20,12**
RAPT + PMFAs79	$6,19$	**6,74**	$7,87$	$8,24$	$20,02$	$20,20$
RAPT + PMFAs912	$6,38$	$6,84$	$8,59$	$8,89$	$20,46$	$20,35$
YIN	$22,35$		$17,44$		$36,86$	
YIN + PMFAs13	$12,10$	*24*, *37*	$16,82$	*28*, *59*	$23,18$	$32,79$
YIN + PMFAs24	$7,73$	$11,42$	$9,56$	$15,82$	$20,41$	$22,58$
YIN + PMFAs34	$7,14$	$8,39$	**8,43**	$11,16$	$20,26$	$21,03$
YIN + PMFAs37	**6,98**	$8,60$	$8,61$	$11,50$	$20,10$	$20,70$
YIN + PMFAs48	$7,06$	$7,70$	$8,68$	$10,06$	**20,09**	$20,23$
YIN + PMFAs68	$7,14$	**7,15**	$8,75$	**8,99**	$20,16$	**20,03**
YIN + PMFAs79	$7,13$	$7,19$	$8,87$	$9,06$	$20,35$	**20,03**
YIN + PMFAs912	$7,44$	$7,55$	$9,60$	$9,73$	$20,76$	$20,54$
SHRp	$25,16$		$22,45$		$38,85$	
SHRp + PMFAs13	$12,94$	$24,98$	$18,28$	*31*, *87*	$25,11$	$35,24$
SHRp + PMFAs24	$8,64$	$12,02$	$9,49$	$16,87$	**23,09**	$24,47$
SHRp + PMFAs34	**8,02**	$9,09$	**8,28**	$11,30$	$23,11$	$22,97$
SHRp + PMFAs37	$8,15$	$9,17$	$8,87$	$11,71$	$23,82$	$23,11$
SHRp + PMFAs48	**8,02**	$8,20$	$8,67$	$9,62$	$23,30$	$22,93$
SHRp + PMFAs68	$8,16$	**7,83**	$8,85$	**8,58**	$23,69$	**22,77**
SHRp + PMFAs79	$8,32$	$7,86$	$9,21$	$8,73$	$24,04$	$22,89$
SHRp + PMFAs912	$8,91$	$8,35$	$10,37$	$9,73$	$24,91$	$23,76$

Tabla D.1: *GER (%) sobre DB1, donde sXY indica la configuración S_{max} para la primera y segunda pasada del algoritmo de programación dinámica. En cursiva se indican los resultados peores que los de referencia y en* ***negrita*** *los mejores por barrido*

El primer experimento está orientado a analizar el comportamiento del PMFA sobre DB1 con respecto a diferentes configuraciones del algoritmo (S_{max} y ventaneo, según la frecuencia de muestreo f_s) y estilos de habla expresivos. Teniendo en cuenta los trabajos de Goncharoff y Gries (1998) y Alías y Iriondo (2001a), el valor de S_{max} para la segunda pasada del algoritmo de programación dinámica debería de ser mayor que el valor usado en la primera. En las Tablas D.1 y D.2 se resumen los resultados

obtenidos por PMFA, usando las medidas GER y GPMER respectivamente, sobre DB1 para un barrido de S_{max} para dos configuraciones de ventanas. Se observa, tanto para GER como para GPMER, que los resultados superan los valores de referencia a pesar de que la entrada sea PDA o PMA, exceptuando las configuraciones extremas de S_{max}, tales como $s13$, y a pesar del estilo expresivo utilizado. Un examen más detallado de la tabla de resultados muestra que la configuración de análisis de 5 ms obtiene mejores resultados ($s34$ y $s37$ como mejores pares S_{max}) que de 10 ms ($s68$ como mejor par de S_{max}). En términos de expresividad, los mayores errores aplicando PMFA se obtienen para el estilo sensual, debido principalmente a la presencia de voz susurrante, seguido por el estilo alegre, como resultado de su elevado valor de F_0 medio y desviación y, finalmente, el menor de los errores se consigue para el estilo neutro. Nótese que aunque los mejores resultados se obtengan cuando se combina RAPT + PMFA, también se consiguen mejoras notables con YIN y SHRp, como por ejemplo algunos de los mejores resultados para GER y GPMER se dan cuando PMFA se utiliza sobre estos algoritmos.

El segundo experimento está orientado a validar los resultados obtenidos en el primero sobre la base de datos Keele, después de haberla remuestreado a 16 KHz para poder ser comparada con DB1. Los valores de *pitch* incluidos en DB2 tienen un umbral inferior del que no pasarán, 70 Hz para hombres y 120 Hz para mujeres, para evitar la presencia de valores de referencia incorrectos (Kasi y Zahorian, 2002). Sin embargo, como contrapartida al trabajo de Kasi y Zahorian (2002) estas ventanas no son excluidas de la comparación. Si el valor evaluado p, correspondiente a esas ventanas, está fuera de ese umbral se considerará como un error, de ahí que la Tabla D.3 pudiera diferir respecto los resultados obtenidos en trabajos previos, que computan estas ventanas en términos de errores de sonoridad (Sun, 2002). La configuración PMFAs34 con una ventana de 5 ms ha sido aplicada a la referencia de PMA y PDA. Como resultado, se ha obtenido una reducción relativa media de un $75\,\%$ y un $57\,\%$ de GER para locutoras (F) y locutores (M) respectivamente. Por tanto, los resultados de referencia son drásticamente mejorados por el PMFA. Al igual que en el anterior experimento, la combinación RAPT + PMFA es la mejor ($87\,\%$ y $65\,\%$ de mejora relativa en mujeres y hombres), a pesar de que con YIN y SHRp + PMFA también se consiguen resultados positivos.

Los experimentos llevados a cabo han mostrado que el uso de PMFA ha mejorado las prestaciones de cualquiera de los algoritmos de entrada utilizados, para prácticamente cualquier configuración y expresividad. Sin embargo, el ventaneo de 5 ms más la configuración del algoritmo de programación dinámica s34 han obtenido los mejores resultados cuando RAPT fue utilizado como PMA de entrada ($f_s = 16$ KHz). Además, s68 ha sido el mejor S_{max} para un ventaneo de 10 ms, debido a la mitad de la velocidad de ventaneo.

Respecto a la medida de evaluación, la nueva medida presentada GPMER, basada en GER, muestra que los valores absolutos más bajos de GPMER respecto a GER son debidos al elevado número de comparaciones (cada dos marcas de *pitch* en comparación a ventanas). GPMER ha sido definido como un ajuste fino de GER, ya que

Medida	Neutro		Alegre		Sensual	
Ventana	**5 ms**	**10 ms**	**5 ms**	**10 ms**	**5 ms**	**10 ms**
RAPT	7,86		10,37		29,26	
RAPT + PMFAs13	6,86	*20,67*	*15,47*	*34,00*	13,76	24,58
RAPT + PMFAs24	2,86	7,08	6,04	*15,70*	10,93	14,28
RAPT + PMFAs34	2,30	3,72	**4,88**	8,64	10,37	12,45
RAPT + PMFAs37	**2,19**	4,06	5,39	9,55	**9,66**	12,99
RAPT + PMFAs48	2,43	3,24	5,89	7,56	10,28	12,55
RAPT + PMFAs68	2,28	**2,62**	5,88	**6,59**	9,81	**12,09**
RAPT + PMFAs79	2,32	2,64	6,54	7,09	9,83	12,16
RAPT + PMFAs912	2,61	2,95	8,95	9,41	10,18	12,57
YIN	5,16		8,06		22,06	
YIN + PMFAs13	*8,09*	*22,38*	*16,73*	*34,88*	13,97	*24,00*
YIN + PMFAs24	3,44	*8,03*	7,28	*17,07*	11,70	13,61
YIN + PMFAs34	**2,87**	4,41	**6,21**	*10,04*	**11,36**	**12,05**
YIN + PMFAs37	3,07	4,83	7,61	*11,07*	12,33	12,50
YIN + PMFAs48	3,14	3,75	7,80	*9,25*	12,28	12,13
YIN + PMFAs68	3,20	**3,14**	8,04	***8,42***	12,49	12,14
YIN + PMFAs79	3,30	3,23	*8,71*	*8,95*	12,72	12,33
YIN + PMFAs912	3,75	3,80	*10,85*	*11,21*	13,44	13,09
SHRp	8,86		12,15		25,55	
SHRp + PMFAs13	8,28	22,92	*16,72*	*36,28*	15,97	*27,34*
SHRp + PMFAs24	4,00	8,17	5,96	*16,62*	**14,79**	14,48
SHRp + PMFAs34	**3,46**	4,63	**4,69**	8,83	14,98	**14,20**
SHRp + PMFAs37	4,04	5,12	6,68	9,86	16,88	14,92
SHRp + PMFAs48	3,93	4,13	6,48	7,58	16,10	15,07
SHRp + PMFAs68	4,10	**3,71**	6,91	**6,43**	16,74	15,14
SHRp + PMFAs79	4,33	3,80	7,91	7,13	17,35	15,44
SHRp + PMFAs912	5,09	4,46	10,76	9,74	18,47	16,78

Tabla D.2: *GPMER (%) sobre DB1, donde sXY indica la configuración S_{max} para la primera y segunda pasada del algoritmo de programación dinámica. En cursiva se indican aquellos resultados peores que los de referencia y en* ***negrita*** *los mejores por barrido*

Método	F1	F2	F3	F4	F5	Media
RAPT	6,62	4,29	5,44	7,68	2,01	5,21
RAPT + PMFAs34	0,61	0,43	0,20	0,93	0,44	0,52
YIN	3,72	1,07	1,88	4,21	0,38	2,25
YIN + PMFAs34	1,69	0,54	0,47	1,40	0,27	0,87
SHRp	10,85	6,53	10,56	20,71	8,15	11,36
SHRp + PMFAs34	0,88	0,86	0,95	4,38	1,14	1,64
Método	**M1**	**M2**	**M3**	**M4**	**M5**	**Media**
RAPT	22,93	17,42	4,72	14,29	8,33	13,54
RAPT + PMFAs34	12,28	5,15	0,89	2,47	3,10	4,78
YIN	12,02	17,47	1,85	7,62	6,89	9,17
YIN + PMFAs34	11,72	4,48	0,89	2,60	6,19	5,18
SHRp	29,30	21,29	16,91	24,97	25,37	23,57
SHRp + PMFAs34	13,97	7,65	2,33	7,17	12,67	8,75

Tabla D.3: *GER (%) para los locutores femeninos (F1 a F5) y masculinos (M1 a M5) del corpus Keele con PMFAs34 y ventana de 5 ms*

GPMER evita ventanas erróneas ausentes, por ejemplo conteniendo segmentos sobremarcados más un segmento inframarcado (una ventana donde existe una transición de sonoridad) que consigue un valor de periodicidad media próximo al de referencia. Además, el patrón logrado por GPMER, sobre DB1 para los algoritmos de referencia (Neutro<Alegre<Sensual), está mejor correlacionado con los actuales resultados que GER, según los diferentes niveles de dificultad.

D.4. Herramienta para el análisis del etiquetado

Con el objetivo de posibilitar el análisis del etiquetado realizado a partir de los puntos anteriores, se implementó una interfaz de usuario para el análisis del corpus bajo estudio.

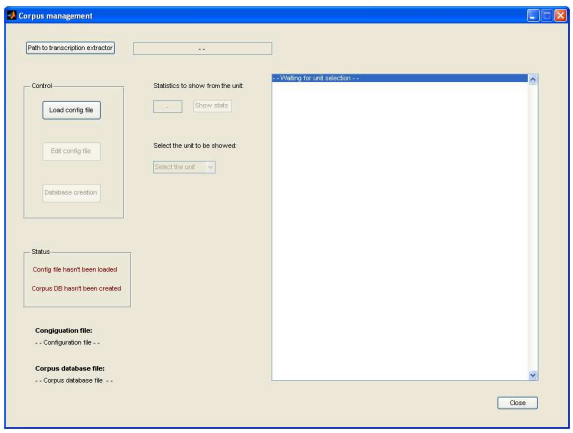


Figura D.4: *Pantalla principal de la herramienta de análisis de corpus*

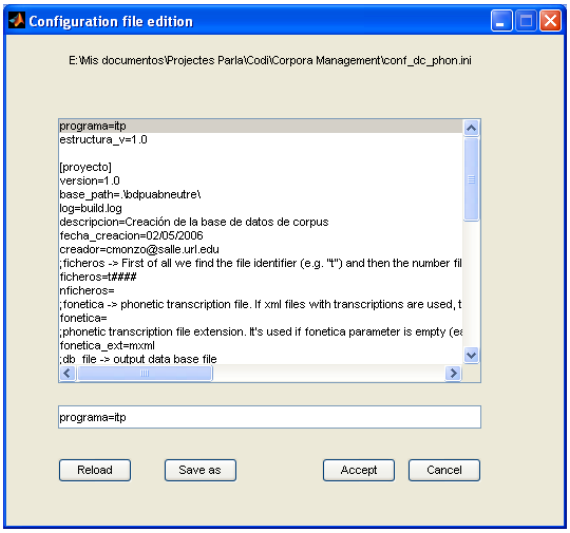


Figura D.5: *Edición del archivo de configuración del analizador del etiquetado*

En la Figura D.4 se muestra la pantalla principal desde la cual se accederá a las distintas opciones de las que se dispone. Para iniciar el análisis se necesita el archivo de configuración que contendrá la información necesaria para comenzar a trabajar, que puede ser editado directamente desde la interfaz (Figura D.5), y posteriormente se precisa de la creación de una base de datos donde se tendrá el corpus de voz preparado para su análisis, de modo que una vez creada se podrá cargar en la interfaz las veces que se desee (Figura D.6).

El software, desarrollado en Matlab, está preparado para soportar todos los corpus que se han ido realizando en castellano, catalán e inglés en diferentes proyectos en los que ha trabajado el grupo de investigación, soportando nuevas funcionalidades como la utilización de archivos XML para contener las transcripciones fonéticas.

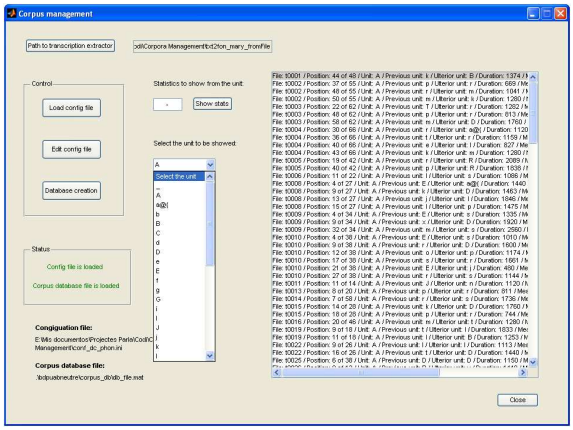


Figura D.6: *Pantalla de carga de la base de datos para el análisis del corpus*

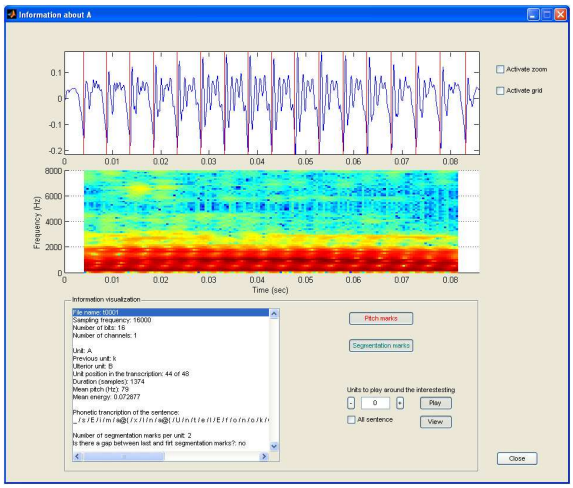


Figura D.7: *Ventana de análisis de una unidad del corpus*

El análisis se realiza a partir de la elección de las unidades que se desean estudiar y a partir de aquí se da la posibilidad de acceder a la frase a la cual pertenece dicha unidad, se muestran todos los datos referentes a ella, se pueden visualizar tanto las marcas de *pitch* como las relativas a la segmentación, se permite la realización de análisis espectrales y por último se da la posibilidad al usuario de escuchar la unidad bajo estudio de manera aislada o bien junto con el número de unidades vecinas deseadas (Figura D.7).

APÉNDICE E

Desambiguación en la transcripción fonética de acrónimos

Este apéndice muestra el trabajo desarrollado en la transcripción fonética de acrónimos, de forma que estos puedan ser pronunciados correctamente por un sistema de CTH.

E.1. Planteamiento del problema

A continuación se aborda el problema de encontrar acrónimos en la entrada del sistema de Conversión de Texto en Habla (CTH), de modo que deben ser procesados para su correcta pronunciación. Esta desambiguación se lleva a cabo dentro del bloque de Procesamiento del Lenguaje Natural (PLN).

En todas las áreas del lenguaje y de las tecnologías del habla se trata, en mayor o menor grado, con texto real. En cualquier caso, el texto que se puede encontrar contiene elementos no deseados, es decir, elementos como por ejemplo números, abreviaturas, acrónimos o fechas. Estos son Palabra no Estándar —*Non-Standard Word*— (NSW), con que sus propiedades no se corresponden con las que tendría el resto del texto, incluyendo su pronunciación si se utiliza en una aplicación de CTH. Es por esta razón que las NSW típicamente son normalizadas por el PLN, para que su comportamiento sea lo más parecido al de una palabra estándar, tarea que no resulta fácil por depender en muchas ocasiones del contexto y tipo de texto.

En el caso concreto de los sistemas de CTH, aparece el problema de convertir estas palabras en habla. Esto es particularmente importante para alcanzar el objetivo de acceso universal, desde texto en periódicos convencionales a texto incluido en web o correos electrónicos, de manera que la calidad del habla generada se vea incrementada en comparación al no tratamiento de estos casos.

Para afrontar este problema se desarrolló el trabajo aquí presentado, dentro del ámbito de tratamiento de NSW orientado a la mejora de la calidad del habla generada por parte de un sistema de CTH, resultando en una publicación en revista nacional

(Monzo et al., 2006). El trabajo realizado forma parte del PLN en cuanto se trata de crear la transcripción fonética más correcta de acrónimos para que sea leída adecuadamente en castellano. A pesar de que la metodología podría ser extrapolable a otras lenguas, no se ha constatado con experimentos. Dentro de la problemática descrita sobre NSW, el foco del trabajo se centró únicamente en el de tratar con acrónimos, detectados previamente en el texto de entrada al sistema de CTH. Por tanto, el objetivo que se persigue es el de disponer de un sistema automático de transcripción de acrónimos que realice la normalización de los textos dentro del bloque de PLN, en lugar de un diccionario de excepciones, táctica basada en disponer de una lista de palabras conocidas cuya transcripción fonética sea alterada por la indicada. El hecho de tener un diccionario de excepciones es justificable en el caso de tener una aplicación de voz ad hoc o de dominio restringido, cosa que no siempre será así, necesitándose por tanto de una herramienta que facilite la tarea de reutilizar el diseño del sistema de CTH de un dominio a otro totalmente diferente.

Antes de continuar pasemos a ver la definición que hace la Real Academia Española[1] de la lengua sobre el término acrónimo:

> *"Vocablo formado por la unión de elementos de dos o más palabras, constituido por el principio de la primera y el final de la última, p. ej., ofi(cina infor)mática, o, frecuentemente, por otras combinaciones, p. ej., so(und) n(avigation) a(nd) r(anging), Ban(co) es(pañol) (de) (crédi)to".*

En nuestro caso se considera en un sentido más amplio, donde se pueden encontrar aquellos que se leen como una palabra más de la lengua en su totalidad, parcialmente o directamente hay que deletrear cada una de las siglas utilizadas en su construcción.

La transcripción fonética se encarga de la conversión de un texto (grafemas) a sus fonemas correspondientes. Esta transcripción se realiza habitualmente mediante la aplicación de reglas (p. ej. castellano) o el uso de diccionarios (p. ej. inglés). Uno de los elementos clave del transcriptor fonético es el proceso que realiza la normalización del texto. Este proceso es el encargado de tratar los números, abreviaturas y acrónimos para obtener la transcripción fonética correspondiente a sus equivalentes escritos o extendidos (p. ej. 12 → doce o Sr. → señor). La normalización se puede realizar a partir de diccionarios de excepciones, donde se guarda la conversión de grafema a fonema, con el problema de que por su poca flexibilidad, para cualquier cambio en el corpus se debe realizar una comprobación manual. Por contra, automatizando del proceso de normalización, el tratamiento de las excepciones se realizaría de modo transparente al resto del sistema de CTH. El proceso de desambiguación de qué es y qué no es un acrónimo se supone previo a este trabajo, pudiéndose realizar mediante:

i. Etiquetas XML, o equivalentes, cuando se conocen a priori los textos a sintetizar (Alías et al., 2005).

[1] http://www.rae.es/rae.html

ii. Un sistema automático de desambiguación de textos (Mikheev, 2002).

La información base que se usó en la creación del sistema de transcripción de acrónimos fueron los distintos grafemas que los forman y sus relaciones, tal y como se desprende del trabajo presentado por Sejnowski y Rosenberg (1987). Por lo que hace referencia a la decisión de cómo se debe de pronunciar cada uno de los acrónimos, se plantea la opción de utilizar árboles de decisión, de manera que se asocie a cada grafema de entrada una cierta pronunciación ("clasificación") siguiendo el trabajo de Mikheev (2002), en el que se realiza desambiguación de abreviaturas mediante esta técnica de aprendizaje.

Debido al gran volumen de información con el que trabajar, es decir, los diferentes grafemas de la lengua de interés y las combinaciones que se pueden dar entre ellos, a la hora de crear los distintos acrónimos, se planteó la necesidad de codificar la información para, de este modo, ayudar a compactarla, aumentando así la eficacia de la herramienta de clasificación utilizada. En concreto, este trabajo se basa en la codificación Soundex (NARA, 1995), que es un sistema de indexación usado para codificar apellidos dentro del sistema de censo norteamericano, de manera que en lugar de estar basado en cómo se deletrea el apellido se basa en cómo suena. En la Tabla E.1 se presenta la tabla de codificación Soundex, donde se puede observar que los grafemas A, E, I, O, U, H, W e Y son obviados. Para nuestros objetivos, dicha codificación fue alterada y testada tal y como se muestra más adelante.

Código	Grafemas representados
1	B, F, P, V
2	C, G, J, K ,Q, S, X, Z
3	D, T
4	L
5	M, N
6	R

Tabla E.1: *Tabla de codificación Soundex (NARA, 1995)*

E.2. Descripción del sistema

A continuación pasemos a ver la descripción del sistema de transcripción automática de acrónimos propuesto. Se muestran los diferentes planteamientos realizados para abordar el problema, se presenta el algoritmo utilizado y los resultados obtenidos a través de las distintas configuraciones propuestas.

El problema de los acrónimos radica en su lectura debido a que no siguen unas reglas predefinidas, por ser en cada caso, la combinación de grafemas quien las fijará. Las diversas opciones que se han detectado para su posible lectura son las siguientes:

1. **Normal**: como una palabra más de la lengua (p. ej. ONU).
2. **Deletreo**: deletreando la palabra (p. ej. FBI).

3. **Híbrida**: lectura híbrida donde se combina la normal más el deletreo (p. ej. PSOE).

Es a partir de técnicas de ML de donde se extrae el conocimiento necesario para realizar la generalización de los casos de entrenamiento. Entre las distintas técnicas que pueden ser utilizadas se optó por el uso de árboles de decisión y, entre las distintas aproximaciones para su modelado, se decidió usar el algoritmo C4.5 (Quinlan, 1993), principalmente por la simplicidad que presenta su funcionamiento frente a otros. Para su aplicación se utilizó la herramienta informática Weka (Witten y Frank, 2005), donde se pueden encontrar multitud de algoritmos de ML, facilitando el proceso de experimentación. El algoritmo C4.5 (J4.8 en Weka), se trata de un árbol de decisión que acepta tanto datos numéricos como nominales. Permite realizar poda de sus ramas consiguiendo así variar el grado de generalización de la solución obtenida, correspondiéndose una mayor poda con una mayor generalización.

En cuanto a los datos utilizados para el entrenamiento y test del sistema, se recopilaron acrónimos de manera que cubrieran un amplio espectro de dominios para dar así mayor variabilidad a los datos de entrenamiento. Los dominios considerados de mayor interés, por el volumen de acrónimos que se suelen encontrar son: "Tecnología" (informática e internet) y "Periodismo" (política y economía), añadiéndose a ellos una serie de "Varios", que dan cobertura a otros ámbitos como son la educación o la medicina. En la Tabla E.2 se presentan los diferentes dominios y la cobertura ofrecida, es decir, el porcentaje de pertenencia de los acrónimos a cada uno de ellos.

Dominio	**Número de acrónimos**	**Cobertura**
Tecnología	1862	54 %
Periodismo	1432	41 %
Varios	174	5 %
Total	3468	100 %

Tabla E.2: *Dominio de aplicación y cobertura de los acrónimos usados*

Tras la recopilación de acrónimos se preparó la información de forma que el algoritmo de ML la pudiera tratar. Para ello, se seleccionó un subconjunto de casos entre todos los posibles y se crearon las correspondencias entre acrónimo y transcripción, grafema a grafema mediante la participación de un experto, de forma que dicha información fuera la utilizada para el entrenamiento y posterior test del sistema de transcripción automática. Debido que a priori se desconocía la correspondencia entre los grafemas y su transcripción fonética (p.ej. P → /pE/ o /p/) el procedimiento a seguir fue el de asegurar un número mínimo de apariciones de todos los grafemas de la lengua de interés dentro de los acrónimos, de igual forma que también se tuvo en cuenta el tamaño (número de grafemas que lo forman) de los mismos. La razón de por qué se tuvo en cuenta su tamaño fue debido a que no es lo mismo uno considerado largo (p. ej. 6 grafemas), que tiene más posibilidades de que ciertas partes se lean de forma normal o híbrida, que uno corto (p. ej. 3 grafemas) donde lo más probable es que deba ser deletreado.

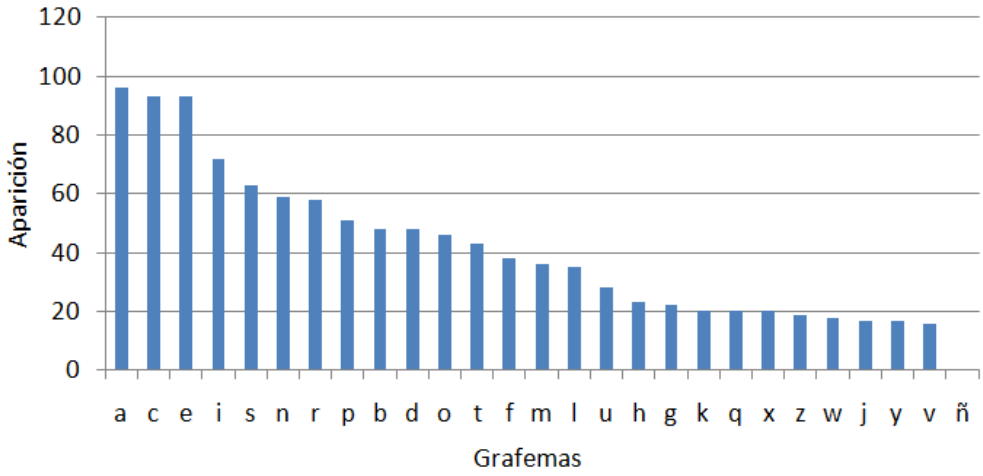


Figura E.1: *Aparición de grafemas*

Para la elección de aquellos acrónimos que serían usados en los experimentos se utilizó un algoritmo *greedy* (François y Boëffard, 2002) sobre el conjunto de datos disponible. Este algoritmo selecciona los acrónimos que cumplen con una serie de requisitos de entrada, siendo éstos el número de apariciones por grafema y su tamaño (criterio de corto y largo). Al algoritmo se le pasaron los 27 grafemas del castellano, es decir el listado con los grafemas de interés, juntamente con el número de apariciones mínimo que se deseaba que hubiera de cada uno de ellos (ajustado empíricamente a 10 apariciones por grafema en acrónimos cortos y 10 más en largos). Finalmente, después de eliminar posibles redundancias, debidas a la selección usando los dos criterios de corto y largo, se obtuvieron 281 acrónimos con un total de 1099 grafemas (con la distribución mostrada en la Figura E.1), con los que se realizó el entrenamiento y test del sistema de ML.

Acrónimo	Graf. interés	Graf. anterior	Graf. posterior	Trasncripción
ZCS	Z	#	C	/TEta/
	C	Z	S	/TE/
	S	C	#	/Ese/

Tabla E.3: *Representación del acrónimo "ZCS" usando una ventana de tamaño igual a 3*

Una vez seleccionados los acrónimos un experto realizó las transcripciones, obteniéndose 48 posibles clases de salida diferentes, es decir transcripciones distintas. Cada grafema que formaba el acrónimo disponía de una representación en el espacio de transcripciones, por tanto el número de grafemas coincidía con el de símbolos representando la transcripción (facilitando el proceso de pasar la información al algoritmo de ML). Las transcripciones utilizadas estaban basadas en SAMPA (Wells, 1997), modificada para representar la tonicidad de las vocales (mayúscula si son tónicas). Si por cualquier razón alguno de sus grafemas no disponía de transcripción se le

asoció el símbolo /-/ (p. ej. HI-FI → /-/ /I/ /-/ /f/ /i/). Si el acrónimo incorporaba símbolos distintos a los grafemas (como "/" o "-") se les asoció /-/. Los datos se pasaron al algoritmo de ML mediante un ventaneo del acrónimo, de forma que el grafema de interés quedaba posicionado en el centro de la ventana, con lo que su tamaño fue un parámetro a estudiar. Mediante el símbolo "#" se indicó que el grafema de interés estaba en un extremo del acrónimo y, por tanto, al ventanear y ser centrado en esa ventana no se dispone de información de los extremos. Este proceso de representación de la información se ejemplifica en la Tabla E.3.

(a) *Versión inicial*

Código	Grafema
UNO	B F P V
DOS	C G J K Q S X Z
TRES	D T
CUAT	L
CINC	M N Ñ
SEIS	R
SIET	H
OCHO	A E I O U W Y
NUEV	/ - #

(b) *Versión final*

Código	Grafema
UNO	B P V D T K
DOS	C J S Z L R
TRES	X Q
CUAT	F G
CINC	M N Ñ
SEIS	H
SIET	A E I O U
OCHO	W
NUEV	Y
DIEZ	/ - #

Tabla E.4: *Versión inicial y final de Soundesp*

A pesar de tener representados cada uno de los grafemas de los diferentes acrónimos, surgió un problema debido a la excesiva variabilidad de los datos de entrada al árbol de decisión. Al trabajar con todas las combinaciones que pueden darse en los acrónimos de entrenamiento, la dimensión del problema era demasiado elevada y podía provocar que los resultados de clasificación no fueran satisfactorios. Así pues, el razonamiento fue el siguiente, si el problema radica en la elevada casuística existente, se planteó la disminución de la dimensión del problema aplicando una codificación sobre los grafemas anteriores y posteriores al de interés. Con este fin se utilizó la codificación Soundex, la cual se adaptó al castellano a partir del trabajo presentado por García (2002), realizando ajustes finos sobre esta, pasándose a denominar Soundesp. En la Tabla E.4a se muestra la versión inicial, más próxima a Soundex, y en la Tabla E.4b la versión final, donde la agrupación se realiza mediante un ajuste más fino de los sonidos con comportamiento similar, para así mejorar el rendimiento del sistema. A modo de ejemplo, en la Tabla E.5 se presenta el caso presentado en la Tabla E.3, usando la versión final de la codificación Soundesp.

Acrónimo	Graf. interés	Cód. graf. ant.	Cód. graf. post.	Trasncripción
ZCS	Z	DIEZ	DOS	/TEta/
	C	DOS	DOS	/TE/
	S	DOS	DIEZ	/Ese/

Tabla E.5: *Ejemplo de aplicación de la versión final de Soundesp*

E.3. Resultados

Antes de pasar a presentar los resultados obtenidos veamos las distintas configuraciones con las que se ha trabajado:

- Barrido del tamaño de la ventana: 3, 5 y 7 grafemas.
- Variación en la representación de los datos, sin utilizar ningún tipo de codificación sobre los grafemas anterior y posterior al de interés y utilizando la versión inicial y final de Soundesp.
- Uso de la información de tonicidad de las vocales.

Se utilizó el algoritmo C4.5 con poda del árbol y $10-$*fold cross validation*. Una vez realizas las pruebas sobre las distintas configuraciones y, por tanto, conocida la configuración óptima, esta fue validada a partir de la aplicación de los siguientes algoritmos de ML:

- C4.5 con y sin poda del árbol.
- IBk (Aha et al., 1991).
- NaiveBayes (John y Langley, 1995).

Configuración	Transcripción 1 (%)	Transcripción 2 (%)
Grafemas 3	$78,82$	$77,92$
Soundesp ini - 3	$84,38$	$83,21$
Soundesp final - 3	$84,29$	$83,84$
Grafemas 3 acentos	$67,15$	$66,25$
Soundesp ini - 3 acentos	$72,08$	$72,44$
Soundesp final - 3 acentos	$72,44$	$72,53$
Grafemas 5	$78,28$	$77,83$
Soundesp ini - 5	$82,5$	$82,56$
Soundesp final - 5	$83,75$	$83,12$
Grafemas 5 acentos	$67,15$	$66,16$
Soundesp ini - 5 acentos	$70,65$	$71,10$
Soundesp final - 5 acentos	$71,81$	$72,35$
Grafemas 7	$78,64$	$78,01$
Soundesp ini - 7	$81,68$	$81,43$
Soundesp final - 7	$82,41$	$82,23$
Grafemas 7 acentos	$66,97$	$65,98$
Soundesp ini - 7 acentos	$70,65$	$71,63$
Soundesp final - 7 acentos	$72,80$	$73,25$

Tabla E.6: *Porcentaje de instancias correctamente clasificadas para diferentes configuraciones y transcripciones*

La medida empleada para decidir la mejor configuración fue la del tanto por ciento de instancias correctamente clasificadas, es decir las transcripciones que el

sistema ha realizado correctamente. Además del porcentaje de clasificación correcta, otras medidas derivadas de la aplicación del árbol de decisión son: la precisión (P), la cobertura (C) y la $F1$ (Sebastiani, 2002) (Apéndice F.7).

Uno de los parámetros que se testaron fue el tamaño de ventana, de los que se eligieron tres configuraciones distintas: 3, 5 y 7; otro fue la codificación utilizada: versión inicial y final; y por último se tuvo, o no, en cuenta la tonicidad de las vocales, de manera que se relajaron las exigencias del transcriptor. La sintaxis que se siguió en cada uno de los casos fue "Grafemas" si no se usaba codificación Soundesp y "Soundesp" en caso contrario, con "ini" o "final" según fuera el caso de versión inicial o final, "tamaño de ventana" y "acentos" en el caso de tenerse en cuenta la tonicidad de las vocales.

En la Tabla E.6 se pueden observar dos columnas de resultados, en la primera se muestran los resultados suponiendo el acrónimo a principio de frase ("Transcripción 1"), mientras que en la segunda se considera el acrónimo en mitad de la misma ("Transcripción 2"), de forma que se suavizaron los requisitos por encontrar para el mismo grafema una sola posible salida independientemente de su posición en el acrónimo. Esta diferenciación hace que algunos fonemas, como por ejemplo, los oclusivos tengan diferente transcripción dependiendo de su posición en la frase.

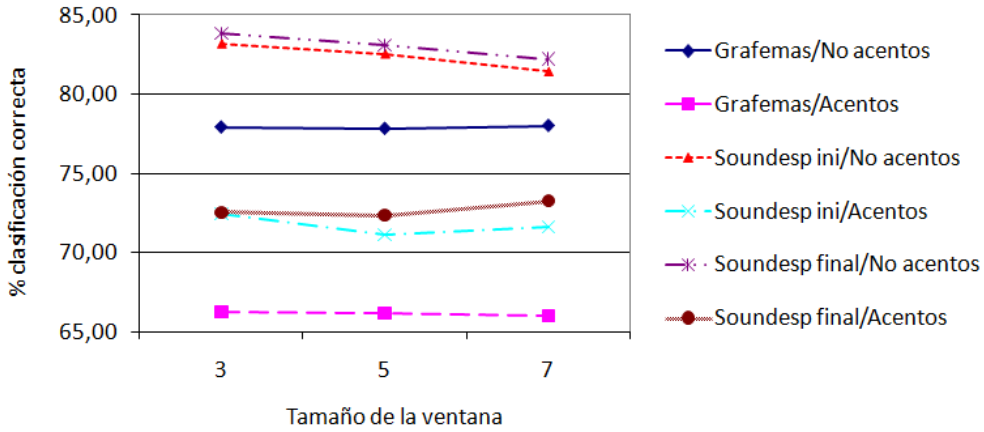


Figura E.2: *Porcentaje de instancias correctamente clasificadas en función de la configuración de entrada*

De los resultados presentados se desprende que no existen diferencias significativas entre el uso de la "Transcripción 1" y "Transcripción 2", por tanto fue la "Transcripción 2" la configuración que se utilizó para la realización del resto de experimentos, por ser una configuración más acorde con la realidad, puesto que los acrónimos, típicamente, los encontraremos dentro del texto en cualquier posición y no solamente al inicio de frase. Así pues, el siguiente paso fue estudiar cuál era la mejor configuración de entrada al algoritmo (Figura E.2). Se observa que entre las diferentes codificaciones, "Soundesp final" proporciona una mejora sobre el porcentaje de clasificación en torno a un $6\,\%$ respecto a no usar codificación. De este análisis se concluyó que el tamaño de las ventanas no es un parámetro determinante por no

existir grandes diferencias en los resultados obtenidos, pero que el hecho de usar la información de tonicidad de las vocales se tendrá que tener en cuenta según sean las necesidades concretas de la aplicación.

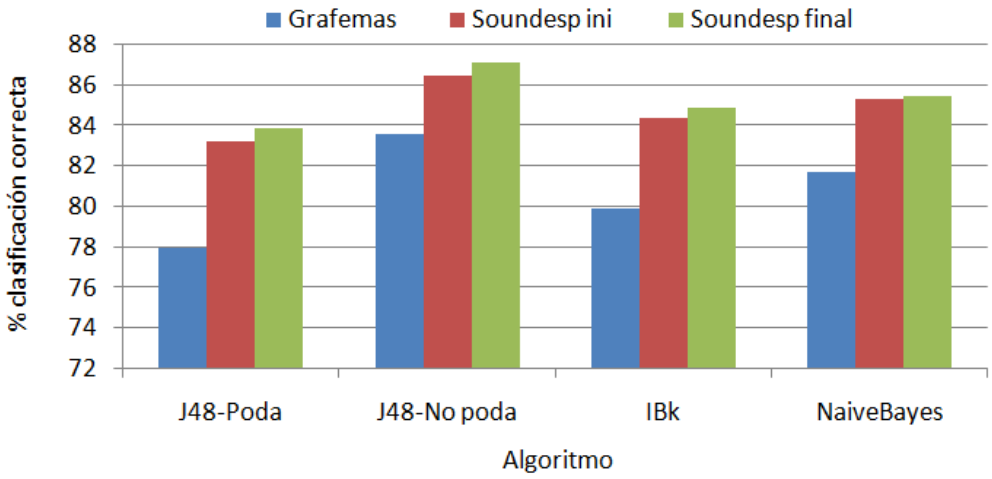


Figura E.3: *Porcentaje de clasificación correcta para distintos algoritmos de aprendizaje*

Por último se planteó la validación de los resultados para las distintas configuraciones mediante un barrido de algoritmos: C4.5 con y sin poda del árbol, IBk y NaiveBayes, utilizando la medida de test $10-$*fold Cross-Validation*. Para ello se consideró únicamente el caso de tamaño de ventana igual a 3, la no utilización de la información de tonicidad y las tres posibilidades de tratamiento de la información de entrada (sin codificación y usando Soundesp en sus dos configuraciones). Los resultados se muestran en la Figura E.3, pudiéndose observar como el uso de Soundesp para ambas versiones (inicial y final) aumentó el rendimiento del sistema para cualquiera de los métodos de aprendizaje. Asimismo, el hecho de no aplicar poda en C4.5 mejora los resultados, a pesar de que esto implicará una reducción de la generalización del algoritmo.

Para terminar se resume la configuración óptima, consiguiéndose un porcentaje de clasificación correcta próxima al $85\,\%$:

- Tamaño de ventana igual a 3.
- Codificación Soundesp versión final.
- Transcripción sin tener en cuenta la tonicidad de las vocales.

APÉNDICE F

Herramientas estadísticas

En este apéndice se introduce al lector a los conceptos y las herramientas estadísticas utilizados a lo largo de la presente tesis. Pretende ser una guía rápida de consulta para facilitar la comprensión de las explicaciones, análisis y conclusiones extraídas durante la lectura de la misma.

F.1. Modelo Autorregresivo (AR)

La idea básica que está detrás del modelo AR es, que para ciertos tipos de procesos, el valor actual de la variable puede ser expresado como una combinación de un número finito de valores del pasado. El número de valores que serán necesarios dependerá de la naturaleza del proceso.

Cada uno de los valores se aproxima por tanto a una combinación lineal (recta de regresión) de P muestras pasadas, donde $v(n)$ representa el error de dicha aproximación (Ecuación F.1).

$$u(n) = -\sum_{k=1}^{P} a_k + u(n-k) + v(n) \tag{F.1}$$

Con las siguientes consideraciones:

- Siendo los parámetros del modelo: $a_1, a_2, \cdots, a_P, \sigma_v^2$.
- La función de transferencia del modelo es la presentada en la Ecuación F.2

$$U(z) = \frac{V(z)}{H_A(z)} \Leftrightarrow V(z) = U(z) \cdot H_A(z),\ H_A(z) = 1 + \sum_{k=1}^{P} a_k \cdot z^{-k} \tag{F.2}$$

La interpretación que se hace a partir de las consideraciones anteriores es la siguiente:

- Análisis/Descorrelación. Dado un proceso AR $u(n)$, puede ser filtrado con el filtro FIR $H_A(z)$ para dar lugar a un proceso blanco $v(n)$, también llamado "innovación del proceso" o "error de predicción" (predicción lineal).

- Síntesis/Correlación. La dinámica propia del proceso $u(n)$ se puede generar filtrando la innovación $v(n)$ con un filtro IIR $\frac{1}{H_A(z)}$.

F.2. Linear Predictive Coding (LPC)

LPC es una herramienta, generalmente utilizada en el procesamiento de la señal de voz, para representar la envolvente espectral de una señal de habla digital de una forma comprimida, usando la información de un modelo predictivo lineal. Esta es una de las técnicas de análisis del habla más potentes y uno de los métodos más útiles para codificar el habla con buena calidad usando una baja velocidad de bit, proveyendo de una precisa estimación de los parámetros del habla.

Los principios básicos son los siguientes:

- LPC empieza con la suposición que la señal de habla se produce por un resonador situado al final de un tubo. La glotis produce el zumbido, que queda caracterizado por su intensidad (sonía) y frecuencia (tono). El tracto vocal forma el tubo, que se caracteriza por sus resonancias (formantes).

- LPC analiza la señal de habla estimando los formantes, eliminando sus efectos (filtrado inverso) y estimando la intensidad y frecuencia de la señal restante (residuo).

- Los valores que describen los formantes y el residuo pueden ser almacenados o transmitidos. LPC sintentiza la señal de habla mediante el proceso inverso: usa el residuo para crear una fuente de señal, los formates para crear un filtro y hace pasar la fuente de señal por este, dando como resultado la señal de habla.

- La señal de habla varía con el tiempo, es por eso que este proceso se realiza en cortos espacios de tiempo llamados *frames*.

El problema principal del LPC es estimar los formantes. La solución básica es una ecuación en diferencias, que expresa cada muestra de la señal como una combinación lineal de las muestras anteriores. A esta ecuación se la conoce como "predictor lineal". Los coeficientes de esta ecuación en diferencias (los "coeficientes de predicción") caracterizan a los formantes y su estimación se realiza minimizando el error cuadrático medio entre la señal predicha y la actual.

F.3. Estadística descriptiva

La estadística descriptiva se refiere a la recolección, presentación, descripción, análisis e interpretación de una colección de datos. Esencialmente consiste en resumirlos con uno o dos elementos de información (medidas descriptivas) que caracterizan la totalidad de los mismos, obteniendo un conjunto de conclusiones sobre si mismos, sin sobrepasar el conocimiento proporcionado por éstos.

F.4. Boxplot

En estadística descriptiva (Apéndice F.3), el *boxplot* (Frigge et al., 1989), *box plot* o diagrama *box-and-whisker* (en castellano conocido como "diagrama de caja" o "diagrama de caja-y-bigote"), es una buena forma de representar gráficamente a grupos de datos numéricos. Esta representación resume la información mediante 5 valores: mediana, primer cuartil, tercer cuartil, valor mímino y valor máximo. Presenta al mismo tiempo información sobre la tendencia central, dispersión y simetría de los datos. Además, permite identificar con claridad y de forma individual, observaciones que se alejan de manera poco usual del resto de los datos, conocidas como valores atípicos (*outiers* en inglés).

Los *boxplots* muestran diferencias entre poblaciones sin hacer ninguna suposición sobre la distribución que siguen los datos (son no-paramétricos). El gráfico consiste en un rectángulo (caja) de cuyos lados superior e inferior se derivan, respectivamente, dos segmentos: uno hacia arriba y uno hacia abajo (bigotes). Los valores que quedan fuera de esta caja más los bigotes serán los valores atípicos. En lo que se refiere a su representación gráfica, pueden ser dibujados horizontal o verticalmente.

F.5. $t-$test

El $t-$test evalúa si las medias de dos grupos son estadísticamente diferentes. Este análisis es apropiado cuando se desea comparar las medias de dos grupos.

La ecuación para el $t-$test (Ecuación F.3) es una proporción. El numerador es la diferencia entre las dos medias, mientras que el denominador es una medida de la variabilidad o dispersión de los datos. Este denominador se calcula como la raíz cuadrada de la suma de las varianzas divididas por el número de datos de cada grupo.

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \tag{F.3}$$

Una vez calculado el valor de t, se busca en una tabla de significación para comprobar si el margen es suficientemente grande como para decir que las diferencias entre grupos no ha sido fortuita, asignando un valor de riesgo (nivel alfa, en muchos trabajos establecido en $0,05$). Si resulta serlo, se puede concluir que la diferencia entre las medias para dos grupos es diferente. Este valor de significación suele ser dado por los programas de análisis estadístico, evitando así tener que buscar en la tabla.

El $t-$test y ANOVA son matemáticamente equivalentes y, por tanto, deberían obtener los mismos resultados.

F.6. Test de Wilcoxon

El test de Wilcoxon es una alternativa no-paramétrica al $t-$test (Apéndice F.5) para el caso de dos muestras relacionadas o medidas repetidas sobre una única muestra.

De igual forma que el $t-$test, el test de Wilcoxon implica comparaciones de diferencias entre medidas, de forma que requiere que los datos sean medidos en un cierto intervalo de medida. Sin embargo, no es necesario hacer ninguna suposición sobre la forma de la distribución de las medidas. Por tanto puede ser utilizado bajo suposiciones de distribución que $t-$test no puede satisfacer.

Empezando con un conjunto de pares de valores de X_a y X_b con un tamaño de muestra n_a y n_b respectivamente, el test de Wilcoxon puede ser calculado del siguiente modo:

1. Se colocan todas las observaciones en una única serie ordenada.
2. Se suman los rangos para las observaciones que provienen de la muestra X_a.
3. Se puede calcular W_a (Ecuación F.4):

$$W_a = R_a - \frac{n_a(n_a+1)}{2} \tag{F.4}$$

4. W_b puede ser obtenido siguiendo los pasos del 1 al 3.
5. Calculados W_a y W_b, el valor más pequeño de los dos resultados es el utilizado para consultar la tabla de significación.

F.7. Matriz de confusión, precisión, cobertura y medida F

Una matriz de confusión (Kohavi y Provost, 1998) es una herramienta de visualización que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases. Si en los datos de entrada el número de muestras de clases diferentes cambia mucho, la tasa de error del algoritmo no es representativa de lo bien que este realiza su tarea.

Para solucionar este problema, se utilizan las medidas de precisión (P) (Ecuación F.5), la cobertura (C) (Ecuación F.6) y la medida F (F_β) (Ecuación F.7) (*precision*, *recall* y F*-measure* en inglés), que son medidas estadísticas utilizadas en la evaluación de algoritmos de recuperación de información. Normalmente, precisión y cobertura no son analizados de forma aislada, sino que se combinan en una única medida que las pondera, dando una visión conjunta de ambas (F_β).

$$P = \frac{VP}{VP + FP} \tag{F.5}$$

$$C = \frac{VP}{VP + FN} \tag{F.6}$$

$$F_\beta = (1 + \beta^2) \cdot \frac{P \cdot C}{\beta^2 \cdot P + C} \tag{F.7}$$

Donde $VP =$"Verdadero Positivo" o clasificación correcta, $FP =$"Falso Positivo" o clasificación incorrecta debido a asociar la salida del clasificador a esa clase cuando esto no es cierto y, por último, $FN =$"Falso Negativo" o clasificación incorrecta debido a asociar la salida a una clase diferente no siendo esto cierto.

Una medida típica de F_β es el caso de $\beta = 1$, dando el mismo peso tanto a la precisión como a la cobertura, siendo conocida como medida $F1$ (Ecuación F.8) (Sebastiani, 2002). Lo interesante es maximizar la medida de $F1$ para conseguir una alta precisión y cobertura en la clasificación.

$$F1 = 2 \cdot \frac{P \cdot C}{P + C} \tag{F.8}$$

F.8. Linear Discriminant Analysis (LDA)

Análisis Discriminante Lineal —*Linear Discriminant Analysis*— (LDA) es un método usado en estadística y en Aprendizaje Automático —*Machine Learning*— (ML) para hallar la combinación lineal de características que mejor separa dos o más clases de objetos o eventos. Esta combinación resultante puede ser usada como un clasificador lineal o para reducir la dimensionalidad antes de la clasificación. Por lo tanto, su propósito principal es el de clasificar objetos (p. ej. tipos de persona en grupos basados en un conjunto de características, como edad o género, que los describen). En general, la asignación de objeto a grupo se realiza en base a observaciones hechas sobre dicho objeto.

LDA está estrechamente ligado con ANOVA y análisis de regresión, intentando expresar una variable dependiente como una combinación lineal de otras características o medidas. En estos dos métodos, la variable dependiente es una cantidad numérica, mientras que para LDA es una variable categórica, es decir una etiqueta.

También está relacionado con Análisis de Componentes Principales —*Principal Component Analysis*— (PCA) y análisis de factor, en los que ambos buscan combinaciones lineales que mejor expliquen los datos. Por un lado LDA trata de modelar explícitamente la diferencia entre clases de datos. Por otro lado, PCA no tiene en cuenta ninguna diferencia de la clase, y el análisis de factor construye las combinaciones de características basadas en diferencias más que en similitudes.

F.9. SMO

SMO implementa el algoritmo de optimización mínima secuencial (Witten y Frank, 2005) para entrenar una Máquina de Soporte Vectorial —*Support Vector Machine*— (SVM) (Vapnik, 1995). Estos algoritmos extienden las características de los modelos lineales, ya que permiten distinguir entre clases que presentan límites de decisión no lineales. Para ello se transforman los datos originales, de forma no lineal, en un nuevo espacio de mayor dimensión. En este nuevo espacio se construye un modelo lineal que pueda representar un límite de decisión no lineal en el espacio original.

F.10. J48

J48 implementa la versión pública del algoritmo de clasificación basada en árboles de decisión C4.5 revisión 8, previa a la comercialización de la versión C5.0 (Witten y Frank, 2005). Estos árboles clasifican un nuevo caso mediante la evaluación, en cada nodo del modelo, de los parámetros que definen el caso que se pretende clasificar. Los casos que, partiendo de la raíz, llegan a una determinada hoja reciben la clasificación que la hoja indica.

F.11. Naïve Bayes

Naïve Bayes (John y Langley, 1995) es un clasificador probabilístico que parte de la premisa de que cada par parámetro-valor de un mismo ejemplo es independiente del resto. A cada par parámetro-valor se le asigna una probabilidad de pertenencia a una clase. Para ello se divide el número de ejemplos de cada clase en los que aparece ese par entre el número de ejemplos que pertenecen a esa clase. Para clasificar un caso nuevo se calcula la probabilidad de pertenencia de ese caso a cada clase, clasificándolo en la clase donde dicha probabilidad sea mayor, adoptando así un criterio de estimación máxima a posteriori. Esta probabilidad de pertenencia se calcula como el producto de la probabilidad de pertenencia a cada clase de cada uno de los pares parámetro-valor que definen el caso que se desea clasificar.

Bibliografía

Abdi, H. (2003). "Least squares". En: M. Lewis-Beck, A. Bryman y T. Futing (Eds.), *Encyclopedia for research methods for the social sciences*, pp. 559–561. Sage, Thousand Oaks (CA). 46, 117, 123

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh University Press, Edinburgh. 34, 35

Aha, D., Kibler, D. y Albert, M. (1991). "Instance-based learning algorithms". *Machine Learning*, **6**, pp. 37–66. 233

Alías, F. y Iriondo, I. (2001a). "Asignación automática de marcas de pitch basada en programación dinámica". En: *Procesamiento del Lenguaje Natural*, volumen 27, pp. 225–231. Jaén, Spain. 100, 221

Alías, F. y Iriondo, I. (2001b). "Segmentador de fonemas en catalán basado en DHMM". En: *XVI Simposium Nacional de la Unión Científica (URSI'2001)*, pp. 149–150. Madrid, Spain. 94

Alías, F. y Iriondo, I. (2002). "La evolución de la Síntesis del Habla en Ingenería la Salle". En: *II Jornadas en Tecnología del Habla*, Granada, Spain. 83, 121, 132

Alías, F., Iriondo, I., Formiga, L., Gonzalvo, X., Monzo, C. y Sevillano, X. (2005). "High quality Spanish restricted-domain TTS oriented to a weather forecast application". En: *The 9th European Conference on Speech Communication and Technology (Interspeech'2005)*, pp. 2573–2576. Lisboa, Portugal. 71, 77, 98, 228

Alías, F., Monzo, C. y Socoró, J. C. (2006). "A Pitch Marks Filtering Algorithm based on Restricted Dynamic Programming". En: *The 9th International Conference on Spoken Language Processing (Interspeech'2006)*, pp. 1698–1701. Pittsburgh, USA. 100, 220

Alku, P. (1992). "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering". *Speech Communication*, **11**, pp. 109–118. 38

Alku, P., Bäckström, T. y Vilkman, E. (2002). "Normalized amplitude quotient for parameterization of the glottal flow". *Journal of the Acoustical Society of America (JASA)*, **12(2)**, pp. 701–710. 41

Alku, P., Story, B. y Airas, M. (2004). "Evaluation of an inverse filtering technique using physical modeling of voice production". En: *The 8th International Conference on Spoken Language Processing (ICSLP'2004)*, pp. 497–500. Jeju Island, South-Korea. 38

Alku, P. y Vilkman, E. (1994). "Estimation of the glottal pulseform based on discrete all-pole modeling". En: *The 3rd International Conference on Spoken Language Processing (ICSLP'1994)*, pp. 1619–1622. Yokohama, Japan. 38

Alter, K., Rank, E., Kotz, S. A., Toepel, U., Besson, M., Schirmer, A. y Friederici, A. D. (2003). "Affective encoding in the speech signal and in event-related brain potentials". *Speech Communication*, **40**, pp. 61–70. 44

Ananthapadmanabha, T. V. (1984). "Acoustic analysis of voice source dynamics". En: *STL-QPSR 2-3, Speech, Music and Hearing, Royal Institute of Technology*, pp. 1–24. Stockholm, Sweden. 39

Arnold, M. B. (1960). *Emotion and personality*. volumen 1-2. Columbia University Press, New York. 59

Aronson, A. E. y Bless, D. M. (2009). *Clinical voice disorders*. Thieme Medical Publishers, Inc., 4ª edición. 52

Atal, B. S. y Hanauer, S. L. (1971). "Speech analysis and synthesis by linear prediction of the speech wave". *Journal of the Acoustical Society of America (JASA)*, **50(2B)**, pp. 637–655. 26

Audibert, N., Vincent, D., Aubergé, V. y Rosec, O. (2006). "Expressive Speech Synthesis: Evaluation of a Voice Quality Centered Coder on the Different Acoustic Dimensions". En: *Speech Prosody*, Dresden, Germany. 158

Averill, J. R. (1980). "A constructivist view of emotion". En: R. Plutchik & H. Kellerman (Ed.), *Emotion: Theory, research and experience*, volumen 1, pp. 305–339. Academic Press, New York. 59

Aylett, M. P. y Yamagishi, J. (2008). "Combining statistical parameteric speech synthesis and unit-selection for automatic voice cloning". En: *LangTech 2008*, Brisbane, Australia. 80

Baken, R. J. (1987). *Clinical Measurements of Speech and Voice*. College-Hill Press, Boston, MA. 40

Banse, R. y Scherer, K. R. (1996). "Acoustic profiles in vocal emotion expression". *Journal of Personality and Social Psycology*, **70(3)**, pp. 614–636. 35, 44, 46, 86

Barner, K. E. (1996). "Nonlinear estimation of DEGG signals with applications to speech pitch detection". En: *The 4th International Conference on Spoken Language Processing (ICSLP'1996)*, pp. 2243–2246. Philadelphia, USA. 96

Barra-Chicote, R., Montero, J. M., Macías-Guarasa, J., Lufti, S., Lucas, J. M., Fernandez, F., D'haro, L. F., San-Segundo, R., Ferreiros, J., Cordoba, R. y Pardo, J. M. (2008). "Spanish Expressive Voices: Corpus for Emotion Research in Spanish". En: *The 6th International Conference on Language Resources and Evaluation (LREC'2008)*, Marrakech, Morocco. 91

Bäckström, T. (2004). *Linear Predictive Modelling of Speech-Constraints and Line Spectrum Pair Decomposition*. Tesis doctoral, Helsinki University of Technology, Espoo, Finland. 14, 15

Biemans, M. (2000). *Gender variation in voice quality*. Tesis doctoral, Netherland Graduate School of Linguistics, Utrecht, The Netherlands. 18, 34, 35, 36, 50

Black, A. W. (2003). "Unit Selection and Emotional Speech". En: *The 8th European Conference on Speech Communication and Technology (Eurospeech'2003)*, volumen 3, pp. 1649–1652. Geneva, Switzerland. 84, 91

Black, A. W. y Campbell, N. (1995). "Optimising selection of units from speech databases for concatenative synthesis". En: *The 4th European Conference on Speech Communication and Technology (Eurospeech'1995)*, volumen 1, pp. 581–584. Madrid, Spain. 76

Black, A. W. y Font Llitjós, A. (2002). "Unit Selection without a phoneme set". En: *IEEE 2002 Workshop on Speech Synthesis*, pp. 207–210. Santa Monica, USA. 95

Black, A. W. y Lenzo, K. (2001). "Optimal Data Selection for Unit Selection Synthesis". En: *The 4th ISCA Workshop on Speech Synthesis*, Perthshire, Scotland. 76

Bänziger, T. y Scherer, K. R. (2003). "A study of perceived vocal features in emotional speech". En: *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis (VOQUAL'2003)*, pp. 169–172. Geneva, Switzerland. 86

Boersma, P. (1998). *Functional Phonology Formalizing the interactions between articulatory and perceptual drives*. Tesis doctoral, University of Amsterdam. 75

Boersma, P. (2001). "Praat, a system for doing phonetics by computer". *Glot International*, **5:9/10**, pp. 341–345. 44, 107, 108, 165

Boula de Mareüil, P., Célérier, P. y Toen, J. (2002). "Generation of emotions by a morphing technique in English, French and Spanish". En: *Speech Prosody*, pp. 187–190. Aix-en-Provence, France. 85

Bozkurt, B., Ozturk, O. y Dutoit, B. (2003). "Text design for TTS speech corpus building using a modified greedy selection". En: *The 8th European Conference on Speech Communication and Technology (Eurospeech'2003)*, pp. 277–280. Geneva, Switzerland. 93

Browman, C. P. y Goldstein, L. (1986). "Towards an articulatory phonology". *Phonology Yearbook*, **3**, pp. 219–252. 75

Bulut, M., Narayanan, S. S. y Syrdal, A. (2002). "Expressive speech synthesis using a concatenative synthesizer". En: *The 7th International Conference on Spoken Language Processing (ICSLP'2002)*, pp. 1265–1268. 84

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. y Weiss, B. (2005). "A Database of German Emotional Speech". En: *The 9th European Conference on Speech Communication and Technology (Interspeech'2005)*, pp. 1517–1520. Lisboa, Portugal. 102

Burkhardt, F. y Sendlmeier, W. F. (2000). "Verification of acoustical correlates of emotional speech using formant synthesis". En: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pp. 151–156. Northern Ireland.. 85, 88

Cabral, J. y Oliveira, L. (2005). "Pitch-synchronous time-scaling for prosodic and voice quality transformations". En: *The 9th European Conference on Speech Communication and Technology (Interspeech'2005)*, pp. 1137–1140. Lisboa, Portugal. 63, 158

Cabral, J. y Oliveira, L. (2006). "EmoVoice: a System to Generate Emotions in Speech". En: *The 9th International Conference on Spoken Language Processing (Interspeech'2006)*, pp. 1798–1801. Pittsburgh, USA. 35, 71, 127

Cahn, J. E. (1989). *Generating Expression in Synthesized Speech*. Proyecto Final de Carrera, Massachusetts Institute of Technology. 84

Cahn, J. E. (1990). "The generation of affect in synthesized speech". *Journal of the American Voice I/O Society*, **8**, pp. 1–19. 85, 88

Calzada, À. (2008). *Estudi d'esquemes de modificació de les característiques vocals*. Proyecto Final de Carrera, Enginyeria i Arquitectura La Salle (Universitat Ramon Llull). 155, 160, 165, 204

Calzada, À. (2010). *Expressive Synhesis based on Harmonic plus Stochastic model*. Diploma de Estudios Avanzados, Univertat Ramon Llull. 155, 160, 165, 204

Campbell, N. (2000). "Databases of emotional speech". En: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pp. 34–38. Newcastle, Northern Ireland, UK. 92

Campbell, N. (2002). "Recording techniques for capturing natural everyday speech". En: *The 3rd International Conference on Language Resources and Evaluation (LREC'02)*, pp. 2029–2032. Las Palmas de Gran Canaria, Spain. 89, 92

Campbell, N. (2003). "Towards Synthesising Expressive Speech; Designing and Collecting Expressive Speech Data". En: *The 8th European Conference on Speech Communication and Technology (Eurospeech'2003)*, pp. 1637–1640. 84

Campbell, N. (2005). "Developments in corpus-based speech synthesis: approaching natural conversational speech". *IEICE Transactions on Information and Systems*, **E88-D(3)**, pp. 376–383. 92

Campbell, N. y Marumoto, T. (2000). "Automatic labelling of voice-quality in speech databases for synthesis". En: *The 6th International Conference on Spoken Language Processing (ICSLP'2000)*, volumen 4, pp. 468–471. Beijing, China. 85, 88

Campbell, N. y Mokhtari, P. (2003). "Voice Quality: the 4th Prosodic Dimension". En: *The 15th International Congress of Phonetic Sciences (ICPhS'2003)*, pp. 2417–2420. Barcelona, Spain. 34, 35, 86

Camps, J., Bailly, G. y Martí, J. (1992). "Synthèse à partir du texte pour le catalan". En: *19èmes Journées d'Études sur la Parole*, pp. 329–333. Brussels, France. 82

Chan, D. S. F. y Brookes, D. M. (1989). "Variability of excitation parameters derived from robust closed phase glottal inverse filtering". En: *The 1st European Conference on Speech Communication and Technology (Eurospeech'1989)*, p. 33.1. Paris, France. 38

Cornelius, R. R. (2000). "Theoretical Approaches to Emotion". En: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pp. 3–10. Newcastle, Northern Ireland, UK. 59, 60, 65

Cowie, R. y Cornelius, R. R. (2003). "Describing the emotional states that are expressed in speech". *Speech Communication*, **40**, pp. 5–32. 58, 60, 61

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M. y Schröder, M. (2000). "'FEELTRACE': an instrument for recording perceived emotion in real time". En: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pp. 19–24. Belfast, Northern Ireland. xvii, 62

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. y Taylor, J.G. (2001). "Emotion Recognition in Human-Computer Interaction". *IEEE Signal Processing Magazine*, **18(1)**, pp. 32–80. xxi, 2, 35, 63, 64, 66

Cranen, B. y Boves, L. (1985). "Pressure measurements during speech production using semiconductor miniature pressure transducers: impact on models for speech production". *Journal of the Acoustical Society of America (JASA)*, **77(4)**, pp. 1543–1551. 37

da S. Maia, R., Zen, H., Tokuda, K., Kitamura, T. y Resende Jr., F. G. V. (2003). "Towards the development of a Brazilian Portuguese Text-to-Speech System Based on HMM". En: *The 8th European Conference on Speech Communication and Technology (Eurospeech'2003)*, pp. 2465–2468. Geneva, Switzerland. 78

Darwin, C. (1872). "Darwin Online: Darwin's Publications". `http://darwin-online.org.uk/pdf/1873_Expression_F1143.pdf` 59

Davis, S. B. (1976). "Computer evaluation of laryngeal pathology based on inverse filtering of speech". *Informe técnico*, Santa Barbara, CA. 30

de Cheveigné, A. y Kawahara, H. (2002). "YIN, a fundamental frequency estimator for speech and music". *Journal of the Acoustical Society of America (JASA)*, **111(4)**, pp. 1917–1930. 95, 96, 220

Dejonckere, P. H., Remacle, M., Fresnel-Elbaz, E., Woisard, V., Crevier, L. y Millet, B. (1998). "Reliability and clinical relevance of perceptual evaluation of pathological voices". En: *Revue de Laryngologie Otologie Rhinologie*, volumen 119, pp. 247–248. 52, 53

Depalle, P. y Helie, T. (1997). "Extraction of spectral peak parameters using a short-time Fourier transform modeling and no sidelobe windows". En: *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, p. 4. 204

Devillers, L., Vidrascu, L. y Lamel, L. (2005). "Challenges in real-life emotion annotation and machine learning based detection". *Neural Networks*, **18**, pp. 407–422. 90

Dikshit, P., Zahorian, S. A. y Nagulapati, S. (2005). "A two-phase pitch marking method for TD-PSOLA synthesis". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2005)*, volumen 1, pp. 233–236. Philadelphia, USA. 220

Ding, W., Kasuya, H. y Adachi, S. (1994). "Simultaneous estimation of vocal tract and voice source parameters with application to speech synthesis". En: *The 3rd International Conference on Spoken Language Processing (ICSLP1994)*, pp. 159–162. Yokohama, Japan. 38

Douglas-Cowie, E., Campbell, N., Cowie, R. y Roach, P. (2003). "Emotional speech: towards a new generation of databases". *Speech Communication*, **40**, pp. 33–60. 90

Drioli, C., Tisato, G., Cosi, P. y Tesser, F. (2003). "Emotions and Voice Quality: Experiments with Sinusoidal Modeling". En: *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis (VOQUAL'2003)*, pp. 127–132. Geneva, Switzerland. 35, 36, 37, 44, 46, 50, 63, 71, 86, 107, 135, 138, 154

Dudley, H. y Tarnoczy, T. H. (1950). "The Speaking Machine of Wolfgang von Kempelen". *Journal of the Acoustical Society of America (JASA)*, **22(2)**, pp. 151–166. 74

Dutoit, T. y Leich, H. (1996). "MBR-PSOLA: Text-To-Speech synthesis based on an MBE re-synthesis of the segments database". *Speech Communication*, **13**, pp. 435–440. 83

Dutoit, T. y Stylianou, Y. (1997). *Text-to-Speech Synthesis*. Oxford University Press. 70, 83

Edgington, M. (1997). "Investigating the limitations of concatenative synthesis". En: *The 5th European Conference on Speech Communication and Technology (Eurospeech'1997)*, pp. 593–596. 88

Ekman, P. (1999). "Basic Emotions". En: T. Dalgleish y M. Power (Ed.), *Handbook of Cognition and Emotion*, Wiley & Sons, Ltd., Sussex, U. K.. 58, 60

Erro, D. (2008). *Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models*. Tesis doctoral, Universitat Politècnica de Catalunya, Barcelona, Spain. 79, 154, 155, 156

Esquerra, I. (2006). "Síntesis de habla emocional por selección de unidades". En: *IV Jornadas en Tecnología del Habla*, pp. 161–165. 84

Fabre, P. (1957). "Un procédé électrique percutané d'inscription de l'accolement glottique au cours de la phonation: glottographie de fréquence; premiers résultats". *Bulletin de l'Académie Nationale de Médecin*, **141**, pp. 66–69. 38

Fairbanks, G. (1960). *Voice and articulation drillbook*. Harper, New York. 51, 53

Faúndez, M. (2001). *Sistemas de Comunicaciones*. Marcombo. 45

Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton and Co, 's Gravenhage. 12, 25, 43

Fant, G. (1979a). "Glottal source and excitation analysis". *Informe técnico*, STL-QPSR 1, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden. 39

Fant, G. (1979b). "Vocal source analysis - a progress report". *Informe técnico*, STL-QPSR 3-4, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden. 39

Fant, G. (1982). "The voice source - acoustic modeling". *Informe técnico*, STL-QPSR 4, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden. 39

Fant, G., Kruckenberg, A., Gustafson, K. y Liljencrants, J. (2002). "A new approach to intonation analysis and synthesis of Swedish". En: *Speech Prosody*, pp. 283–286. Aix en Provence, France. 122

Fant, G., Liljencrants, J. y Lin, Q. (1985). "A four-parameter model of glottal flow". *Informe técnico*, STL-QPSR 4, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden. 13, 37, 39

Flanagan, J. L. (1972). *Speech Analysis, Synthesis and Perception*. Springer, Berlin-Heidelberg-New York. 25

Fourcin, A. J. y Abberton, E. (1971). "First applications of a new Laryngograph". *Medical and Biological Illustration*, **21**, pp. 172–182. 38

François, H. y Boëffard, O. (2002). "The greedy algorithm and its application to the construction of a continuous speech database". En: *The 3rd International Conference on Language Resources and Evaluation (LREC'2002)*, volumen 5. Las Palmas, Spain. 92, 97, 231

Fröhlich, M., Michaelis, D. y Strube, H. W. (2001). "SIM—simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals". *Journal of the Acoustical Society of America (JASA)*, **110(1)**, pp. 479–488. 38

Frigge, M., Hoaglin, D. C. y Iglewicz, B. (1989). "Some Implementations of the Boxplot". *American Statistician*, **43(1)**, pp. 50–54. 239

Frokjaer-Jensen, B. y Prytz, S. (1976). "Registration of voice quality". *Bruel and Kjar Technical Review*, **(3)**, pp. 3–17. 30

Frokjaer-Jensen, B. y Thorvaldsen, P. (1968). "Construction of a Fabre Glottograph". *Annual Report of the Institute of Phonetics, University of Copenhagen*, **3**, pp. 1–8. 38

Fujisaki, H. y Ljungqvist, M. (1986). "Proposal and evaluation of models for the glottal source waveform". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2005)*, volumen 11, pp. 1605–1608. Tokyo, Japan. 39

García, M. (2007). *Voice Quality Determination Using Inverse LPC-Filtering*. Proyecto Final de Carrera, Enginyeria i Arquitectura La Salle (Universitat Ramon Llull), Barcelona, Spain. 18

García, P. (2002). *Consulta a textos digitalizados: implementacion y analisis en el contexto de las colecciones especiales de la UDLA*. Proyecto Final de Carrera, Universidad de las Américas, Cholula, Mexico. 232

Garrido, J. M. (1991). *Modelización de patrones melódicos del español para la síntesis y el reconocimiento del habla.* Universitat Autònoma de Barcelona, Bellaterra. 63, 86

Gelfer, M. P. (1988). "Perceptual attributes of voice: Development and use of rating scales". *Journal of Voice*, **2(4)**, pp. 320–326. 52, 53

Gerratt, B. R., Hanson, D. G. y Berke, G. S. (1986). "Glottographic measures of laryngeal function in individuals with abnormal motor control". En: K. Harris, C. Sasaki y T. Baer (Eds.), *Vocal Fold Physiology: Laryngeal Function in Phonation and Respiration*, College Hill Press, San Diego, CA. 40

Gerratt, B. R., Hanson, D. G. y Berke, G. S. (1988). "Laryngeal configuration associated with glottography". *American Journal of Otolaryngology*, **9**, pp. 173–179. 40

Gerratt, B. R. y Kreiman, J. (2001). "Measuring vocal quality with speech synthesis". *Journal of the Acoustical Society of America (JASA)*, **110(5)**, pp. 2560–2566. 54

Gerratt, B. R. y Kreiman, J. (2003). "Voice Quality, Perceptual Evaluation of". En: Raymon D. Kent (Ed.), *The MIT encyclopedia of communication disorders*, pp. 78–80. The MIT Press. 51, 52, 53

Gil, J. (2007). *Fonética para profesores de español: de la teoría a la práctica*. Manuales de formación de profesores de español 2/L. Editorial ARCO/LIBROS S.L.. 32, 33, 63

Gobl, C., Bennett, E. y Ní Chasaide, A. (2002). "Expressive synthesis: how crucial is voice quality?" En: *IEEE 2002 Workshop on Speech Synthesis*, 11–13, pp. 91–94. 57, 86

Gobl, C. y Ní Chasaide, A. (2003). "The role of voice quality in communicating emotion, mood and attitude". *Speech Communication*, **40**, pp. 189–212. XVII, 35, 37, 38, 40, 41, 50, 63, 68, 69, 86

Goncharoff, V. y Gries, P. (1998). "An algorithm for accurately marking pitch pulses in speech signals". En: *IASTED International Conference on Signal and Image*, pp. 281–284. Las Vegas, USA. 101, 102, 221

Gonzalvo, X., Iriondo, I., Socoró, J. C., Alías, F. y Monzo, C. (2007a). "HMM-based Spanish speech synthesis using CBR as F0 estimator". En: *ISCA Tutorial and Research Workshop on Non Linear Speech Processing (NOLISP'2007)*, pp. 7–10. Paris, France. 182

Gonzalvo, X., Iriondo, I., Socoró, J. C., Alías, F. y Monzo, C. (2007b). "Mixing HMM-Based Spanish Speech Synthesis with a CBR for Prosody Estimation". En: *Advances in Nonlinear Speech Processing*, volumen 4885/2007 de *Lecture Notes in Computer Science (LNCS)*, pp. 78–85. Springer, Heidelberg. 182

Gonzalvo, X., Socoró, J. C., Iriondo, I., Monzo, C. y Martínez, E. (2007c). "Linguistic and Mixed Excitation Improvements on a HMM-based speech synthesis for Castilian Spanish". En: *The 6th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis (SSW6)*, pp. 362–367. Bonn, Germany. 182

Gonzalvo, X., Taylor, P., Monzo, C., Iriondo, I. y Socoró, J. C. (2009). "High Quality Emotional HMM-Based Synthesis In Spanish". En: *ISCA Tutorial and Research Workshop on Non Linear Speech Processing (NOLISP'2009)*, Vic, Spain. 182

Gonzalvo, X., Taylor, P., Monzo, C., Iriondo, I. y Socoró, J. C. (2010). "High Quality Emotional HMM-Based Synthesis in Spanish". En: *Advances in Nonlinear Speech Processing*, volumen 5933/2010 de *Lecture Notes in Computer Science (LNCS)*, pp. 26–34. Springer, Heidelberg. 182

Guaus, R., Gudayol, F. y Martí, J. (1996). "Conversión Texto-Voz mediante síntesis PSOLA". En: *Jornadas Nacionales de Acústica*, pp. 355–358. Barcelona. 82

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J. y Wedin, L. (1980). "Perceptual and acoustic correlates of abnormal voice qualities". *Acta Otolaryngologica*, **90**, pp. 441–451. 46

Hammarberg, B. y Gauffin, J. (1995). "Vocal fold physiology: Voice quality control". En: M. Hirano y O. Fujimura (Eds.), *Perceptual and acoustic characteristics of quality differences in pathological voices as related to physiological aspects*, pp. 283–303. Singular Publishing Group, San Diego. 52, 53

Hanson, D. G., Gerratt, B. R., Karin, R. R. y Berke, G. S. (1988). "Glottographic measures of vocal fold vibration: An examination of laryngeal paralysis". *Laryngoscope*, **98**, pp. 541–549. 40

Hanson, D. G., Jiang, J., D'Agostino, M. M. y Herzon, G. (1995). "Clinical measurement of mucosal wave velocity using simultaneous photoglottography and laryngostroboscopy". *American Laryngological, Rhinological and Otological Society, Inc.*, **5**, pp. 340–349. 40

Harbeck, S., Kießling, A., Kompe, R., Niemann, H. y Nöth, E. (1995). "Robust pitch period detection using dynamic programming with an ANN cost function". En: *Proceedings of EuroSpeech*, volumen 2, pp. 1337–1340. Madrid, Spain. 96, 101, 220

Hedelin, P. (1984). "A glottal LPC-vocoder". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'1984)*, volumen 9, pp. 21–24. San Diego, USA. 39

Heuft, B., Portele, T. y Rauth, M. (1996). "Emotions in time domain synthesis". En: *Proceedings of the 4th International Conference of Spoken Language Processing*, volumen 3, pp. 1974–1977. Philadelphia, USA. 88

Hiki, S., Imaizumi, S., Hirano, M., Matsushita, H. y Kakita, Y. (1976). "Acoustical analysis for voice disorders". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'1976)*, pp. 613–616. 31

Hirano, M. (1975). "Phonosurgery: Basic and clinical investigations". *Otologia Fukuoka*, **21**, pp. 229–440. 31

Hirano, M. (1981). "Clinical Examination of Voice". En: Godfrey E. Arnold, Fritz Winckel y Barry D. Wyke (Eds.), *Disorders of Human Communications*, volumen 5. Springer-Verlag Wien, New York. XXI, 2, 8, 19, 20, 21, 51, 53

Hirano, M., Kakita, Y., Matsushita, H., Hiki, S. y Imaizumi, S. (1977a). "Correlation between parameters related to vocal vibration and acoustical parameters in voice disorders". *Practica Otologica (Kyoto)*, **70**, pp. 393–403. 31

Hirano, M., Kakita, Y., Matsushita, H., Hiki, S. y Imaizumi, S. (1977b). "Psychoacoustic parameters in voice disorders". *Practica Otologica (Kyoto)*, **70**, pp. 525–531. 31

Hirano, M., Matsushita, H., Hiki, S. y Kakita, Y. (1976). "Acoustic analysis for voice disorders: A basic conception for the use of acoustic measurements for the diagnosis in voice disorders". *Practica Otologica (Kyoto)*, **69**, pp. 267–271. 31

Hornbeck, R. W. (1975). *Numerical Methods*. Quantum, New York. 40

Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A. y Nogueiras, A. (2002). "Interface databases: Design and collection of a multilingual emotional speech database". En: *The 3rd International Conference on Language Resources and Evaluation (LREC'02)*, pp. 2019–2023. Las Palmas de Gran Canaria, Spain. 91

Hunt, A. y Black, A. W. (1996). "Unit selection in a concatenative speech synthesis system using a large speech database". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'1996)*, pp. 373–376. Atlanta, Canada. 76

Iida, A., Campbell, N., Iga, S., Higuchi, F. y Yasumura, M. (2000). "A speech synthesis system with emotion for assisting communication". En: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pp. 167–172. Northern Ireland. 88

Imaizumi, S., Hiki, S., Hirano, M. y Matsushita, H. (1980). "Analysis of pathological voices with a sound spectrograph". *Journal of the Acoustical Society of Japan*, **36**, pp. 9–16. 31

Iribar, Alexander (2008a). "Física de los sonidos del lenguaje. La fonética acústica". `http://paginaspersonales.deusto.es/airibar/Fonetica/Apuntes/03.html` 10, 11

Iribar, Alexander (2008b). "La producción de los sonidos del lenguaje. La fonética articulatoria". `http://paginaspersonales.deusto.es/airibar/Fonetica/Apuntes/02.html` 9

Iriondo, I. (2008). *Producción de un corpus oral y modelado prosódico para la síntesis del habla expresiva*. Tesis doctoral, Universitat Ramon Llull. 33, 63, 64, 86, 134

Iriondo, I., Alías, F. y Melenchón, J. (2002). "Un modelo híbrido orientado a la síntesis multimodal del habla". En: *Procesamiento del Lenguaje Natural*, volumen 29, pp. 159–163. Valladolid. 83

Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J. M., Bernadas, D., Oliver, J. M., Tena, D. y Longhi, L. (2000). "Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques". En: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pp. 161–166. Northern Ireland. 83, 85

Iriondo, I., Planet, S., Socoró, J. C., Alías, F., Monzo, C. y Martínez, E. (2007a). "Expressive speech corpus validation by mapping subjective perception to automatic classification based on prosody and voice quality". En: *The 16th International Congress of Phonetic Sciences (ICPhS'2007)*, pp. 2125–2128. Saarbrücken, Germany. 71, 139, 152, 178

Iriondo, I., Planet, S., Socoró, J. C., Martínez, E., Alías, F. y Monzo, C. (2009). "Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification". *Speech Communication*, **51**, pp. 744–758. 138, 139, 152, 158, 178, 179, 181

Iriondo, I., Planet, S., Socoró, J.C. y Alías, F. (2007b). "Objective and subjective evaluation of an expressive speech corpus". En: *ISCA Tutorial and Research Workshop on Non Linear Speech Processing (NOLISP'2007)*, pp. 15–18. Paris, France. 63, 180

Iriondo, I., Socoró, J. C. y Alías, F. (2007c). "Prosody modelling of Spanish for expressive speech synthesis". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2007)*, volumen 4, pp. 821–824. Honolulu, HI, USA. 2, 51, 71, 86, 93, 121, 132, 156, 158, 160, 182

ITU-P.800 (1996). "Methods for subjective determination of transmission quality". 51, 121, 133, 163, 166, 183

James, W. (1884). "Classics in the History of Psychology". `http://psychclassics.yorku.ca/James/emotion.htm` 59

Jiang, J. J., Shuangyi, T., Dalal, M., Chi-Haur, W. y Hanson, D. G. (1998). "Integrated Analyzer and Classifier of Glottographic Signals". *IEEE Transactions on Rehabilitation Engineering*, **6(2)**, pp. 227–234. 40

John, G. H. y Langley, P. (1995). "Estimating Continuous Distributions in Bayesian Classifiers". En: *The 11th Conference on Uncertainty in Artificial Intelligence*, pp. 338–345. Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Mateo, USA. 233, 242

Johnston, J. D. (1988). "Transform coding of audio signals using perceptual noise criteria". *IEEE Journal on Selected Areas in Communications*, **6(2)**, pp. 314–323. 46

Juslin, P. N. y Laukka, P. (2003). "Communication of emotions in vocal expression and music performance: Different channels, same code?" *Psychological Bulletin*, **129(5)**, pp. 770–814. 64, 65

Kakita, Y., Hirano, M., Matsushita, H., Hiki, S. y Imaizumi, S. (1977a). "Acoustical parameters relevant to diagnosis in voice disorders". *Practica Otologica (Kyoto)*, **70**, pp. 269–276. 31

Kakita, Y., Hirano, M., Matsushita, H., Hiki, S. y Imaizumi, S. (1977b). "Differentiation of laryngeal diseases using acoustical analysis". *Practica Otologica (Kyoto)*, **70**, pp. 729–739. 31

Karjalainen, M. (2008). "Kommunikaatioakustiikka". `http://www.acoustics.hut.fi/teaching/S-89.3320/KA4.pdf` XVII, 9

Kasi, K. y Zahorian, S. A. (2002). "Yet another algorithm for pitch tracking". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2002)*, pp. 361–364. Orlando, USA. 222

Kasuya, H., Maekawa, K. y Kiritani, S. (1999). "Joint estimation of voice source and vocal tract parameters as applied to the study of voice source dynamics". En: *The 14th International Congress of Phonetic Sciences (ICPhS'1999)*, pp. 2505–2512. San Fracisco, USA. 38

Kawahara, H., Masuda-Katsuse, I. y de Cheveigné, A. (1999). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction". *Speech Communication*, **27**, pp. 187–207. 172

Kawanami, H., Iwami, Y., Toda, T., Saruwatari, H. y Shikano, K. (2003). "GMM-based voice conversion applied to emotional speech synthesis". En: *The 8th European Conference on Speech Communication and Technology (Eurospeech'2003)*, pp. 2401–2404. Geneva, Switzerland. 184

Keller, E. (2003). "Voice characteristics of MARSEC speakers". En: *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis (VOQUAL'2003)*, pp. 97–102. 56

Keller, E. (2005). "The Analysis of Voice Quality in Speech Processing". *Lecture Notes in Computer Science (LNCS)*, **3445**, pp. 54–73. 2, 17, 18, 34, 40, 50, 54, 55, 57, 135

Kempe, André (1999). "Experiments in unsupervised entropy-based corpus segmentation". En: *Workshop on Computational Natural Language Learning (CoNLL'99)*, pp. 7–13. Bergen, Norway. 93

Kitzing, P. y Löfqvist, A. (1975). "Subglottal and oral pressure during phonation - preliminary investigation using a miniature transducer system". *Medical and Biological Engineering*, **13**, pp. 644–648. 37

Klasmeyer, G. (2000). "An automatic description tool for time-contours and long-term average voice features in large emotional speech databases". En: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pp. 66–71. 56

Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer". *Journal of the Acoustical Society of America (JASA)*, **67(3)**, pp. 971–995. 73

Klatt, D. H. y Klatt, L. C. (1990). "Analysis, synthesis, and perception of voice quality variations among female and male talkers". *Journal of the Acoustical Society of America (JASA)*, **87(2)**, pp. 820–857. 39

Kohavi, R. y Provost, F. (1998). "Glossary of Terms". *Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*, **30(2–3)**, pp. 271–274. 240

Koike, Y. (1969). "Vowel amplitude modulations in patients with laryngeal diseases". *Journal of the Acoustical Society of America (JASA)*, **45(4)**, pp. 839–844. 29, 45

Koike, Y. (1973). "Application of some acoustic measures for the evaluation of laryngeal dysfunction". *Studia Phonologica*, **7**, pp. 17–23. 28

Kounoudes, A., Naylor, P. A. y Brookes, M. (2002). "The DYPSA algorithm for estimation of glottal closure instants in voiced speech". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2002)*, volumen 1, pp. 349–352. Orlando, USA. 220

Kreiman, J. y Gerratt, B. R. (1996). "The perceptual structure of pathologic voice quality". *Journal of the Acoustical Society of America (JASA)*, **100(3)**, pp. 1787–1795. 52

Kreiman, J. y Gerratt, B. R. (1998). "Validity of rating scale measures of voice quality". *Journal of the Acoustical Society of America (JASA)*, **104(3)**, pp. 1598–1608. 52

Laroche, J., Stylianou, Y. y Moulines, E. (1993). "HNS: Speech modification based on a harmonic+noise model". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'1993)*, volumen 2, pp. 550–553. Minneapolis, USA. 79, 203

Laukka, P. (2004). *Vocal Expression of Emotion. Discrete-emotions and Dimensional Accounts*. Tesis doctoral, Uppsala Universitet. xxi, 68

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge University Press. xvii, 17, 34, 35, 52, 53, 56

Laver, J. (1991). *The Description of Voice Quality in General Phonetic Theory*. Edinburgh University Press. 17, 56

Laver, J. (2000). "Phonetic evaluation of voice quality". En: Ball M. J. (ed) Kent R. D. (Ed.), *Voice quality measurement*, pp. 37–48. Singular Thomson Learning, San Diego. 52, 53

Laver, J. (2002). *Principles of phonetics*. Cambridge University Press. 57

Lee, C. K. y Childers, D. G. (1991). "Some acoustical, perceptual, and physiological aspects of vocal quality". En: J. Gauffin y B. Hammarberg (Eds.), *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms*, pp. 233–242. Singular Publishing Group, San Diego. 38

Lee, J-Y., Jeong, S., Hahn, M. y Choi, H-S. (2008). "Automatic voice quality measurement based on efficient combination of multiple features". En: *The 2nd International Conference on Bioinformatics and Biomedical Engineering (ICBBE'2008)*, pp. 1272–1275. Shanghai, China. 122

Lieberman, P. (1963). "Some acoustic measures of the fundamental periodicity of normal and pathologic larynges". *Journal of the Acoustical Society of America (JASA)*, **35(3)**, pp. 344–353. 28, 29, 44

Lieberman, P. y Blumstein, S. E. (1998). *Speech physilogy, speech perception, and acoustic phonetics*. Cambridge University Press. 57

Lilliefors, H. W. (1967). "On the Komogorov-Smirnov test for normality with mean and variance unknown". *Journal of the American Statistical Association*, **62**, pp. 399–402. 127

Lin, C-Y. y Jang, J-S. R. (2004). "A two-phase pitch marking method for TD-PSOLA synthesis". En: *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP'2004)*, pp. 1189–1192. Jeju Island, South Korea. 220

Ljungqvist, M. y Fujisaki, H. (1985). "A method for simultaneous estimation of voice source and vocal tract parameters based on linear predictive analysis". *Transactions of the Committee on Speech Research, Acoustical Society of Japan*, **S85-21**, pp. 153–160. 38

Llisterri, J., Aguilar, L., Garrido, J. M., Machuca, M. J., Marín, R., de la Mota, C. y Ríos, A. (1999). "Fonética y tecnologías del habla". En: J. M. Blecua, G. Clavería, C. Sánchez y J. Torruella (Eds.), *Filología e informática. Nuevas tecnologías en los estudios filológicos*, pp. 449–479. Editorial Milenio, Barcelona: Seminario de Filología e Informática, Departamento de Filología Española, Universidad Autónoma de Barcelona. 8

Lugger, M. y Yang, B. (2006a). "Classification of different speaking groups by means of voice quality parameters". En: *ITG-Fachtagung Sprachkommunikation*, Kiel, Germany. 42, 138

Lugger, M. y Yang, B. (2006b). "Robust estimation of voice quality parameters under real world disturbances". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2006)*, volumen 1, pp. 1097–1100. Toulouse, France. 37, 42, 107

Lugger, M. y Yang, B. (2007). "The relevance of voice quality features in speaker independent emotion recognition". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2007)*, volumen 4, pp. 17–20. Honolulu, Hawaii, USA. 138

Macías, J. (2006). "El sistema de producción de habla". `http://insn.die.upm.es/docs/INSN0506-TemaProduccionDeHabla-JMG-v46.pdf` 8, 10, 14, 15

Makhoul, J. (1975). "Linear prediction: A tutorial review". En: *Proceedings of the IEEE*, volumen 63, pp. 561–580. 26

Markel, J. D. y Gray, A. H. (1976). *Linear Prediction of Speech*. Springer, New York. 26

Martí, J. (1985). *Estudi acústic del català i síntesi automàtica per ordinador*. Tesis doctoral, Universitat de València. 82

Martí, J. (1987). *Síntesis del habla: Evolución histórica y situación actual*. Marcombo, Barcelona. 82

Martí, J. (1990). "Estado actual de la síntesis de voz". *Estudios de Fonética Experimental*, **4**, pp. 147–168. 82

Möbius, B. (2000). "Corpus-based speech synthesis: methods and challenges". En: *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS)*, volumen 6(4), pp. 87–116. 76

McKenna, J. y Isard, S. (1999). "Tailoring Kalman filtering toward speaker characterisation". En: *The 6th European Conference on Speech Communication and Technology (Eurospeech'1999)*, pp. 2793–2796. Budapest, Hungary. 38

Michaelis, D., Gramss, T. y Strube, H. W. (1997). "Glottal to noise excitation ratio - a new measure for describing pathological voices". *Acustica/acta acustica*, **83**, pp. 700–706. 45, 110

Mikheev, A. (2002). "Periods, Capitalized Words, etc". En: *Association for Computational Linguistics (ACL)*, volumen 28(3), pp. 289–318. Philadelphia, USA. 229

Milenkovic, P. (1986). "Glottal Inverse Filtering by Joint Estimation of an AR System with a Linear Input Model". *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-34(1)**, pp. 28–42. 38

Montero, J. M., Gutiérrez-Arriola, J., Colás, J., Enríquez, E. y Pardo, J.M (1999). "Analysis and modelling of emotional speech in Spanish". En: *The 14th International Congress of Phonetic Sciences (ICPhS'1999)*, pp. 957–960. San Francisco, USA. 35, 88

Montero, J. M., Gutiérrez-Arriola, J., Palazuelos, S., Enríquez, E., Aguilera, S. y Pardo, J. M. (1998). "Emotional speech synthesis: from speech database to TTS". En: *The 5th International Conference on Spoken Language Processing (ICSLP'1998)*, pp. 923–926. Sydney, Australia. 35, 85, 88, 91

Montoya, N. (1998). "El papel de la voz en la publicidad audiovisual dirigida a los niños". *Zer. Revista de estudios Comunicación*, **(4)**, pp. 161–177. 96

Monzo, C., Alías, F., Iriondo, I., Gonzalvo, X. y Planet, S. (2007). "Discriminating Expressive Speech Styles by Voice Quality Parameterization". En: *The 16th International Congress of Phonetic Sciences (ICPhS'2007)*, pp. 2081–2084. Saarbrücken, Germany. 107, 135, 180, 184

Monzo, C., Alías, F., Morán, J. A. y Gonzalvo, X. (2006). "Transcripción fonética de acrónimos en castellano utilizando el algoritmo C4.5". *Procesamiento del Lenguaje Natural*, **37**, pp. 275–282. 228

Monzo, C., Calzada, À., Iriondo, I. y Socoró, J. C. (2010). "Expressive Speech Style Transformation: Voice Quality and Prosody Modification Using a Harmonic plus Noise Model". En: *Speech Prosody*, Chicago, USA. 154

Monzo, C., Formiga, L., Adell, J., Iriondo, I., Alías, F. y Socoró, J. C. (2008a). "Adaptación del CTH-URL para la competición Albayzin 2008". En: *V Jornadas en Tecnología del Habla*, pp. 87–90. Bilbao, España. 93, 121, 132

Monzo, C., Iriondo, I. y Martínez, E. (2008b). "Procedimiento para la medida y la modificación del jitter y del shimmer aplicado a la síntesis del habla expresiva". En: *V Jornadas en Tecnología del Habla*, pp. 58–61. Bilbao, España. 108, 154, 158

Moon, T. K. y Stirling, W. C. (2000). *Mathematical methods and algorithms for Signal Processing*. Prentice Hall. 204

Moulines, E. y Charpentier, F. (1990). "Pitch-Sinchronous waveform processing techniques for text-to-speech synthesis using diphones". *Speech Communication*, **9**, pp. 453–467. 80, 82, 95

Mozziconacci, S. J. L. (1998). *Speech Variability and Emotion: Production and Perception*. Tesis doctoral, Technical University Eindhoven. 85

Mozziconacci, S. J. L. y Hermes, D. J. (1999). "Role of intonation patterns in conveying emotion in speech". En: *The 14th International Congress of Phonetic Sciences (ICPhS'1999)*, pp. 2001–2004. San Francisco, USA. 85, 88

Murray, I. R. y Arnott, J. L. (1993). "Toward the simulation of Emotion in Synthetic Speech: A Review of The Literature of Human Vocal Emotion". *Journal of the Acoustical Society of America (JASA)*, **93(2)**, pp. 1097–1108. xxi, 64, 65, 84, 86

Murray, I. R. y Arnott, J. L. (1995). "Implementation and testing of a system for producing emotion-by-rule in synthetic speech". *Speech Communication*, **16(4)**, pp. 369–390. 84, 85, 87, 88

Murray, I. R., Edgington, M. D., Campion, D. y Lynn, J. (2000). "Rule-based emotion synthesis using concatenated speech". En: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pp. 173–177. Northern Ireland. 85

Murty, G. E., Carding, P. N. y Kelly, P. J. (1991a). "The effect of intensity on combined glottography". *Clinical Otolaryngology*, **16(4)**, pp. 399–400. 40

Murty, G. E., Carding, P. N., Kelly, P. J. y Lancaster, P. (1991b). "The effect of frequency on combined glottography". *Clinical Otolaryngology*, **16(3)**, pp. 298–301. 40

Murty, G. E., Carding, P. N. y Lancaster, P. (1991c). "An outpatient clinic system for glottographic measurement of vocal fold vibration". *British Journal of Disorders of Communication*, **26**, pp. 115–123. 40

NARA (1995). "Using the Census Soundex". En: *U.S. National Archives and Records Administration*, Washington, D.C., USA. xxiv, 229

Navas, E., Hernáez, I. y Luengo, I. (2006). "An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS". *IEEE Transanctions Audio Speech Language Process*, **14(4)**, pp. 1117–1127. 93

Naylor, P. A., Kounoudes, A., Gudnason, J. y Brookes, M. (2007). "Estimation of Glottal Closure Instants in Voiced Speech Using the DYPSA Algorithm". *IEEE Transactions on Acoustics, Speech and Signal Processing*, **15(1)**, pp. 34–43. 38

Núñez, F., Corte, P., Suárez, C., Señaris, B. y Sequeiros, G. (2004). "Evaluación perceptual de la disfonía: correlación con los parámetros acústicos y fiabilidad". *Acta otorrinolaringológica española: Órgano Oficial de la Sociedad Española de Otorrinolaringología y Patología Cérvico-Facial*, **55(6)**, pp. 282–287. 44, 50, 120

Peng, H., Long, F. y Ding, C. (2005). "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27(8)**, pp. 1226–1238. 153

Planet, S., Iriondo, I., Martínez, E. y Montero, J. A. (2008). "TRUE: an online testing platform for multimedia evaluation". En: *The 2nd International Workshop on EMOTION: Corpora for Research on Emotion and Affect at the 6th Conference on Language Resources & Evaluation (LREC 2008)*, Marrakech, Morocco. 133, 162

Planet, S., Iriondo, I., Socoró, J. C., Monzo, C. y Adell, J. (2009). "GTM-URL Contribution to the INTERSPEECH 2009 Emotion Challenge". En: *The 10th Annual Conference of the International Speech Communication Association (Interspeech'2009). Special session: Emotion Challenge*, pp. 316–319. Brighton, United Kingdom. ISSN 1990-9772. 151

Plante, F., Meyer, G. y Ainsworth, W. A. (1995). "A pitch extraction reference database". En: *The 4th European Conference on Speech Communication and Technology (Eurospeech'1995)*, pp. 837–840. Madrid, Spain. 220

Plutchik, R. (1958). "Outlines of a new theory of emotion". *Transactions of the New York Academy of Sciences*, **20**, pp. 394–403. 60

Plutchik, R. (2001). "The nature of emotions". *American Scientist*, **89(4)**, pp. 344–350. xvii, 58, 60, 61

Pérez, H. E. (2003). "Frecuencia de fonemas". *Revista electrónica de Tecnología del Habla (e-RTH)*, **(1)**. xxi, xxi, 97, 98

Price, P. J. (1989). "Male and female voice source characteristics: inverse filtering results". *Speech Communication*, **8**, pp. 261–277. 39

Qi, Y. Y. y Bi, N. (1994). "Simplified approximation of the four-parameter LF model of voice source". *Journal of the Acoustical Society of America (JASA)*, **96(2)**, pp. 1182–1185. 39

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo. 230

Rank, E. y Pirker, H. (1998). "Generating emotional speech with a concatenative synthesizer". En: *The 5th International Conference on Spoken Language Processing (ICSLP'1998)*, volumen 3, pp. 671–674. Sydney, Australia. 85, 88

Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J., Bernadas, D., Oliver, J. y Longhi, L. (1999). "Modelización acústica de la expresión emocional en el español". En: *Procesamiento del Lenguaje Natural*, volumen 25, pp. 159–166. Lleida, Spain. 83

Rosenberg, A. E. (1971). "Effect of Glottal Pulse Shape on the Quality of Natural Vowels". *Journal of the Acoustical Society of America (JASA)*, **49(2B)**, pp. 583–590. 13, 39

Rothenberg, M., Carlson, R., Granström, B. y Lindqvist-Gauffin, J. (1975). *A three-parameter voice source for speech synthesis*. Fant, G. (Ed.), Proceedings of the Speech Communication Seminar, Stockholm, 1974, Vol. 2. Almqvist and Wiksell, Stockholm, Sweden. 39

Rubin, P. y Vatikiotis-Bateson, E. (1998). "Measuring and modeling speech production". En: S. L. Hopp, M. J. Owren y C. S. Evans (Eds.), *Animal Acoustic Communication*, capítulo 8, pp. 251–290. Springer-Verlag. 14

Ruinskiy, D. y Lavner, Y. (2008). "Stochastic models of pitch jitter and amplitude shimmer for voice modification". En: *Electrical and Electronics Engineers in Israel, 2008. IEEEI 2008. IEEE 25th Convention of*, pp. 489–493. Eilat, Israel. 127

Sagisaka, Y. (1988). "Speech synthesis by rule using an optimal selection of non-uniform synthesis units". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'1988)*, pp. 679–682. New York, USA. 76

Sagisaka, Y., Naiki, N., Iwahashi, N. y Mimura, K. (1992). "ATR - v - TALK speech synthesis system". En: *The 2nd International Conference on Spoken Language Processing (ICSLP'1992)*, volumen 1, pp. 483–486. Banff, Canada. 76

Scherer, K. R. (1986). "Vocal affect expression: a review and a model for future research". *Psichological Bulletin*, **99**, pp. 143–165. 59, 62

Scherer, K. R. (1989). *Vocal correlates of emotional arousal and affective disturbance*. Handbook of Psychophysiology: Emotion and social behavior. Wiley, London. 35, 46

Scherer, K.R. (1999). *Appraisal theory*. New York. 58, 88

Schoentgen, J. (1993). "Modelling the glottal pulse with a selfexcited threshold autoregressive model". En: *The 3rd European Conference on Speech Communication and Technology (Eurospeech'1993)*, pp. 107–110. Berlin, Germany. 39

Schröder, M. (2001). "Emotional Speech Synthesis: a Review". En: *The 7th European Conference on Speech Communication and Technology (Eurospeech'2007)*, pp. 561–564. 84

Schröder, M. (2004). *Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis*. Tesis doctoral, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University, Saarbrücken, Germany. xxi, xxi, 61, 84, 85, 86, 87, 88, 89, 90

Schuller, B., Steidl, S. y Batliner, A. (2009). "The Interspeech 2009 Emotion Challenge". En: *The 10th Annual Conference of the International Speech Communication Association (Interspeech'2009)*, pp. 312–315. 103, 151, 152

Sebastiani, F. (2002). "Machine Learning in Automated Text Categorization". *ACM Computing Surveys*, **34(1)**, pp. 1–47. 138, 234, 241

Sejnowski, T. J. y Rosenberg, C. R. (1987). "Parallel networks that learn to pronounce English text". *Journal of Complex Systems*, **1(1)**, pp. 145–168. 229

Sáenz-Lechón, N., Godino-Llorente, J. I., Osma-Ruiz, V., Blanco-Velasco, M. y Cruz-Roldán, F. (2006). "Automatic Assessment of Voice Quality According to the GRBAS Scale". En: *The 28th IEEE EMBS Annual International Conference*, pp. 2478–2481. New York City, USA. 53

Severin, F., Bozkurt, B. y Dutoit, T. (2005). "HNR extraction in voiced speech, oriented towards voice quality analysis". En: *The 13th European Signal Processing Conference (EUSIPCO'2005)*, Antalya,Turkey. 45

Slyh, R. E., Nelson, W. T. y Hansen, E. G. (1999). "Analysis of mrate, shimmer, jitter, and F0 contour features across stress and speaking style in the SUSAS database". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'1999)*, volumen 4, pp. 2091–2094. Phoenix, USA. 95, 120

Sondhi, M. M. (1975). "Measurement of the glottal waveform". *Journal of the Acoustical Society of America (JASA)*, **(57)**, pp. 228–232. 26

Sproat, R., Black, A. W., Chen, S., Shankar, S. Kumar, Ostendorf, M. y Richards, C. (1999). "Normalization of non-standard words: WS'99 final report". *Informe técnico*, The Center for Language and Speech Processing, Johns Hopkins University. 70, 215

Stallo, J. (2000). *Simulating emotional speech for a talking head*. Tesis doctoral, School of Computing, Curtin University of Technology, Australia. 85, 87, 88

Steidl, S. (2009). *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Logos Verlag, Berlin. 103

Stevens, K. y Hanson, H. (1994). "Classification of glottal vibration from acoustic measurements". *Vocal Fold Physiology*, pp. 147–170. 37, 42

Stevens, S. S., Volkmann, J. y Newman, E. B. (1937). "A Scale for the Measurement of the Psychological Magnitude Pitch". *Journal of the Acoustical Society of America (JASA)*, **8(3)**, pp. 185–190. 33

Strik, H., Jansen, J. y Boves, L. (1992). "Comparing methods for automatic extraction of voice source parameters from continuous speech". En: *The 2nd International Conference on Spoken Language Processing (ICSLP'1992)*, volumen 1, pp. 121–124. Banff, Canada. 38

Stylianou, Y. (1996). *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modifications*. Tesis doctoral, École Nationale Supérieure des Télécommunications, Paris, France. 79, 83

Stylianou, Y. (1998). "Concatenative Speech Synthesis using a Harmonic plus Noise Model". En: *The 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, pp. 261–266. Jenolan Caves, Australia. 79

Stylianou, Y. (1999). "Assessment And Correction Of Voice Quality Variabilities In Large Speech Databases For Concatenative Speech Synthesis". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'1999)*, pp. 377–380. 50

Stylianou, Y. (2001). "Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis". *IEEE Transactions on Speech and Audio Processing*, **9(1)**, pp. 21–29. 154

Stylianou, Y., Laroche, J. y Moulines, E. (1995). "High-quality speech modification based on a harmonic + noise model". En: *The European Conference on Speech Communication and Technology (Eurospeech'1995)*, pp. 451–454. 79, 86

Sun, X. (2000). "A pitch determination algorithm based on Subharmonic-to-Harmonic Ratio". En: *The 6th International Conference on Spoken Language Processing (ICSLP'2000)*, volumen 4, pp. 676–679. Beijing, China. 96

Sun, X. (2002). "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2002)*, volumen 1, pp. 333–336. Orlando, USA. 220, 221, 222

Swerts, M. y Veldhuis, R. (2001). "The effect of speech melody on voice quality". *Speech Communication*, **33**, pp. 297–303. 120

Takeda, K., Abe, K. y Sagisaka, Y. (1990). "On unit selection algorithms and their evaluation in non-uniform speech synthesis". En: *ESCA Workshop on Speech Synthesis*, pp. 35–38. Autrans, France. 76

Talkin, D. (1995). *A Robust Algorithm for Pitch Tracking (RAPT)*. capítulo 14, pp. 495–518. Speech Coding and Synthesis, Elsevier Science, Amsterdam, NL. 100, 152, 220

Talkin, D. y Rowley, J. (1990). "Pitch-synchronous analysis and synthesis for TTS systems". En: *ESCA Workshop on Speech Synthesis*, pp. 55–58. Autrans, France. 38

Taylor, P. (2006). "Unifying unit selection and hidden Markov model speech synthesis". En: *The 9th International Conference on Spoken Language Processing (Interspeech'2006)*, pp. 1758–1761. Pittsburgh, USA. 78

Taylor, P. (2009a). *Distinctive Features and Phonological Theories*. Cambridge University Press, New York. 75

Taylor, P. (2009b). *Text-to-Speech Synthesis*. Cambridge University Press. 74

Tesser, F., Cosi, P., Drioli, C. y Tisato, G. (2005). "Emotional Festival-MBROLA Synthesis". En: *The 9th European Conference on Speech Communication and Technology (Interspeech'2005)*, pp. 505–508. Lisboa, Portugal. 51, 86

Toda, T. y Tokuda, K. (2005). "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis". En: *The 9th European Conference on Speech Communication and Technology (Interspeech'2005)*, pp. 2801–2804. Lisboa, Portugal. 183

Tokuda, K., Kobayashi, T. y Imai, S. (1995). "Speech parameter generation from HMM using dynamic features". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'1995)*, volumen 1, pp. 660–663. 77

Tokuda, K., Zen, H. y Black, A. W. (2002). "An HMM-based speech synthesis system applied to English". En: *IEEE 2002 Workshop on Speech Synthesis*, pp. 227–230. Santa Monica, USA. 78

Torruella, J. y Llisterri, J. (1999). *Diseño de corpus textuales y orales*. Filología e informática. Nuevas tecnologías en los estudios filológicos. Milenio, Barcelona. Seminario de Filología e Informática, Departamento de Filología Española, Universidad Autónoma de Barcelona. 215

Trapp, T. K., Berke, G. S., Bell, T. S., Hanson, D. G. y Ward, P. (1989). "Effect of vocal fold augmentation on laryngeal vibration in simulated recurrent laryngeal nerve paralysis: A study of teflon and phonogel". *American Laryngological, Rhinological and Otological Society, Inc.*, **98**, pp. 220–227. 40

Türk, O. y Schröder, M. (2008). "A Comparison of Voice Conversion Methods for Transforming Voice Quality in Emotional Speech Synthesis". En: *The 9th Annual Conference of the International Speech Communication Association (Interspeech'2008)*, pp. 2282–2285. Brisbane, Australia. 35

Türk, O., Schröder, M., Bozkurt, B. y Arslan, L. M. (2005). "Voice quality interpolation for emotional text-to-speech synthesis". En: *The 9th European Conference on Speech Communication and Technology (Interspeech'2005)*, pp. 797–800. Lisboa, Portugal. 63

Tryon, Warren W. (1997). "Introduction to the Bidirectional Associative Memory Model: Implications for Psycopathology, Treatment, and Research". En: G. Matthews (Ed.), *Cognitive Science Perspectives on Personality and Emotion, 124*, pp. 65–122. North-Holland, Amsterdam, The Netherlands. 60

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc.. 242

Vatikiotis-Bateson, E. y Yehia, H. (1997). "Unified physiological model of audible-visible speech production". En: *The 5th European Conference on Speech Communication and Technology (Eurospeech'1997)*, pp. 2031–2034. Rhodes, Greece. 75

Veenevan, D. E. (1985). "Automatic Glottal Inverse Filtering from Speech and Electroglottographic Signals". *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-33(2)**, pp. 369–377. 38

Veldhuis, R. (1998). "A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation". *Journal of the Acoustical Society of America (JASA)*, **103(1)**, pp. 566–571. 39

Veldhuis, R. (2000). "Consistent Pitch Marking". En: *The 6th International Conference on Spoken Language Processing (ICSLP'2000)*, volumen 3, pp. 207–210. Beijing, China. 95

Verma, A. y Kumar, A. (2005). "Introducing Roughness in Individuality Transformation through Jitter Modeling and Modification". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2005)*, volumen 1, pp. 5–8. Philadelphia, USA. 120

von Kempelen, W. (1791). *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine*. J. B. Degen, Viena. XVII, 74

Vroomen, J., Collier, R. y Mozziconacci, S. J. L. (1993). "Duration and intonation in emotional speech". En: *The 3rd European Conference on Speech Communication and Technology (Eurospeech'1993)*, pp. 577–580. Berlin, Germany. 88

Wells, J.C. (1997). "SAMPA computer readable phonetic alphabet". En: D. Gibbon, R. Moore y R. R. Winski (Eds.), *Handbook of Standards and Resources for Spoken*, volumen Part IV, section B. Mouton de Gruyter, Berlin and New York. 97, 231

Wilhelms-Tricarico, R. (1995). "Physiological modeling of speech production: Methods for modeling soft-tissue articulators". *Journal of the Acoustical Society of America (JASA)*, **97(5)**, pp. 3085–3098. 75

Wilson, F. B. (1977). *Voice disorders*. Austin. 52, 53

Witten, I. H. y Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2ª edición. 230, 242

Wokurek, W. y Pützer, M. (2003). "Automated corpus based spectral measurement of voice quality parameters". En: *The 15th International Congress of Phonetic Sciences (ICPhS'2003)*, pp. 2173–2176. Barcelona, Spain. 42

Wong, D. Y., Markel, J. D. y Gray, A. H. (1979). "Least squares glottal inverse filtering from the acoustic speech waveform". *IEEE Transactions on Acoustics, Speech and Signal Processing*, **27(4)**, pp. 350–355. 16, 38

Yamagishi, J., Masuko, T. y Kobayashi, T. (2004). "HMM-based expressive speech synthesis – Towards TTS with arbitrary speaking styles and emotions". En: *Special Workshop in MAUI (SWIM), Lectures by Masters in Speech Processing*, pp. Conference CD–ROM, 1.13 (4 pages). 86

Yamagishi, J., Ogata, K., Nakano, Y., Isogai, J. y Kobayashi, T. (2006). "HSMM-Based Model Adaptation Algorithms for Average-Voice-Based Speech Synthesis". En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2006)*, volumen 1, pp. I–I. Toulouse, France. 184

Yanagihara, N. (1967a). "Hoarseness: Investigation of the physiological mechanisms". *Annals of Otology, Rhinology, and Laryngology*, **76**, pp. 472–489. 29

Yanagihara, N. (1967b). "Significance of harmonic changes and noise components in hoarseness". *Journal of Speech and Hearing Research*, **10**, pp. 531–541. 29

Yanushevskaya, I., Gobl, C. y Ní Chasaide, A. (2005). "Voice quality and f0 cues for affect expression: implications for synthesis". En: *The 9th European Conference on Speech Communication and Technology (Interspeech'2005)*, pp. 1849–1852. Lisboa, Portugal. 57, 86

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. y Kitamura, T. (1999). "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis". En: *The 6th European Conference on Speech Communication and Technology (Eurospeech'1999)*, pp. 2374–2350. Budapest, Hungary. 78

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. Andrew, Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. y Woodland, P. (2006). *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department. 217

Zemlin, W. R. (1988). *Speech and Hearing Science: Anatomy & Physiology*. 3rd ed. Englewood Cliffs, NJ: Prentice-Hall. 40

Zen, H., Toda, T., Nakamura, M. y Tokuda, K. (2007). "Details of Nitech HMM-based speech synthesis system for the blizzard challenge 2005". *IEICE Transactions on Information and Systems*, **E90-D(1)**, pp. 325–333. 182

Zwicker, E. (1961). "Subdivision of the audible frequency range into critical bands". *Journal of the Acoustical Society of America (JASA)*, **33(2)**, pp. 248–248. 42

Universitat Ramon Llull

Aquesta Tesi Doctoral ha estat defensada el dia ____ d _____ de ____

al Centre Escola Tècnica Superior d'Enginyeria Electrònica i Informàtica La Salle

de la Universitat Ramon Llull

davant el Tribunal format pels Doctors sotasignants, havent obtingut la qualificació:

President/a

_____

Vocal

_____

Vocal

_____

Vocal

_____

Secretari/ària

_____

Doctorand/a
Carlos Manuel Monzo Sánchez

C. Claravall, 1-3
08022 Barcelona
Tel. 936 022 200
Fax 936 022 249
E-mail: urlsc@sec.url.es
www.url.es