Universitat Ramon Llull

TESI DOCTORAL

Títol	Reconeixement afectiu automàtic mitjançant l'anàlisi de paràmetres acústics i lingüístics de la parla espontània
Realitzada per	Santiago Planet García
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. 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

Als meus pares

Resum

Aquesta tesi tracta el reconeixement automàtic d'emocions espontànies basat en l'anàlisi del senyal de veu. Es duu a terme dins del Grup de recerca de Tecnologies Mèdia (GTM) d'Enginyeria i Arquitectura La Salle, i té el seu origen en un moment en què existeixen obertes diverses línies d'investigació relacionades amb la síntesi afectiva però cap d'elles relacionada amb el seu anàlisi. La motivació és, doncs, millorar la interacció persona-màquina aportant un mòdul d'anàlisi a l'entrada dels sistemes que permeti, posteriorment, generar una resposta adequada mitjançant els mòduls de síntesi a la sortida dels mateixos. El centre d'atenció se situa en l'expressivitat afectiva, intentant dotar d'habilitats d'intel·ligència emocional a sistemes d'intel·ligència artificial, amb l'objectiu de què la interacció persona-màquina sigui el més similar possible a la comunicació humana.

En primer lloc es duu a terme un anàlisi preliminar basat en locucions enregistrades en condicions ideals. L'expressivitat vocal en aquest cas és actuada i les gravacions responen a un guió previ que estableix a priori l'etiqueta que descriu el contingut afectiu de les mateixes. Tot i què aquest no és el paradigma de la interacció en un entorn natural, aquest primer pas serveix per provar les primeres aproximacions a la parametrització dels corpus, els mètodes de selecció de paràmetres i la seva utilitat en l'optimització dels procediments, així com la viabilitat de considerar el sistema de reconeixement afectiu com un exercici de classificació categòrica. Tanmateix, permet comparar els resultats obtinguts en aquest escenari amb els que s'obtinguin posteriorment a l'escenari natural. Si bé es podria considerar que la utilitat d'un marc de treball com el que aquí es proposa no interès més enllà de l'exercici de comprovació abans explicat, a aquesta tesi es proposa un sistema basat en aquest plantejament la finalitat del qual és la validació automàtica d'un corpus de veu expressiva destinat a síntesi. En síntesi sí és necessari que el corpus estigui gravat en condicions òptimes perquè serà emprat per a la generació de noves locucions.

En segon lloc la tesi aprofundeix en l'anàlisi del corpus FAU Aibo, un corpus multilocutor de veu expressiva espontània enregistrat en alemany que recull un conjunt de gravacions d'interaccions naturals d'un grup de nens i nenes amb un robot que té instal·lat un micròfon. En aquest cas el plantejament és completament diferent a l'anterior començant per la definició del mateix corpus, en el que les locucions no responen a un guió previ i les etiquetes afectives són assignades posteriorment considerant l'avaluació

subjectiva de les mateixes. Tanmateix, el grau d'expressivitat emocional d'aquestes locucions és inferior al de les enregistrades per un actor o una actriu en tant que són espontànies i les emocions, donat que es generen de forma natural i no es treballa en base a un guió, no responen necessàriament a una definició prototípica. Tot això sense tenir en consideració que les condicions de gravació no són les mateixes que les que es tindrien en un estudi de gravació professional. En aquest escenari els resultats són molt diferents als que s'obtenen a l'escenari anterior i, per tant, es fa necessari un estudi més acurat. En aquest sentit es plantegen dues parametritzacions, una a nivell acústic i una altra a nivell lingüístic, en tant que la segona podria no resultar tan afectada pels elements que poden degradar la primera, tals com soroll o d'altres artefactes. Es proposen diferents esquemes de classificació de complexitat variable malgrat que, molt sovint, els sistemes més senzills proporcionen resultats bons. També es proposen diferents conjunts de paràmetres intentant aconseguir una selecció de dades tan petita com sigui possible que pugui realitzar un reconeixement afectiu automàtic de manera eficaç.

Els resultats obtinguts de l'anàlisi de les expressions espontànies posen de manifest la complexitat del problema plantejat i suposen valors inferiors als obtinguts a partir de corpus enregistrats en condicions ideals. Malgrat tot, els esquemes proposats aconsegueixen resultats que superen els publicats fins la data actual a estudis realitzats en condicions anàlogues i obren, per tant, la porta a investigacions futures dins d'aquest àmbit.

PARAULES CLAU: Anàlisi de la parla expressiva, reconeixement d'emocions, parla espontània, tecnologies de la parla, interacció persona-màquina

Resumen

Esta tesis aborda el reconocimiento automático de emociones espontáneas basado en el análisis de la señal de voz. Se realiza dentro del Grup de recerca de Tecnologies Mèdia (GTM) de *Enginyeria i Arquitectura La Salle*, teniendo su origen en un momento en el que existen abiertas varias líneas de investigación relacionadas con la síntesis afectiva pero ninguna relacionada con su análisis. La motivación es mejorar la interacción persona-máquina aportando un módulo de análisis en la entrada de los sistemas que permita, posteriormente, generar una respuesta adecuada a través de los módulos de síntesis en la salida de los mismos. El centro de atención se sitúa en la expresividad afectiva, intentando dotar de habilidades de inteligencia emocional a sistemas de inteligencia artificial con el objetivo de lograr que la interacción persona-máquina se asemeje, en la mayor medida posible, a la comunicación humana.

En primer lugar se realiza un análisis preliminar basado en locuciones grabadas en condiciones ideales. La expresividad vocal en este caso es actuada y las grabaciones responden a un guion previo que determina a priori la etiqueta que describe el contenido afectivo de las mismas. Si bien este no es el paradigma de la interacción en un entorno realista, este primer paso sirve para probar las primeras aproximaciones a la parametrización de los corpus, los métodos de selección de parámetros y su utilidad en la optimización de los procedimientos, así como la viabilidad de considerar el sistema de reconocimiento afectivo como un ejercicio de clasificación categórica. Asimismo, permite comparar los resultados obtenidos en este escenario con los que se obtengan posteriormente en el escenario realista. Si bien pudiera considerarse que la utilidad de un marco de trabajo como el aquí propuesto carece de interés más allá del mero ejercicio de comprobación citado, en esta tesis se propone un sistema basado en este planteamiento cuya finalidad es la validación automática de un corpus de voz expresiva destinado a síntesis, ya que en síntesis sí es necesario que el corpus esté grabado en condiciones óptimas puesto que será empleado para la generación de nuevas locuciones.

En segundo lugar la tesis profundiza en el análisis del corpus FAU Aibo, un corpus multilocutor de voz expresiva espontánea grabado en alemán a partir de interacciones naturales de un grupo de niños y niñas con un robot dotado de un micrófono. En este caso el planteamiento es completamente distinto al anterior partiendo de la definición del propio corpus, en el que las locuciones no responden a un guion previo y las etiquetas afectivas

se asignan posteriormente a partir de la evaluación subjetiva de las mismas. Asimismo, el grado de expresividad emocional de estas locuciones es inferior al de las grabadas por un actor o una actriz en tanto que son espontáneas y las emociones, dado que se generan de forma natural, no responden necesariamente a una definición prototípica. Todo ello sin considerar que las condiciones de grabación no son las mismas que las que se obtendrían en un estudio de grabación profesional. En este escenario los resultados son muy diferentes a los obtenidos en el escenario anterior por lo que se requiere un estudio más detallado. En este sentido se plantean dos parametrizaciones, una a nivel acústico y otra a nivel lingüístico, ya que la segunda podría no verse tan afectada por los elementos que pueden degradar la primera, tales como ruido u otros artefactos. Se proponen distintos sistemas de clasificación de complejidad variable a pesar de que, a menudo, los sistemas más sencillos producen resultados buenos. También se proponen distintas agrupaciones de parámetros intentando conseguir un conjunto de datos lo más pequeño posible que sea capaz de llevar a cabo un reconocimiento afectivo automático de forma eficaz.

Los resultados obtenidos en el análisis de las expresiones espontáneas ponen de manifiesto la complejidad del problema tratado y se corresponden con valores inferiores a los obtenidos a partir de corpus grabados en condiciones ideales. Sin embargo, los esquemas propuestos logran obtener resultados que superan los publicados hasta la fecha en estudios realizados en condiciones análogas y abren, por lo tanto, la puerta a investigaciones futuras en este ámbito.

PALABRAS CLAVE: Análisis del habla expresiva, reconocimiento de emociones, habla espontánea, tecnologías del habla, interacción persona-máquina

Summary

The topic of this thesis is the automatic spontaneous emotion recognition from the analysis of the speech signal. It is carried out in the Grup de recerca de Tecnologies Mèdia (GTM) of *Enginyeria i Arquitectura La Salle*, and was started when several research lines related to the synthesis of emotions were in progress but none of them were related to its analysis. The motivation is to improve human-machine interaction by developing an analysis module to be adapted as an input to the devices able to generate an appropriate answer at the output through their synthesis modules. The highlight is the expression of emotion, trying to give emotional intelligence skills to artificial intelligence systems. The main goal is to make human-machine interaction more similar to human communication.

First, we carried out a preliminary analysis of utterances recorded under ideal conditions. Vocal expression was, in this case, acted and the recordings followed a script which determined the descriptive label of their emotional content. Although this was not the paradigm of interaction in a realistic scenario, this previous step was useful in testing the first approaches to parameterisation of corpora, feature selection methods and their utility optimizing the proposed procedures, and in determining whether the consideration of the emotion recognition problem as a categorical classification exercise is viable. Moreover, it allowed the comparison of the results in this scenario with the results obtained in the realistic environment. This framework can be useful in some other contexts, additionally to this comparison utility. In this thesis we propose a system based on it with the goal of validating automatically an expressive speech corpus for synthesis. In the synthesis field, corpora must be recorded under optimal conditions to create new speech utterances.

Second, we present an analysis of the FAU Aibo corpus, a multi-speaker corpus of emotional spontaneous speech recorded in German from the interaction of a group of children with a robot which had an incorporated microphone. In this case, the approach was different because of the definition of the corpus. The recordings of the FAU Aibo corpus did not follow a script and the emotion category labels were assigned after a subjective evaluation process. Moreover, the emotional content of these recordings was lower than in those recorded by actors, because of their spontaneity; emotions were not prototypical because they were generated naturally, not following a script. Furthermore, recording conditions were not the same as in a professional recording studio. In this sce-

nario, results were very different to those obtained in the previous one. For this reason a more accurate analysis was required. In this sense we used two parameterisations, adding linguistic parameters to the acoustic information since the first one could be more robust to noise or some other artefacts than the second one. We considered several classifiers of different complexity although, frequently, simple systems get the better results. Moreover, we defined several sets of features trying to get a reduced set of data able to work efficiently in the automatic emotion recognition task.

Results related to the analysis of the spontaneous emotions confirmed the complexity of the problem and revealed lower values than those associated to the corpus recorded under ideal conditions. However, the schemas got better results than those published so far in works carried out under similar conditions. This opens a door to future research in this area.

KEYWORDS: Expressive speech analysis, emotion recognition, spontaneous speech, speech technologies, human-computer interaction

Agraïments

Aquesta tesi és fruit de l'esforç dut a terme durant molt temps. Al llarg dels anys han estat moltes les persones que en major o menor mesura han ajudat al fet que aquesta tesi sigui avui una realitat i fins a la contribució més petita, realitzada sempre de forma constructiva, ha suposat per a mi un important impuls per seguir endavant en aquest recorregut.

Vull agrair als meus pares el seu esforç per aconseguir que avui hagi pogut arribar al final d'aquest trajecte. Vull agrair-los de tot cor el seu suport i la seva comprensió al llarg de tot el temps invertit per dur a terme els meus estudis i, en aquesta darrera etapa, per a la realització d'aquesta tesi. Vull agrair-los l'amor i l'afecte incondicional que m'han donat al llarg de tota la meva vida i l'educació que m'han ofert. Perquè aquest també és el resultat del seu gran esforç i del seu treball dur i sense ells mai hagués pogut arribar fins a aquí.

A la meva família. Als que estan i als que van marxar sense poder veure aquest treball finalitzat i sempre em van fer costat.

A Ignasi Iriondo, director d'aquesta tesi doctoral, per les seves indicacions, el seu suport, la seva amistat i la seva insistència, per haver-se involucrat plenament en aquest treball i haver aconseguit que tirés endavant malgrat els esculls del camí.

A Joan Claudi Socoró, professor, company i amic, per tota l'ajuda, comprensió i suport oferts al llarg d'aquest temps en tantes i tan diverses situacions i circumstàncies i per haver pogut comptar amb ell en tot moment.

A Jose Antonio Morán per la motivació, per haver cregut en mi i per introduir-me en el món de la recerca. I juntament amb Elisa Martínez per la seva dedicació en les fases inicials d'aquest treball.

A tots i cadascun dels companys que m'han acompanyat durant la meva trajectòria en La Salle i, especialment, a aquells amb els quals vaig compartir la meva estada en el que va ser el Departament de Comunicacions i Teoria del Senyal, en el GPMM i, posteriorment, en el GTM. Entre ells hi ha persones amb les quals he compartit moltes hores d'estudi i treball a les quals els agraeixo la seva dedicació i companyerisme, com Jordi Adell, Francesc Alías, Rosa Maria Alsina, Àngel Calzada, Germán Cobo, Lluís Formiga,

David García, Xavier Gonzalvo, José Antonio Montero, Carlos Monzo, Xavier Sevillano i Alexandre Trilla. La trajectòria ha estat llarga i han estat moltes les persones a les quals vull agrair els moments compartits. No obstant això, resulta impossible citar-les a totes sense que em deixi noms en el tinter: Simó, David M., Carles, Marc, Mireia, David B., Gemma, Borja, Oriol, Pere, Lluís, Diego, Berta, Ester, ...

A Gabriel. A Alberto, Ana, Cristina, Gemma, Íñigo, Isa, Jaume, Laura, Lluís, Luisa, Marta i Mónica, per haver estat a prop sempre i, especialment, en aquests darrers anys. I també a tots els amics que no he citat però que no per això són menys importants per a mí, amb els quals he compartit tants moments i m'han mostrat el seu afecte en moltes ocasions i encara segueixen fent-ho.

Als organismes públics i privats que han finançat parcialment el treball de recerca presentat en aquesta tesi.

I a totes aquelles persones a les quals no he anomenat però amb les quals he estudiat, treballat o compartit vivències en tot aquest temps i han contribuït a la meva formació humana i professional.

A tots vosaltres, moltes gràcies.

Índex

Índex de figures

Índex de taules

Índex d'algorismes

Sigles, acrònims i símbols

AJAX *Asynchronous JavaScript And XML*

ANEW *Affective Norms for English Words*

ASR *Automatic Speech Recogniser*

ASSESS *Automatic Statistical Summary of Elementary Speech Structures*

CMOS *Comparison Mean Opinion Score*

CSV *Comma-Separated Values*

DAL *Dictionary of Affect in Language*

DMOS *Degradation Mean Opinion Score*

do1000 *Drop-off of spectral energy above 1000Hz*

EMODB *Berlin Database of Emotional Speech*

evMIC *entornos virtuales Multimodales, Inmersivos y Colaborativos*

F0 Freqüència fonamental

GMM *Gaussian-Mixture Models*

GNE *Glottal-to-Noise Excitation Ratio*

GPMM Grup de Recerca en Processament Multimodal

GTM Grup de recerca de Tecnologies Mèdia

Hamml *Hammamberg Index*

HMM *Hidden Markov Models*

HNM *Harmonic plus Noise Model*

HNR *Harmonic-to-Noise Ratio*

INREDIS *INterfaces de RElación entre el entorno y las personas con DIScapacidad*

JSP *Java Server Pages*

KNN *k-Nearest Neighbour*

LAICOM-UAB Laboratori d'Anàlisi Instrumental de la Comunicació de la Universitat Autònoma de Barcelona

LFPC *Log-Frequency Power Coefficients*

LOSO *Leave One Speaker Out*

MAP *Maximum A Posteriori*

MFCC *Mel Frequency Cepstral Coefficients*

MOS *Mean Opinion Score*

MRAA Mòdul de Reconeixement Afectiu Audiovisual

mRMR *Minimal-Redundancy Maximal-Relevance*

pe1000 *Relative Amount of Energy above 1000 Hz*

PF-STAR *Preparing Future Multisensorial Interaction Research*

POS *Part-Of-Speech*

SAM *Self-Assessment Manikin*

SAVE *Síntesis Audiovisual Expresiva*

SES *Spanish Emotional Speech*

SFM *Spectral Flatness Measure*

SL *Simple Logistic*

SVM *Support-Vector Machine*

TRUE *Testing platfoRm for mUltimedia Evaluation*

UAR *Unweighted Average Recall*

VoQ *Voice Quality*

WAR *Weighted Average Recall*

ZCR *Zero-Crossing Rate*

Capítol 1

Introducció

Aquesta tesi s'emmarca dins del programa de doctorat *Les Tecnologies de la Informació i la Comunicació i la seva gestió*. S'ha dut a terme dins del Grup de recerca de Tecnologies Mèdia (GTM) d'Enginyeria i Arquitectura La Salle, pertanyent a la Universitat Ramon Llull, sota la direcció del Dr. Ignasi Iriondo Sanz.

En aquest primer capítol es duu a terme una introducció a la tesi a través de les següents seccions:

- Presentació del context de la recerca realitzada (secció 1.1).
- Definició dels objectius de la tesi (secció 1.2).
- Organització de la tesi (secció 1.3).

1.1 Presentació del context de la recerca realitzada

Serveixi un famós escenari de la ciència ficció com a il·lustrador, i només com a tal, del context d'aquesta tesi. *¿Sueñan los androides con ovejas eléctricas? (Do androids dream of electric sheep?)* (Dick, 1968) és una novel·la curta que va donar origen al clàssic del cinema de ciència ficció *Blade Runner*, dirigit el 1982 per Ridley Scott. A l'escena inicial de la pel·lícula, ambientada en una futurista ciutat de Los Angeles i després de presentar als androides[1] del model Nexus 6 com "més humans que els humans", un policia interroga a un subjecte amb l'ajuda de la màquina de Voigt-Kampff per detectar si és un humà o un androide i, en aquest cas, retirar-ho. Aquesta màquina mesura les respostes emocionals dels individus realitzant un test d'empatia (Prieto, 2008), qualitat que no poseeixen els androides i que, malgrat la seva aparent similitud amb els humans, els diferència d'ells. Aquest androide no supera la prova. Però, existeix algun androide capaç de superar-la?

[1] *Androide* és el terme emprat a la novel·la. A la seva adaptació al cinema el terme emprat és *replicant*.

Són l'empatia i les reaccions emocionals, doncs, els clars indicadors de què un ésser humà realment ho és?

No és aquest l'únic cas en el que la literatura i el cinema recorren a les emocions per plantejar la inquietant pregunta de si aquesta persona que hi ha davant nostre i sembla tan humana realment ho és o no. Així, per exemple, la novel·la de Jack Finney, *Els lladres de cossos (The body snatchers)*, escrita el 1955 i adaptada al cinema en múltiples ocasions (la primera el 1956 per Don Siegel i recentment en la quarta i, fins avui, última versió de 2007 per Oliver Hirschbiegel) planteja una invasió extraterrestre en la qual uns éssers alienígenes suplanten a persones normals sense que aparentment es percebi cap canvi. Tan sols en aparença doncs l'absència d'emocions és el tret característic dels que han estat suplantats. En l'adaptació al cinema de 2007, titulada *Invasió (The invasion)*, un personatge aconsella a la protagonista, mentre camina per un carrer en el que corre el risc de ser descoberta: "No mostri cap emoció, així no podran descobrir-la." Un tema similar al que es planteja a la pel·lícula *El poble dels maleïts (Village of the damned)*, adaptació de la novel·la publicada el 1957 *The Midwich Cuckoos*, de John Wyndham, dirigida el 1960 per Wolf Rilla i en una segona versió de 1995 per John Carpenter.

No ha d'estranyar que la vinculació de l'expressivitat emocional al comportament humà sigui empleada de forma recurrent en aquestes obres, així com que s'empri l'absència d'emocions en una persona com a element per generar terror entre el públic. La comunicació humana consta d'una important part afectiva que forma part del missatge que es transmet, si bé de forma implícita, i mostra un caràcter empàtic que es desenvolupa des d'edat primerenca i que es fonamenta en la interpretació dels senyals transmesos de forma inconscient i que no sempre són verbals (Goleman, 1995). L'empatia suposa un mínim reconeixement i comprensió de l'estat emocional d'una persona per part del seu interlocutor (Decety i Jackson, 2004), i la seva absència dificulta el procés comunicatiu.

Malgrat això, la comunicació persona-màquina, propiciada cada vegada més pels avenços tecnològics i l'accés majoritari de la població a les noves tecnologies, sovint no disposa d'aquesta component afectiva. Així, sol associar-se la comunicació sense emocions a la comunicació amb un robot (Iriondo, 2008), exemple popular per il·lustrar un discurs no emocional. Un dels objectius recents de la comunicació persona-màquina és la millora de l'experiència d'usuari[2], intentant aconseguir que aquesta interacció sigui més similar a la comunicació entre humans. La inclusió de mòduls de reconeixement de veu va ser un dels punts clau en el passat recent per afegir capacitats perceptives als dispositius multimèdia, cosa que va millorar notablement les seves interfícies d'usuari (Canny, 2006).

No obstant això, l'anàlisi d'estats afectius a través de l'estudi del canal implícit de la comunicació[3] pot millorar les aplicacions que requereixen d'una comunicació amb l'usuari fent-les més usables i amigables. Això és així perquè, en general, afegir capacitats

[2]Millora que suposa una accés més còmode a la tecnologia i que, en conseqüència, pot ajudar també a evitar que certs sectors socials quedin exclosos de l'accés a la mateixa.

[3]És a dir, l'anàlisi de com alguna cosa ha estat dita i no només de què és el que s'ha dit.

d'intel·ligència emocional a la intel·ligència artificial dels dispositius i aplicacions pot fer possible que la interacció persona-màquina sigui més similar a la comunicació humana (Picard *et al.*, 2001). Existeix un ampli ventall de contextos en els quals l'anàlisi de la parla i de l'emoció en l'entrada dels sistemes (i també la síntesi de veu emocionada en la sortida dels mateixos) pot ser aplicat, incloent entre ells la generació automàtica de contingut audiovisual, reunions virtuals, sistemes de diàleg automàtics, sistemes de tutoria en àmbits docents, entreteniment o "jocs seriosos"[4].

La figura 1.1 mostra un esquema genèric d'interacció persona-màquina a través de la veu amb reconeixement i síntesi d'emocions. No obstant això, aquest esquema podria ser fàcilment ampliat per representar un sistema que inclogui la modalitat visual i que introdueixi l'anàlisi de l'expressivitat facial, i encara podria incloure altres modalitats per gestionar la interacció. Malgrat això, aquesta tesi focalitza en l'anàlisi del senyal de veu i, per tant, aquest esquema incideix en la modalitat de la parla. Noti's que el sistema automàtic rep la transcripció del missatge dictat per l'usuari i també un indicador de l'estat afectiu del mateix. A partir d'ambdues informacions el sistema pot generar una resposta a tots dos nivells, tant a nivell de missatge (originalment un missatge de text que es convertirà en veu en el mòdul corresponent) com a nivell expressiu emocional (que modularà el senyal de veu generat per acolorir-ho afectivament).

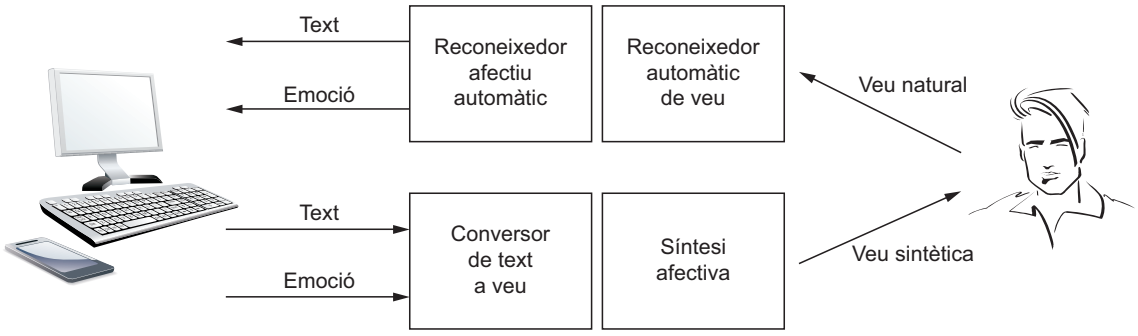


Figura 1.1: Esquema d'interacció persona-màquina a través de la veu amb reconeixement i síntesi d'emocions.

En el moment previ a l'inici de la present tesi doctoral, al GTM s'havien iniciat quatre tesis doctorals en l'àmbit de la síntesi de veu expressiva (la branca inferior de l'esquema de la figura 1.1), les desenvolupades per Iriondo (2008), Gonzalvo (2010), Monzo (2010) i Formiga (2010). Neix, llavors, la necessitat d'avançar en el camp de l'anàlisi afectiva per poder completar l'esquema d'interacció i aprofitar aquests avanços, en la mesura del possible, en els estudis de síntesi ja iniciats. Així doncs, s'inicia una recerca en l'estat de la qüestió que mostra els estudis realitzats per diferents investigadors i posa de manifest l'existència d'estudis basats en veus de naturalesa diversa (enregistraments de veu actuada i espontània, condicions de laboratori i naturals, diversos àmbits d'estudi, etc.), així com la diversitat de resultats obtinguts en funció dels plantejaments adoptats. Després de l'anàlisi inicial dels corpus gravats per actors, s'escull l'àmbit de les emocions

[4]De l'anglès, *Serious games*.

espontànies donada la seva major adequació als processos d'interacció reals, sent aquest, a més, un àmbit menys tractat per la comunitat científica donada la seva complexitat i els resultats d'identificació habitualment baixos que s'obtenen. Tot això serà detallat en els capítols següents.

1.2 Definició dels objectius de la tesi

L'interès d'aquesta tesi doctoral se centra en l'anàlisi de l'expressivitat afectiva dins del procés comunicatiu i, en concret, la que es refereix al senyal de veu. La veu, en si mateixa, transmet informació paralingüística referida a l'emoció del parlant i, en un sentit més ampli, al seu estat d'ànim, les seves intencions o d'altres aspectes que el relacionen amb l'entorn i/o amb el seu interlocutor[5]. Per aquest motiu, i donada l'existència de les línies de recerca en l'àmbit de la síntesi ja obertes en el GTM en el moment del plantejament de la present tesi, la motivació d'aquest treball és l'anàlisi dels factors que proporcionen a la parla els seus trets afectius.

Exposada la motivació de la tesi, es formulen dues preguntes que donen motiu a la recerca duta a terme:

1. Quina informació oral és la més rellevant per dur a terme la identificació de les emocions?
2. Com ha de transferir-se aquesta informació a un sistema automàtic i quins han de ser els seus procediments per aconseguir identificar les emocions en un entorn d'interacció natural?

Per donar resposta a aquestes preguntes es plantegen les següents hipòtesis de treball:

1. Atès que a nivell oral un parlant transmet informació acústica i lingüística, l'anàlisi d'ambdues permetrà aconseguir una millor identificació afectiva que l'anàlisi individual de cadascuna de les mateixes.
2. En un entorn d'interacció natural les locucions dels parlants són espontànies i no sempre es produeixen en condicions òptimes per al seu enregistrament. A més, una aplicació de reconeixement ha d'admetre diversos locutors. Això dificultarà la seva anàlisi a nivell acústic. Mentre que la informació acústica podria no ser suficient

[5] A més de la informació paralingüística, amb les funcions referides en el text, la veu transmet informació lingüística que permet, per exemple, distingir entre una afirmació o una pregunta, i informació extralingüística, relativa a característiques del locutor tals com la seva edat, el seu sexe o el seu estatus socioeconòmic (Escudero, 2003). Aquests dos últims tipus d'informació, no obstant això, no es relacionen directament amb l'estat afectiu del locutor. Vegeu la secció 2.3.3 per a més informació.

per a tasques de reconeixement afectiu en un escenari realista (Batliner *et al.*, 2003) i considerant que la informació lingüística podria veure's afectada en menor mesura per les condicions d'enregistrament[6], l'anàlisi d'aquesta última podria contribuir a millorar els resultats obtinguts a partir de la primera.

3. El reconeixement afectiu automàtic pot considerar-se un exercici de classificació categòrica en el qual les locucions han de rebre una etiqueta que identifiqui l'emoció que representen. Algorismes de classificació de l'estat de la qüestió poden ser útils per a dur a terme aquesta tasca, de forma individual o combinada mitjançant tècniques de fusió.

4. La parametrització de les locucions com a pas previ a la seva classificació donarà lloc a un gran volum de dades. Tècniques de selecció de paràmetres reduiran aquest volum per optimitzar els resultats dels algorismes de classificació.

A partir de les hipòtesis de treball anteriors es defineix l'objectiu general d'aquesta tesi:

- Determinar els esquemes de classificació més adequats i els paràmetres necessaris per classificar fragments de veu espontània segons el seu contingut afectiu.

Per aconseguir l'objectiu general es defineix una sèrie d'objectius específics:

1. Realitzar un exercici previ de reconeixement afectiu automàtic en corpus gravats en condicions de laboratori per observar els resultats obtinguts i assentar la base per als experiments posteriors en corpus de veu espontània. Aquest exercici servirà per comparar els procediments emprats en tots dos casos així com els resultats obtinguts i posar de manifest les problemàtiques específiques.

2. Parametritzar a nivell acústic i lingüístic el corpus de veu espontània, tot creant un conjunt de dades complet que serà dividit en subconjunts de paràmetres rellevants mitjançant tècniques automàtiques i manuals. Això permetrà optimitzar els algorismes de classificació determinant, al mateix temps, quins són els paràmetres més importants per dur a terme la tasca de classificació de forma eficaç.

3. Definir un entorn de treball que permeti avaluar i comparar els mètodes proposats en un escenari multilocutor d'anàlisi d'emocions espontànies. Aquest entorn de treball ha de considerar les característiques pròpies del corpus per definir unes mètriques adequades i ha d'oferir també uns resultats que siguin comparables amb els treballs realitzats per altres investigadors en aquest àmbit.

[6]La informació lingüística en si no es veuria afectada però, si el sistema disposa d'un mòdul de reconeixement de veu, unes condicions no òptimes d'enregistrament poden influir negativament en el procés d'extracció del contingut lingüístic de les locucions. Això sí afectaria, indirectament, al processament d'aquesta informació.

4. Seleccionar els esquemes de classificació més adequats, ja siguin classificadors individuals o estructures de classificació complexes, així com els paràmetres més convenients per a les mateixes. Això inclou la consideració dels esquemes de fusió més eficaços i eficients, tant a nivell de paràmetres com a nivell de decisió, per optimitzar els resultats.

1.3 Organització de la tesi

Aquesta tesi consta de set capítols, els quals es descriuen a continuació.

El capítol 1 conté la introducció de la tesi, considerant el context de la recerca realitzada, els objectius plantejats i el detall dels capítols d'aquesta tesi i la seva organització.

El capítol 2, en primer lloc, fa un repàs de les teories més acceptades sobre les emocions. El concepte d'emoció és complicat de definir per la qual cosa es recullen els plantejaments que diferents perspectives realitzen, al llarg de la història, al voltant d'aquest terme. Tanmateix es presenten tres mètodes per descriure les emocions, incloent els llistats d'emocions, subjectes a la definició creada pels autors de cadascun d'ells, els models *circumplex* i els espais multidimensionals. Cadascun d'aquests mètodes presenta les seves pròpies característiques i tots ells suposen visions diferents de les emocions en si, la qual cosa torna a posar de manifest l'ambigüitat sovint vinculada al concepte d'emoció. En segon lloc, aquest capítol fa un recorregut per diferents mètodes creats per a l'etiquetatge de corpus afectius donat l'estret vincle que s'estableix entre aquests mètodes i la definició que s'adopta del concepte emoció. Aquests mètodes permeten assignar, de forma subjectiva, una etiqueta afectiva a cada element del corpus i inclouen, entre els més estesos, els etiquetatges categòrics i els dimensionals. Finalment, es tracten temes relacionats amb l'expressió i la percepció de les emocions, tals com les reaccions fisiològiques associades, els elements multimodals de la seva expressió i, finalment, detalls relatius a l'expressivitat afectiva a través de la parla.

El capítol 3 descriu les condicions generals per a la creació d'un corpus de veu emocionada. El contingut del capítol incideix en els diferents mètodes per a la generació de la parla emocional donada la importància que tenen els corpus en estudis com els presentats en aquesta tesi, ja que són la base fonamental dels mateixos. De la seva elecció depèn en gran mesura l'estratègia a seguir i l'escollir un corpus o un altre pot suposar una important dificultat afegida, depenent de les condicions del seu enregistrament. Així mateix, l'expressió vocal emprada[7] defineix si l'estudi de reconeixement se centra en condicions pròpies d'un entorn realista o, per contra, s'emmarca en un context propi de laboratori. El capítol detalla també el procés d'extracció de paràmetres perquè aquests corpus puguin ser processats posteriorment per diferents algorismes d'aprenentatge. En aquesta tesi s'empren dos tipus de paràmetres: acústics i lingüístics. Finalment es descriuen els

[7] En aquesta tesi es consideren quatre tipus d'expressió vocal: natural, induïda, estimulada o actuada, tal com assenyala Schröder (2004)

tres corpus que s'analitzaran al llarg d'aquesta tesi tot detallant la parametrització dels mateixos.

Al capítol 4 es recullen i analitzen alguns dels treballs més rellevants realitzats en el camp del reconeixement afectiu basat en el senyal de veu, diferenciant els treballs realitzats al voltant de corpus de veu actuada i al voltant de corpus de veu espontània. S'inclou també l'anàlisi d'alguns estudis importants realitzats a partir d'altres modalitats així com la seva anàlisi conjunta emprant tècniques de fusió. Aquesta anàlisi es planteja com un recorregut des dels estudis de les primeres aproximacions fins a arribar a les tendències més recents, incloent les aplicacions d'aquests treballs en àmbits que són, sovint, molt diferents entre si. Per poder assentar la base d'experimentació, el capítol recull també la definició d'aquells conceptes i mètodes que posteriorment seran emprats. Així, es defineixen les principals modalitats que molts autors analitzen en els treballs més rellevants, s'introdueixen les tècniques de selecció de paràmetres que s'empren amb major freqüència en aquest àmbit, s'expliquen les mètriques emprades per comparar els resultats entre diferents treballs i quins són els mètodes d'avaluació dels procediments de reconeixement automàtic que es proposen.

El capítol 5 centra la seva atenció en els experiments que es duen a terme sobre corpus gravats per actors sota condicions ideals pròpies de laboratori. Es realitzen dos plantejaments diferenciats. El primer d'ells pretén maximitzar el nombre de coincidències entre l'estil afectiu de cada element del corpus i l'etiqueta assignada automàticament pel reconeixedor. El segon pretén maximitzar el nombre de coincidències per a aquelles locucions que també serien correctament identificades de forma subjectiva. Aquest segon plantejament suposa un mapatge de criteris subjectius en tant que no busca obtenir el valor màxim de classificació sinó realitzar la mateixa classificació que un avaluador humà realitzaria. Tots dos plantejaments es corresponen amb una fase preliminar de l'estudi del tema principal d'aquesta tesi, el qual s'aborda en el següent capítol.

El capítol 6 recull els experiments realitzats sobre un corpus d'expressió vocal espontània. El plantejament d'aquests experiments és molt diferent al dels mostrats al capítol 5 donades les característiques d'aquest tipus d'expressió vocal. El capítol s'inicia amb la descripció de la metodologia emprada al llarg de tota l'experimentació, detallant els experiments de reconeixement afectiu automàtic que es duen a terme agrupant-los en tres blocs, segons el seu objecte d'anàlisi: l'estudi de paràmetres acústics i lingüístics de forma independent d'una banda i de forma conjunta mitjançant la seva fusió per una altra. Els experiments de classificació es realitzen des de diferents perspectives, emprant esquemes de classificació de forma individual i també agrupats en estructures més complexes creant estructures jeràrquiques, en cascada o de fusió a nivell de decisió. Ambdues modalitats, l'acústica i la lingüística, s'estudien per separat però també de forma conjunta mitjançant esquemes de fusió a nivell de decisió i a nivell de paràmetres. Així mateix, i atès que el volum de dades és molt elevat, es realitza un procés de selecció de paràmetres tant de forma automàtica com de forma manual. L'objectiu d'aquesta reducció és doble: d'una banda optimitzar el procés de classificació i, per una altra, determinar els paràmetres més rellevants per al procés de classificació. Finalment es realitza una comparació global entre

els resultats obtinguts i els observats en estudis relacionats en l'estat de la qüestió.

Finalment, el capítol 7 exposa les principals conclusions i les línies futures de treball que s'obren i donen continuïtat a les aportacions d'aquesta tesi.

Capítol 2

Fonaments

En aquest capítol es fa un recorregut per les principals teories de les emocions per poder definir el concepte d'emoció i poder arribar a donar una descripció de les mateixes sobre la base de diferents models, incloent els llistats d'emocions, els models *circumplex* i els espais multidimensionals. A continuació s'expliquen les principals formes de representar i avaluar les emocions, la qual cosa serà d'utilitat posteriorment per al correcte etiquetatge dels corpus de veu que s'empraran en els experiments de reconeixement emocional. Finalment, es descriuen l'expressió i la percepció de les emocions de forma general i, en concret, a través de la parla, relacionant els paràmetres propis d'aquesta modalitat amb la modalitat musical, a la qual va estretament lligada en tant que ambdues incideixen, directament, en el sentit auditiu del receptor.

2.1 Teoria de les emocions

Descriure una emoció no és una tasca senzilla. Des de les teories aristotèliques fins a les propostes contemporànies, molts pensadors, filòsofs i científics han intentat dotar al terme *emoció* d'una definició que li proporcioni un significat complet, que realitzi una descripció, que proposi una classificació de les seves diferents variants i que li atorgui una funcionalitat per als éssers vius en general i per a l'ésser humà en particular. No obstant això, i malgrat ser un dels pilars bàsics per a la interacció, són moltes i variades les teories que existeixen sobre aquest tema. Sense pretendre abastar-les totes, aquesta secció descriu les principals definicions que seran d'utilitat per al desenvolupament de la present tesi.

2.1.1 El concepte d'emoció

En els processos d'interacció humana poden distingir-se dos canals diferenciats (Cowie *et al.*, 2001): el primer és l'encarregat de la transmissió de missatges explícits men-

tre que el segon s'encarrega de la transmissió de missatges implícits. Mentre que els primers fan referència al contingut del propi missatge, els segons es refereixen als locutors en si. S'han invertit més esforços, tant a nivell lingüístic com a nivell tecnològic, estudiant el primer canal en detriment del segon si bé actualment la tendència s'està invertint. El reconeixement emocional s'emmarca en l'anàlisi del canal implícit[1].

La pròpia definició de la paraula *emoció* és un tema complex. Goleman (1995) es refereix a la definició "agitació o pertorbació de la ment; sentiment; passió; qualsevol estat mental vehement o agitat". Per a aquest autor, el terme emoció "fa referència a un sentiment i als pensaments, als estats biològics, als estats psicològics i al tipus de tendències a l'acció que ho caracteritzen. Existeixen centenars d'emocions i moltíssimes mescles, variacions, mutacions i matisos diferents entre totes elles. En realitat existeixen més subtileses en l'emoció que paraules per descriure-les". És important assenyalar el concepte de "tendència a l'acció" vinculat a la pròpia arrel etimològica de la paraula "emoció": del llatí *movere* i el prefix *-e*, indicant *moviment cap a*. Segons el Diccionari de la Llengua Catalana de l'Institut d'Estudis Catalans, la primera accepció d'aquesta paraula és "Reacció afectiva, en general intensa, provocada per un factor extern o pel pensament, que es manifesta per una commoció orgànica més o menys visible". Cowie i Cornelius (2003) fan una profunda anàlisi dels termes i conceptes que es relacionen amb l'emoció i la parla. Defineix, entre d'altres, els següents conceptes emocionals:

- Emoció plena (Scherer, 1999): també referenciada com emoció primària (Plutchik, 2001) o emoció bàsica (Ekman, 1999). Descriu l'emoció en la seva forma més intensa.
- Emoció subjacent: denota la coloració emocional que està present en tots els estats mentals. La seva descripció és més complexa que la de les emocions plenes encara que les subjacents estan presents en la comunicació humana en major mesura que les plenes, les quals es donen en poques ocasions.
- Estats emocionals: recullen totes les emocions, des de les subjacents fins a les plenes, incloent tots els estats intermedis significatius en la comunicació humana.

Al mateix temps, Cowie i Cornelius (2003) introdueixen el concepte d'estats relacionats amb l'emoció en els quals les persones no senten pròpiament una emoció encara que sí presenten certs aspectes propis de les emocions, tals com a humor, excitació, certa actitud, etc.

De la mateixa forma que els conceptes anteriors estan relacionats amb la intensitat de l'emoció, diferents conceptes emocionals es poden associar a una escala temporal concreta. En aquest sentit, mentre que les expressions, les actituds o els estats emocionals (com per exemple estar content o trist) acostumen a durar alguns segons o diversos minuts, un estat d'humor (com per exemple irascible o malenconiós) pot durar hores o

[1] La descripció d'aquests canals i la seva funció en el procés comunicatiu es detallen en la secció 2.3

estendre's fins a alguns mesos, un desordre emocional (com per exemple un estat depressiu) es pot perllongar de setmanes a anys i un tret emocional (com taciturn o nerviós) pot durar anys o definir-se durant tota la vida d'una persona com el seu propi caràcter (Cowie *et al.*, 2001).

2.1.2 Teories sobre les emocions plenes

Scherer (1986) va descriure l'emoció com "la interfície de l'organisme cap al món exterior" i va destacar tres funcions principals:

- Reflectir l'avaluació de la rellevància i el significat de l'estímul particular en termes de les necessitats de l'organisme, plans i preferències. És a dir, valorar la situació a la qual s'enfronta un individu.
- Preparar fisiològicament i psicològicament l'organisme per a una acció apropiada, la qual cosa es reflecteix en canvis fisiològics i proporciona a l'individu una tendència a l'acció.
- Comunicar l'estat de l'organisme i les intencions de comportament cap a altres éssers propers a través de l'expressivitat facial, corporal i oral.

Les teories psicològiques contemporànies sobre l'emoció s'emmarquen en diferents perspectives, de les quals podem destacar quatre com a bàsiques. Aquestes perspectives comprenen des de les primeres aproximacions realitzades per Charles Darwin fins a les teories de finals del segle XX i tracten sobre com definir, estudiar i explicar les emocions. Aquestes quatre perspectives, segons la descripció donada per Cornelius (2000), són: Darwiniana, Jamesiana, Cognitiva i Constructivista Social. Encara que cadascuna d'elles estableix les seves pròpies bases referents a com construir teories sobre l'emoció, la seva naturalesa i sobre com dirigir la recerca en aquest àmbit, hi ha coincidències destacables entre elles —sobretot entre la Darwiniana i la Jamesiana— com pot comprovar-se a continuació.

2.1.2.1 La Perspectiva Darwiniana

Des de la perspectiva Darwiniana les emocions són fenòmens desenvolupats com a funcions importants de supervivència i han estat seleccionades com a tal per solucionar certs problemes als quals l'espècie humana ha hagut de fer front. Per aquest motiu els comportaments emocionals són similars en tots els éssers humans i s'assemblen, fins i tot, als d'aquells mamífers amb els quals l'home ha compartit un passat al llarg de l'evolució. Per a Charles Darwin, l'expressió de les emocions està inscrita en la memòria genètica dels individus i per això es dóna de forma innata des de la infància. Els inicis d'aquesta perspectiva es remunten a l'any 1872, quan Darwin publica *La expresión de las emociones en*

los animales y en el hombre (The expression of emotion in man and animals) (Darwin, 1872)[2]. Les seves idees han estat molt influents i el seu llegat en l'estudi de l'emoció en la psicologia i la biologia es basa en:

- Aplicar les teories de l'evolució per selecció natural amb la finalitat d'entendre les expressions emocionals i les pròpies emocions en si.
- Remarcar que les expressions emocionals han d'entendre's en termes de les seves funcions i, per tant, com un valor de supervivència.

Treballs posteriors remarquen el factor d'herència genètica en els comportaments emocionals com, per exemple, el de Peleg *et al.* (2006) en el qual es comprova com dos individus d'una mateixa família manifesten emocions a través de les mateixes expressions facials malgrat que un d'ells és invident. Atès que no ha pogut aprendre l'expressió facial de la seva família a partir de l'observació de la mateixa es corrobora l'herència genètica abans comentada. Matsumoto i Willingham (2009) obtenen resultats equiparables en realitzar un estudi similar observant atletes invidents i les seves expressions facials.

2.1.2.2 La Perspectiva Jamesiana

La perspectiva Jamesiana segueix el postulat de William James que defineix les emocions com a canvis corporals produïts immediatament després de la percepció d'un estímul, segons publicava el 1884 en l'article *What is an emotion?* (James, 1884)[3]. Així doncs, James afirma que els canvis corporals es produeixen tan aviat com es percep una excitació i l'emoció és el sentiment que s'experimenta quan aquests canvis es manifesten. No és possible l'existència d'emocions sense que es produeixin aquests canvis i aquests sempre són previs a l'emoció en si. La coincidència d'aquesta perspectiva amb la Darwiniana es produeix en considerar les emocions com adaptacions a l'entorn basades en la supervivència de l'individu, però difereixen en el fet de què Darwin se centra en les seves manifestacions mentre que James ho fa en la seva pròpia naturalesa. Així és important remarcar que, per a James, el cos respon primer i l'experiència adquirida pels humans constitueix l'emoció, que és la resposta a aquests canvis. A William James és deguda la popular frase "ens sentim tristos perquè plorem, enfadats perquè copegem, espantats perquè tremolem, i no plorem, copegem ni tremolem perquè estiguem tristos, enfadats o espantats"[4]. No obstant això, les recerques realitzades en aquest camp encara no han desvetllat com els canvis corporals serien iniciats per la percepció dels estímuls de l'entorn.

[2]Aquesta obra així com d'altres de Charles Darwin poden ser consultades a la següent pàgina web de lliure accés: http://darwin-online.org.uk

[3]Aquesta obra pot ser consultada a la pàgina web http://psychclassics.yorku.ca

[4]La cita original, en anglès, és: *"(...) the more rational statement is that we feel sorry because we cry, angry because we strike, afraid because we tremble, and not that we cry, strike, or tremble, because we are sorry, angry, or fearful (...)"*

2.1.2.3 La Perspectiva Cognitiva

La perspectiva cognitiva és la més dominant de les quatre gràcies a que aquesta perspectiva ha estat minuciosament incorporada dins de les altres tres. La perspectiva cognitiva moderna es basa en els estudis de les emocions realitzats per Magda Arnold (1903–2002), tot i que els orígens de la mateixa es remunten a més enllà dels filòsofs hel·lenístics. Aquesta perspectiva centra el seu eix indicant que l'emoció i el pensament són inseparables. Concretament, indica que totes les emocions són enjudiciades mitjançant una avaluació dels estímuls que les desencadenen. El procés d'avaluació consisteix en determinar què estímuls de l'entorn són considerats per l'individu com a bons o dolents. Aquesta avaluació comporta una tendència a l'acció que desemboca, finalment, en una resposta emocional.

Aquesta perspectiva discrepa de la Jamesiana en què aquella no especifica mitjançant quin mecanisme es desencadena el canvi corporal al que fa referència, que era la gran crítica d'Arnold a James, i la complementa indicant que aquest desencadenant és el procés avaluador que introdueix. Si James no podia concebre una emoció sense un cos, Arnold no ho pot fer sense una avaluació. Cada emoció es considera subjecta a un patró diferent d'avaluació i depèn de cadascun dels individus considerant variables tals com les característiques particulars de la persona, el seu aprenentatge, el seu temperament, la seva personalitat, el seu estat psicològic i les condicions específiques de la situació en la qual es troba. D'aquesta manera, el procés d'avaluació s'encarrega d'informar a l'organisme de les característiques de l'entorn i proporciona la manera d'actuar enfront d'elles.

2.1.2.4 La Perspectiva Constructivista Social

Aquesta perspectiva és la més recent i també la més controvertida. Refusa assumir que les emocions són fruit d'una adaptació al mitjà i afirma que són productes definits per les regles socials de l'entorn. Són, per tant, construccions socials i és la cultura la que juga un paper crucial en la seva estipulació. A diferència de la perspectiva cognitiva en la qual el procés d'avaluació, que depèn de factors personals, és el desencadenant de l'emoció, per a la perspectiva constructivista és la cultura, en forma de regles socials, la que les defineix. Segons Averill (1980) citat per Cornelius (2000), les emocions no es poden considerar romanents d'un passat psicogenètic així com tampoc poden ser explicades de forma estrictament psicològica. Atès que les emocions són considerades per aquesta teoria com a construccions socials només poden ser enteses a partir d'una anàlisi social.

2.1.3 La descripció de les emocions

Existeixen diverses formes de definir i representar les diferents emocions. Podria semblar que la més senzilla és la d'un model categòric consistent a crear un conjunt limitat d'etiquetes que les descriuen creant un llistat d'emocions. No obstant això, determinar

el nombre d'emocions d'aquest llistat així com determinar quines són les fonamentals i quines no no és una tasca tan simple com pot semblar a ull nu. Això dóna motiu a la creació d'altres representacions alternatives com podrien ser els models *circumplex* o les representacions multidimensionals. Tal com ja s'ha comprovat en l'apartat anterior, no existeix un consens per determinar quina és la definició d'*emoció* més adequada, i l'elecció del model de representació més convenient dependrà tant de l'aplicació en la qual es requereixi l'anàlisi emocional com de la teoria que es prengui com a base.

2.1.3.1 Llistats d'emocions bàsiques

El concepte d'emocions bàsiques (també conegudes com a emocions plenes o primàries) és emprat per moltes de les teories sobre l'emoció, especialment aquelles que segueixen les tradicions Darwiniana i Jamesiana. Es consideren bàsiques per estar directament lligades a la supervivència dels individus sobre la base de patrons evolutius i, també, perquè a partir d'elles es generen totes les altres mitjançant variacions o combinacions de les mateixes. Al mateix temps, es poden definir les emocions secundàries com a emocions més complexes que es formarien com a combinació de les bàsiques, d'igual manera que els colors primaris es barregen en una roda de color per crear d'altres colors. Ara bé, definir quines són aquestes emocions bàsiques no és una tasca fàcil, més quan ni tan sols els investigadors coincideixen en afirmar la seva existència real. No hi ha un criteri únic per definir quines emocions formen part d'aquest conjunt bàsic. Les quatre emocions bàsiques més acceptades generalment són: alegria, tristesa, enuig i por, atès que es considera que estan directament lligades a processos biològics. La major part de les teories emocionals coincideixen en què hi ha un nombre inferior a deu emocions bàsiques, a pesar que estudis més recents (Cowie i Cornelius, 2003) defineixen un conjunt d'entre deu i vint emocions. Altres estudis, com els de Whissell (1989) o Plutchik (1980), amplien aquests conjunts incloent emocions bàsiques i secundàries arribant a crear llistats que inclouen de 107 a 142 etiquetes emocionals, respectivament. La taula 2.1 conté un llistat d'aquestes emocions segons diversos autors i la seva traducció a l'espanyol i al català. Apareixen remarcades les quatre emocions bàsiques anteriorment citades, les quals són comunes en els estudis referenciats. Cal destacar el terme *The Big Six* utilitzat en Cornelius (2000), una ampliació de les quatre emocions anteriors a un total de sis definint un conjunt format per: felicitat (*happiness*), tristesa (*sadness*), por (*fear*), fàstic (*disgust*), enuig (*anger*) i sorpresa (*surprise*).

Un dels principals problemes en la recerca intercultural és la traducció precisa dels termes emocionals. Les connotacions particulars de cada terme impliquen que no hi hagi una solució satisfactòria per a aquest problema. La traducció dels termes de la taula 2.1 està realitzada per Iriondo (2008) amb l'ajuda de l'estudi presentat per Scherer (1988), que recull una llista de descriptors de l'emoció en cinc llengües indoeuropees que és el resultat de l'activitat de recerca d'un equip de psicòlegs de diferents països. La traducció mostrada en l'última columna ve a representar, sota una única etiqueta emocional, la fusió dels conceptes emocionals dels sis treballs analitzats per Cowie i Cornelius (2003). Com

Taula 2.1: Emocions bàsiques segons el llistat de Cowie i Cornelius (2003) (en anglès) i la seva traducció a l'espanyol, reproduïdes d'(Iriondo, 2008). Inclou també la traducció al català.

Lazarus (Lazarus, 1999)	Ekman (Ekman, 1999)	Buck (Buck, 1999)	Lewis-Haviland (Lewis i Haviland, 1993)	Banse-Scherer (Banse i Scherer, 1996)	Cowie et al. (Cowie *et al.*, 1999)	Traducció espanyola (Iriondo, 2008)	Traducció al català
Anger	Anger	Anger	Anger/hostility	Rage/hot anger Irritation/cold anger	Angry	Enfado/enfadado	Enuig/enfadat
Fright	Fear	Fear	Fear	Fear/terror	Afraid	Miedo/atemorizado	Por/atemorit
Sadness	Sadness/distress	Sadness	Sadness	Sadness/dejection Grief/desperation	Sad	Tristeza/triste	Tristesa/trist
Anxiety		Anxiety	Anxiety	Worry/anxiety	Worried	Inquietud/preocupado	Inquietut/preocupado
Happiness	Sensory pleasure	Happiness	Happiness	Happiness Elation (joy)	Happy	Alegría Felicidad/feliz	Alegria Felicitat/feliç
	Amusement		Humour		Amused	Diversión/divertido	Diversió/divertit
	Satisfaction				Pleased	Satisfacción/satisfecho	Satisfacció/satisfet
	Contentment				Content	Contento	Content
		Interested			Interested	Interesado	Interessat
		Curious				Curioso	Curiós
		Surprised				Sorprendido	Sorprès
	Excitement				Excited	Ilusión/excitado	Il·lusió/excitat
		Bored		Boredom/indifference	Bored	Aburrido	Avorrit
					Relaxed	Relajado	Relaxat
		Burnt out				Quemado/estresado	Cremat/estressat
Disgust	Disgust	Disgust	Disgust	Disgust		Asco	Fàstic
	Contempt	Scorn		Contempt/scorn		Desprecio/desdén	Mensypreu/desdeny
Pride	Pride	Pride	Pride			Orgullo	Orgull
		Arrogance				Arrogancia	Arrogància
Jealousy		Jealousy				Celos	Gelosia
Envy		Envy				Envidia	Enveja
Shame	Shame	Shame	Shame	Shame/guilty		Vergüenza	Vergonya
Guilt	Guilt	Guilt	Guilt			Culpabilidad	Culpabilitat
	Embarrassment		Embarrassment			Desconcierto	Desconcert
					Disappointed	Desilusionado	Desil·lusionat
Relief	Relief					Alivio	Alleujament
Hope						Esperanza	Esperança
					Confident	Confiado	Confiat
Love			Love		Loving	Amor/cariñoso	Amor
					Affectionate	Afectuoso	Afectuós
Compassion		Pity				Compasión	Comprensió
		Moral rapture				Entusiasmo	Entusiasme
		Moral indignation				Indignación	Indignació
Aesthetic						Estético	Estètic

es pot comprovar, crea una nova llista de 34 emocions bàsiques, cosa que mostra la complexitat que suposa intentar discretizar l'espai emocional per representar-ho mitjançant un conjunt limitat d'etiquetes emocionals.

Ekman (1999) proposa el concepte de famílies d'emocions en lloc de definir una etiqueta per descriure una emoció concreta, ja que considera que cada emoció no és un únic estat afectiu sinó una família d'estats relacionats. Cada família es caracteritza per un tema que és fruit de l'evolució i unes variacions que són reflex de l'aprenentatge. Ekman proposa una llista amb 15 famílies d'emocions mostrades en la segona columna de la taula 2.1. Així, per exemple, la família *Anger* abastaria les emocions d'enuig, empipament i ràbia, totes amb un tema comú però amb diferents matisos fruit d'elements apresos prèviament pels individus.

En la llista anterior hi ha molts termes que descriuen estats relacionats amb l'emoció més que emocions en si com, per exemple, els estats confiat, avorrit o relaxat. Així doncs, les teories de les emocions es refereixen, en la seva gran majoria, a les emocions bàsiques mentre que les altres etiquetes emocionals descriuen combinacions o variacions d'aquestes emocions bàsiques. No obstant això, no existeix un consens sobre quines han de ser considerades emocions bàsiques.

Les etiquetes esmentades en aquesta secció són sovint també referenciades com a categories, etiquetes categòriques, etiquetes emocionals o classes. En la present tesi tots aquests termes s'empren com a sinònims.

2.1.3.2 Models *circumplex*

Donada la dificultat de crear un llistat únic d'emocions, és a dir, de treballar amb una representació completament discreta de l'univers emocional, s'han creat altres representacions alternatives que, tot i poder ser discretitzades, es basen en una representació contínua d'aquest univers. Així, alguns investigadors opten per representar les emocions mitjançant una estructura circular contínua en la qual poden situar-se les etiquetes emocionals principals. La proximitat de dues categories representa emocions conceptualment similars mentre que les emocions contràries estan situades en posicions diametralment oposades. En les posicions intermèdies apareixerien les combinacions de les emocions adjacents. Aquests models reben el nom de models *circumplex* i es poden equiparar a una roda de color on els colors primaris ocuparien posicions maximalment distanciades i, entre ells, apareixerien les combinacions dels mateixos. El primer model *circumplex* data del 1941, obra d'Harold Schlosberg, obtingut de l'observació de què els errors de reconeixement de l'expressió facial es corresponien a la confusió entre categories adjacents situades sobre una circumferència (Schröder, 2004). El 1958, Robert Plutchik va proposar un model amb vuit emocions bàsiques bipolars: alegria-avorriment, enuig-por, acceptació-fàstic i sorpresa-expectació. Una evolució d'aquesta teoria porta al model *circumplex* tridimensional de (Plutchik, 2001) representat en la figura 2.1, el qual es defineix pels següents aspectes:

- La distància des del centre de la figura fins a l'etiqueta emocional representa la intensitat de l'emoció.

- En els cercles concèntrics es representen les emocions segons el seu grau de similitud de manera que les emocions similars estan properes i les oposades estan separades diametralment.

- Els vuit sectors representen les emocions bàsiques bipolars del model de Plutchik de 1958.

- En els espais en blanc tenen cabuda les emocions que són combinacions de dues emocions primàries com, per exemple, el fàstic i la ràbia per produir odi.

En la figura 2.1 es pot observar que les emocions secundàries es defineixen mitjançant la combinació de les emocions primàries que ocupen posicions adjacents. Així, per exemple, el remordiment es concep com una mescla de tristesa i aversió cap a la pròpia conducta. A més, variant la intensitat de les emocions es poden obtenir nous estats emocionals. D'aquesta forma el temor pot anar des d'una simple aprensió fins a un enorme terror. En la mateixa gràfica també s'observa que hi ha emocions oposades, com la tristesa i la alegria per exemple, que pel fet d'ocupar posicions distants no es poden combinar.

2.1.3.3 Espai multidimensional

La creació d'una llista d'etiquetes per definir estats emocionals suposa treballar en un univers discret i limitat al nombre d'etiquetes creat. D'altra banda, els models *circumplex* mostren una relació entre els estats emocionals segons la seva posició geomètrica dins d'una representació contínua però no defineixen les emocions en si. Una alternativa que pretén considerar un univers continu però, al mateix temps, donar una interpretació tant a les emocions com a la relació que existeix entre elles, es pot trobar en les representacions basades en espais multidimensionals. En aquest tipus de models, els estats emocionals es representen com a coordenades d'un espai de dimensionalitat reduïda. Existeixen nombrosos estudis relacionats amb aquest tipus de representació i, tal com passa amb els models anteriors, la definició d'aquest espai no és senzilla i existeixen diverses alternatives. El principal punt clau és determinar quantes i quines han de ser les dimensions que defineixen aquest espai. La primera referència als models d'espais multidimensionals data del 1896 gràcies al psicòleg alemany Wilhelm Wundt, i van ser emprats per vegada primera en l'àmbit de l'anàlisi emotiva facial per Schlosberg (1954). Cowie i Cornelius (2003) duen a terme una revisió històrica de les diferents dimensions que han considerat diversos autors. Schröder (2004), per la seva banda, duu a terme un estudi més ampli centrat en la descripció dels estats emocionals mitjançant espais multidimensionals, presentant una perspectiva històrica i una descripció del significat d'aquestes emocions i especificant què relació tenen amb el comportament emotiu humà.

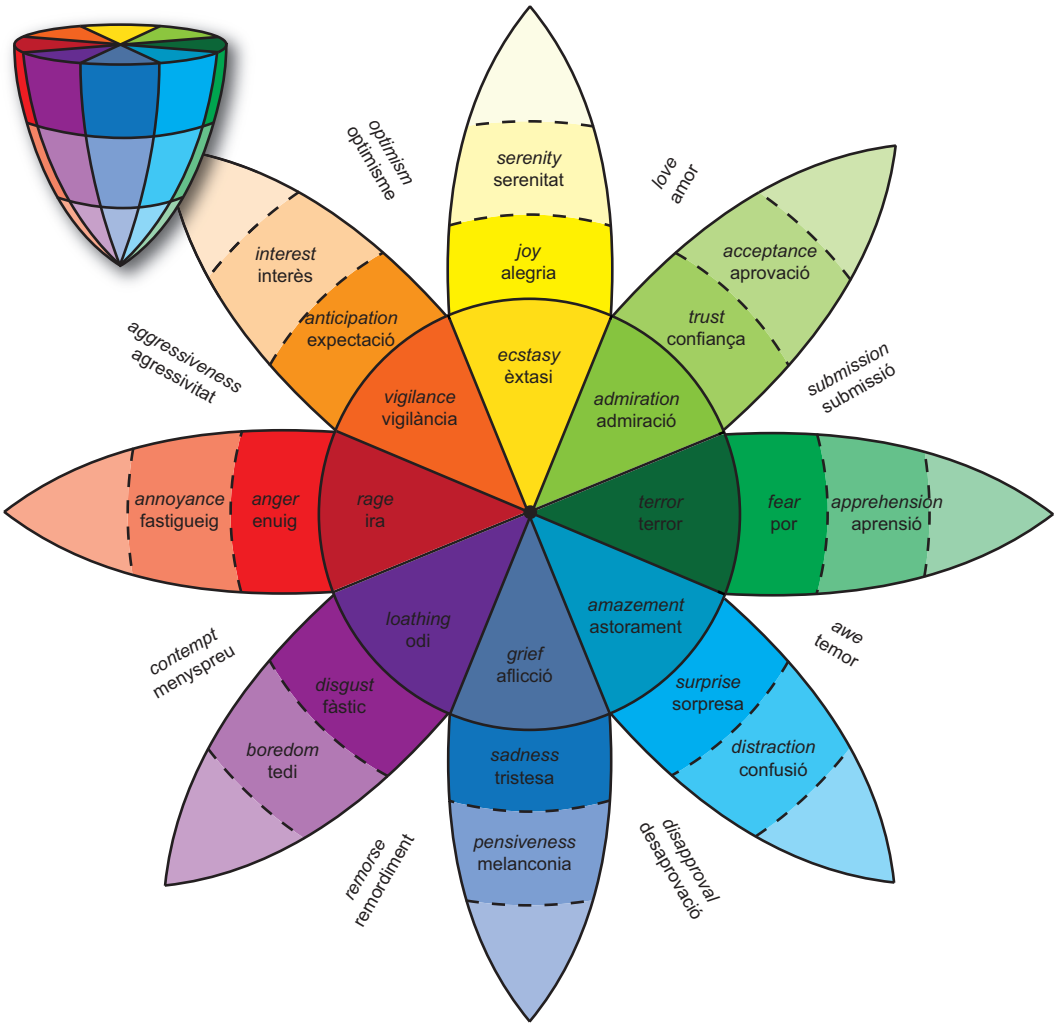


Figura 2.1: Model *circumplex* tridimensional de Plutchik i la seva traducció al català. Adaptat de (Plutchik, 2001) (en anglès).

En general, considerant diversos estudis, el nombre de dimensions de l'espai emocional es limita a dos o fins i tot tres. No obstant això, la definició de cadascuna d'aquestes dimensions difereix d'uns estudis a uns altres. Malgrat les variacions, molt sovint aquests eixos es refereixen al mateix concepte tot i que emprant diferent nomenclatura per a la seva definició. Els tres eixos poden situar-se en qualsevol orientació de l'espai, si bé sempre de forma ortogonal entre si. Les tres dimensions més emprades són:

- Avaluació (també referenciada com a grat, valoració o valència). Defineix un eix positiu-negatiu classificant les emocions segons si són més agradables o més desagradables. Per exemple: l'alegria és una emoció positiva mentre que l'enuig és una emoció negativa.

- Activació (o activitat). Defineix un eix actiu-passiu que indica la presència o absèn-

cia d'energia o tensió. Per exemple, l'enuig se situa en la posició més activa de l'eix mentre que l'avorriment se situa en la posició més passiva.

- Potència (també referenciada com a control o força). Defineix una escala dominant-submís o alta-baixa la qual distingeix les emocions iniciades pel subjecte d'aquelles que són provocades per l'entorn. Per exemple, el menyspreu se situa en la posició més dominant, o alta, de l'escala mentre que la sorpresa se situa en la més submissa, o baixa.

La dimensió de potència no s'estudia de manera tan habitual com les dues anteriors de manera que molt sovint l'espai emocional es representa com una circumferència en la qual es defineixen els eixos vertical i horitzontal com a activació (actiu-passiu) i avaluació (positiu-negatiu), respectivament. A aquest model bidimensional se'l coneix com, simplement, model d'activació-avaluació.

Atès que tant les llistes d'etiquetes emocionals com els models dimensionals permeten la descripció d'emocions, es pot realitzar un mapatge entre tots dos mètodes de representació (Steidl, 2009). La figura 2.2 mostra la disposició de 80 estats afectius en l'espai bidimensional d'activació-avaluació segons la reproducció que Steidl (2009) fa del treball de Scherer (2000). En la figura s'anteposa el concepte en anglès (en cursiva) a la traducció en català atès que, tal com s'ha explicat en la secció 2.1.3.1, en la traducció es poden perdre algunes subtileses del terme original. La disposició dels elements en aquesta gràfica s'obté d'un escalat multidimensional basat en la matriu de similitud creada a partir dels judicis subjectius d'un conjunt d'avaluadors humans. Atès que aquests avaluadors mesuren les similituds de les 80 etiquetes emocionals donades, la posició exacta de les mateixes variaria si canviés tant el seu nombre com la seva especificació (Steidl, 2009).

Existeixen altres models que, bàsicament, consisteixen en el mapatge d'altres etiquetes emocionals sobre aquest mateix espai bidimensional. Tal com s'ha pretès assenyalar al llarg d'aquestes línies, no són millors ni pitjors, tan solament consisteixen en alternatives o diferents punts de vista al voltant d'un mateix assumpte. Així, per exemple, també són representatius els models proposats per Russell (1980), amb el mapatge de 28 etiquetes emocionals, o el basat en la llista de 142 etiquetes de Plutchik (1980) les coordenades del qual en l'espai d'activació-avaluació es poden consultar en (Cowie *et al.*, 2001).

Malgrat la importància de les dimensions anteriorment citades, l'especificitat d'alguns treballs determina, en ocasions, definicions alternatives més adequades que unes altres al problema tractat. Així, per exemple, Batliner *et al.* (2008) proposa el canvi de la dimensió d'activació per la d'interacció per abordar l'anàlisi d'emocions espontànies[5]. Aquest eix situa les emocions depenent de si es dirigeixen a un mateix (com per exemple en el cas de l'alegria) o cap als altres (com en el cas de la irascibilitat).

[5]La major part dels estudis dels quals deriven les dimensions anteriors es basen en corpus actuats, no espontanis.

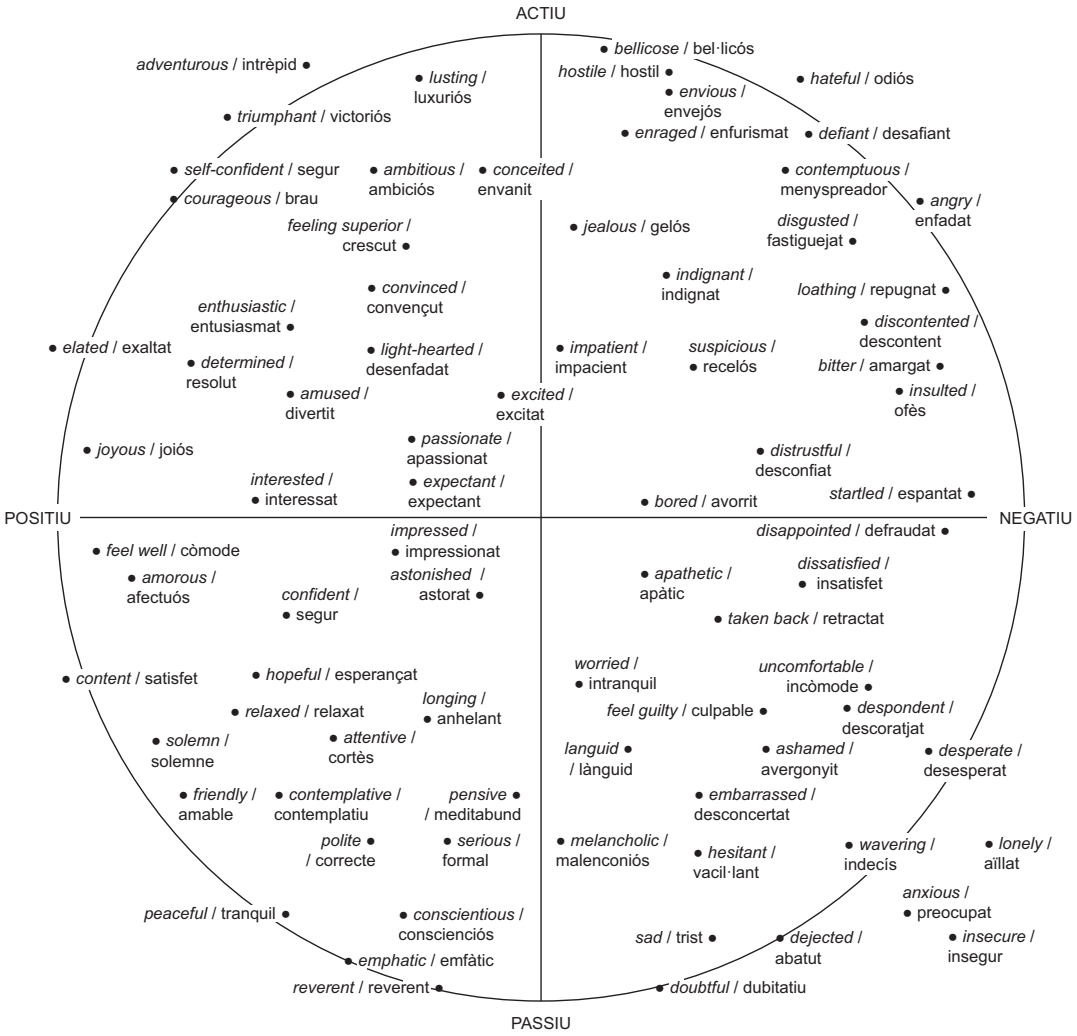


Figura 2.2: Mapatge de 80 etiquetes emocionals en l'espai bidimensional d'activació (eix vertical: actiu-passiu) i avaluació (eix horitzontal: positiu-negatiu) i la seva traducció al català. Adaptat de (Scherer, 2000) segons la reproducció de Steidl (2009) (en anglès).

Els models multidimensionals semblen solucionar els inconvenients dels categòrics. No obstant això, i tal com indiquen Cowie i Schröder (2005), tots dos presenten avantatges i inconvenients. És cert que treballar amb un llistat d'emocions pot resultar molt complex quan es consideren moltes etiquetes i que el model multidimensional permet solucionar aquest problema. No obstant això, aquest últim presenta dificultats per fer distincions significatives entre estats emocionals. Així, en reduir la dimensionalitat de l'espai, es produeix una pèrdua d'informació i emocions dispars com la por i la ira queden situades en posicions properes (atès que ambdues són d'avaluació negativa i activació passiva), mentre que en el model *circumplex* de Plutchik ocupen posicions diametralment oposades.

2.2 Etiquetatge subjectiu d'estímuls

Aquesta tesi està orientada al reconeixement afectiu automàtic. Malgrat això, per poder realitzar aquesta tasca és necessari disposar d'un conjunt de dades convenientment etiquetat. D'aquesta forma, un conjunt de dades etiquetat es converteix en una font d'informació a partir de la qual els algorismes encarregats de dur a terme la tasca de reconeixement automàtic poden aprendre a identificar els diferents estats afectius. A més, gràcies a l'etiquetatge de cadascun dels elements d'aquest conjunt de dades, es pot realitzar la comprovació de si la tasca automàtica realitza correctament la seva comesa comparant l'etiqueta assignada pel sistema i la que cada element porta ja assignada.

No obstant això, realitzar l'etiquetatge dels elements d'un conjunt de dades no és tan fàcil com pot semblar a ull nu. Deixant de banda els estímuls no procedents d'una reacció humana (com podrien ser, per exemple, fotografies o fragments de text) i centrant l'estudi dins de l'àmbit d'aquesta tesi (fragments de locucions pronunciats per un individu), l'etiquetatge d'aquests elements depèn molt sovint de com s'han realitzat els enregistraments. Així, per exemple, si els enregistraments procedeixen d'actors i són, per tant, simulats seguint un guió, llavors no hi ha inconvenient en afirmar, directament, que cada locució es correspon amb l'emoció que se li ha demanat a l'actor que representi. En cas de dubtar de la capacitat interpretativa de l'actor encarregat de l'enregistrament, es podria realitzar una neteja del conjunt de dades eliminant aquelles locucions que, de forma subjectiva, no aconseguissin expressar l'emoció que es requeria[6]. En el cas de què els enregistraments procedeixin d'individus als quals se'ls ha provocat d'alguna forma una emoció sovint és necessari recórrer al autoetiquetatge (és a dir, consultar al propi individu l'emoció que ha experimentat) o a l'avaluació que un grup representatiu de persones facin de l'emoció que observen. Això és així perquè, malgrat haver pretès provocar una determinada emoció, és possible que l'individu hagi experimentat una altra en un grau diferent a l'esperat. En aquest cas, l'anàlisi de centra en l'efecte i no en la causa que provoca la reacció de l'individu, tal com assenyalen Cowie *et al.* (2001).

Recórrer a l'avaluació subjectiva d'un determinat estímul, ja sigui un fragment de veu, una seqüència audiovisual, una fotografia, un text o un estímul de qualsevol altra naturalesa, suposa plantejar la representació dels estats afectius per després identificar-los. Així, per exemple, es poden emprar etiquetes categòriques per identificar les emocions, tal com es plantejava en la secció 2.1.3.1 i en la secció 2.1.3.2. D'altra banda, es pot recórrer a l'ús dels models dimensionals descrits en la secció 2.1.3.3. L'ús d'etiquetes emocionals per descriure les emocions suposa una simplificació de la tasca atès que els avaluadors tan sols han de seleccionar l'etiqueta emocional que consideren que millor descriu l'estat afectiu que es desprèn de l'estímul analitzat. No obstant això, una llista massa extensa d'etiquetes pot complicar molt aquesta tasca. D'altra banda, l'ús de models basats en espais multidimensionals suposa un major grau de coneixement de les teories emocionals per part de l'avaluador. Malgrat això, confereix una major llibertat comparat amb l'ús de

[6] Aquest procediment es correspondria amb un procés de validació del corpus i suposaria una avaluació subjectiva de cada locució del mateix. En la secció 5.2 es proposa un esquema per automatitzar-ho.

les etiquetes categòriques mentre que, habitualment[7], proporciona una forma d'avaluar els estímuls en un univers continu en lloc de discret i, per tant, també més limitat.

2.2.1 Etiquetatge categòric

L'assignació d'etiquetes als estímuls sotmesos a avaluació és, potser, el mecanisme més senzill d'implementar. No fa falta recórrer a eines informàtiques per realitzar un test a un grup d'avaluadors subjectius atès que es pot dur a terme fàcilment en paper. No obstant això, per gestionar les respostes de molts usuaris i també permetre que una sèrie d'estímuls pugui ser avaluada de forma no presencial per usuaris de diverses procedències i/o en moments de temps diferents, es pot recórrer a eines que permetin la realització de tests a través d'Internet[8].

Malgrat això, sovint existeix ambigüitat en l'estat emocional associat a l'estímul avaluat i els tests que limiten les respostes a una sèrie d'etiquetes predefinides i exclusives (conegudes com *decisions dures*[9]) poden confondre a l'avaluador i el poden forçar a donar una resposta inadequada. Per aquest motiu es poden crear tests en els quals els avaluadors hagin de marcar valors numèrics, expressats en forma de percentatges, per a cada etiqueta emocional (coneguts com *decisions toves*[10]). Aquests valors reflectirien el grau de seguretat en la resposta proporcionada. L'eina *eTool* (Steidl, 2009), per exemple, implementada per a l'etiquetatge del corpus FAU Aibo (Steidl, 2009), permet que l'avaluador pugui indicar aquest percentatge per a les diferents etiquetes de les quals disposa per realitzar la seva tasca.

2.2.2 Etiquetatge dimensional

Existeixen diverses eines que es basen en models multidimensionals per dur a terme l'avaluació subjectiva d'estímuls. Encara que no necessàriament, moltes d'elles estan creades amb la idea de ser implementades com a eines informàtiques tot i que algunes es poden adaptar per ser utilitzades en paper. En aquesta secció es destaquen tres d'aquestes eines: *Feeltrace*, la roda d'emocions *Geneva* i els *Self-Assessment Manikins*.

2.2.2.1 *Feeltrace*

Feeltrace (Cowie *et al.*, 2000) és una eina informàtica que permet l'anotació de l'estat emocional d'un individu en temps real mentre percep un estímul audiovisual. Durant

[7] Habitualment perquè encara que els eixos dimensionals són naturalment continus aquests es poden discretizar per facilitar la tasca dels avaluadors.

[8] Vegeu l'apèndix B.

[9] De l'anglès, *hard decisions*.

[10] De l'anglès, *soft decisions*.

la reproducció de l'estímul l'usuari pot anar marcant punts en la circumferència de l'espai d'activació-avaluació mitjançant un cursor per a la seva posterior anàlisi pel que, de forma implícita, s'emmagatzema també una referència temporal durant l'etiquetatge. Tal com descriuen Cowie *et al.* (2001), per indicar a l'avaluador el que significa cada posició del cursor en l'espai s'inclouen diversos indicadors. En primer lloc, els eixos es descriuen segons les dimensions d'activació i avaluació pel que, en els extrems dels mateixos apareixen les llegendes "Molt actiu" i "Molt passiu", i "Molt negatiu" i "Molt positiu", respectivament. D'altra banda, el cursor pren un color diferent segons la seva posició dins dels eixos anteriors, segons l'escala vermell, groc, verd i blau per a cadascun dels quatre punts de tall de la circumferència amb els eixos i la gradació de color corresponent per a posicions intermèdies. Així mateix, al voltant de la circumferència de l'espai d'activació-avaluació apareixen etiquetes que descriuen les emocions associades a aquesta regió de l'espai. Finalment, a l'interior d'aquesta circumferència, s'inclou un conjunt d'etiquetes seleccionades d'una llista procedent d'un vocabulari bàsic d'emocions en anglès (Cowie *et al.*, 1999) situades en les seves coordenades corresponents. Segons els seus autors, aquesta eina té la mateixa potència que un vocabulari de vint paraules, amb l'avantatge de permetre estats intermedis i fer el registre de l'evolució temporal d'un estat emocional a un altre. La figura 2.3 mostra aquesta eina incloent la traducció al català de les etiquetes emocionals.

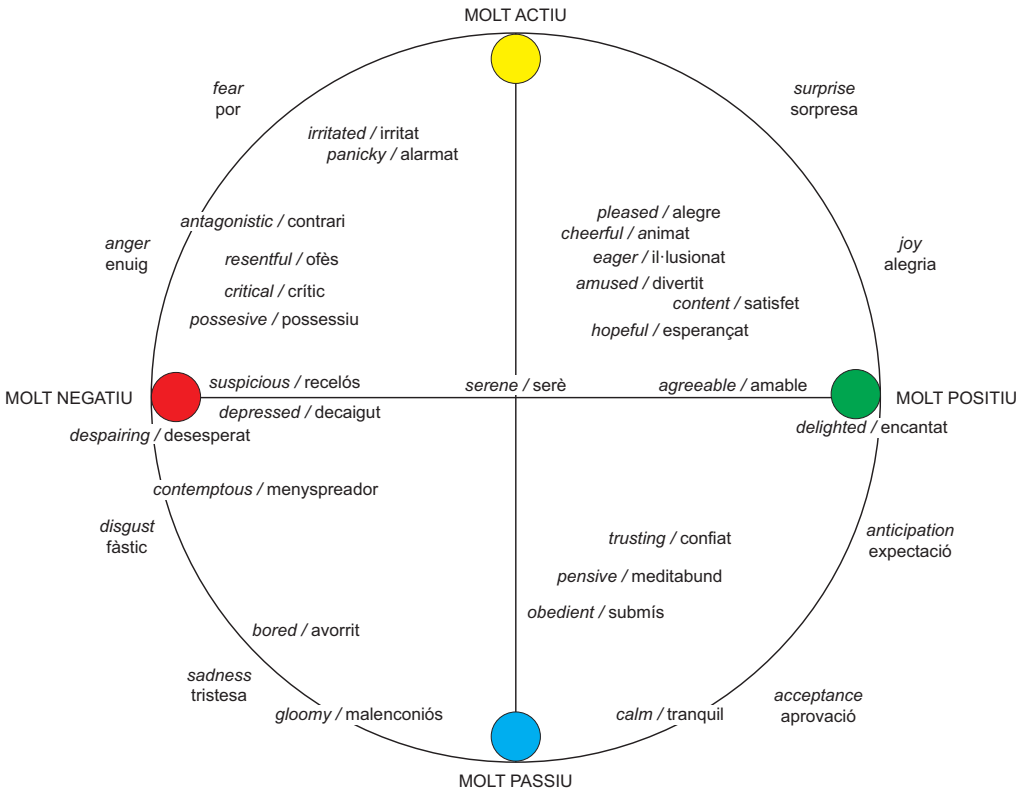


Figura 2.3: Eina *Feeltrace* (Cowie *et al.*, 2000) utilitzada per anotar l'emoció d'un estímul audiovisual en un espai bidimensional. Adaptat de (Cowie *et al.*, 2001) (en anglès).

2.2.2.2 La roda d'emocions *Geneva*

La roda d'emocions *Geneva*[11] (Scherer, 2005) és una eina bastant similar a *Feeltrace* tot i que presenta algunes diferències notables. Aquesta eina no està orientada únicament al seu ús com a eina informàtica, si bé és com principalment es va concebre. Les dimensions representades són control (alt-baix) i valència o avaluació (positiu-negatiu, agradable-desagradable). Tal com es mostra a la figura 2.4, en un cercle definit pels eixos anteriorment citats es distribueixen 16 famílies emocionals creant 16 sectors circulars. En cada sector se situen, des del centre fins als límits del sector, quatre elements en forma de cercles de color. Aquests elements presenten diferents tonalitats cromàtiques per a cada família emocional, així com diferents grandàries i saturacions de color segons la intensitat emocional representada. La grandària dels cercles i la saturació del color decreix segons disminueix la seva distància al centre de la roda representant una menor intensitat d'emoció.

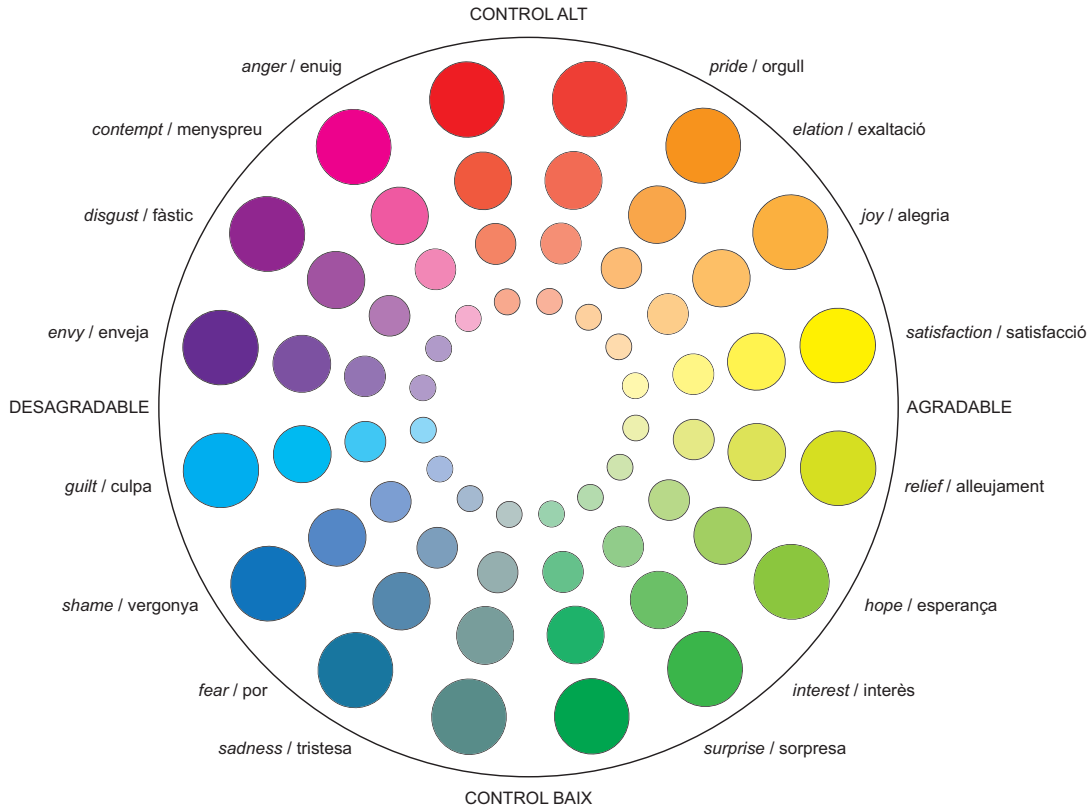


Figura 2.4: Roda d'emocions *Geneva* (Scherer, 2005). Adaptat de (Steidl, 2009) (en anglès).

[11] *Geneva Emotion Wheel*, de l'original, en anglès.

2.2.2.3 *Self-Assessment Manikin*

La tècnica del *Self-Assessment Manikin* (SAM) (Lang, 1980) també està relacionada amb els models dimensionals. SAM és una tècnica d'avaluació pictòrica no verbal per mesurar les dimensions de valència, activació i control d'un determinat estímul en les escales negatiu-positiu, calmat-excitat i dominat-dominant, respectivament. Mentre que *Feeltrace* és una eina orientada al seu ús com a programari, SAM pot emprar-se a més a més com un test en paper. En aquest cas, l'eina no permet el registre directe de l'evolució temporal dels estats emocionals dels individus però el seu ús resulta més senzill que *Feeltrace*. Amb aquesta eina l'usuari ha de marcar en tres escales diferents, una per a cada dimensió avaluada, la resposta emocional que cregui oportuna. Per fer més senzilla aquesta tasca se li proporcionen uns dibuixos representatius de cadascuna de les escales a avaluar. La figura 2.5 mostra la implementació de la tècnica SAM, segons la descripció gràfica que realitzen Bradley i Lang (1994), en l'eina d'avaluació en línia d'estímuls multimèdia *Testing platfoRm for mUltimedia Evaluation* (TRUE) (Planet *et al.*, 2008)[12]. SAM, gràcies al seu fonament gràfic, ve a solucionar els problemes que es deriven de l'escala diferencial semàntica[13] proposada per Mehrabian i Russell (1974), en la qual les tres dimensions s'avaluen a partir de les puntuacions de 18 parelles d'adjectius en una escala de 9 possibles valors. A la incomoditat vinculada a la necessitat d'haver d'avaluar tants elements per a un estímul donat s'afegeix la d'emprar un sistema verbal, desenvolupat originalment en anglès, en individus no angloparlants o amb un desenvolupament lingüístic limitat (per exemple, nens i nenes o persones afàsiques) (Bradley i Lang, 1994).

2.3 Expressió i percepció d'emocions

No hi ha dubte de la importància de l'expressivitat emocional —i en conseqüència, també del reconeixement emocional— en la comunicació humana. Són molts els estudis que han investigat com l'emoció afecta a diferents aspectes del comportament humà, tal com es reflecteix en el resum realitzat per Bartneck (2000), el qual recull experiments i resultats relacionats amb les emocions a nivell facial, vocal, musical i de gestualitat corporal. Un exemple de la importància de les emocions en la comunicació pot detectar-se en la convergència que es produeix, de manera pràcticament involuntària, entre dos interlocutors que sostenen una comunicació. Dos individus que estan en sintonia i volen expressar-ho convergeixen en una sèrie de paràmetres vocals (Giles i Smith, 1979). Per contra, si aquesta convergència no es produeix, els agents de la comunicació mostren indiferència o distanciament (Cowie *et al.*, 2001).

En el procés comunicatiu humà es poden distingir dos canals diferenciats (Cowie *et al.*, 2001):

[12] Per a una explicació detallada de l'eina TRUE, consulteu l'apèndix B.
[13] *Semantic Differential Scale*, de l'original, en anglès.

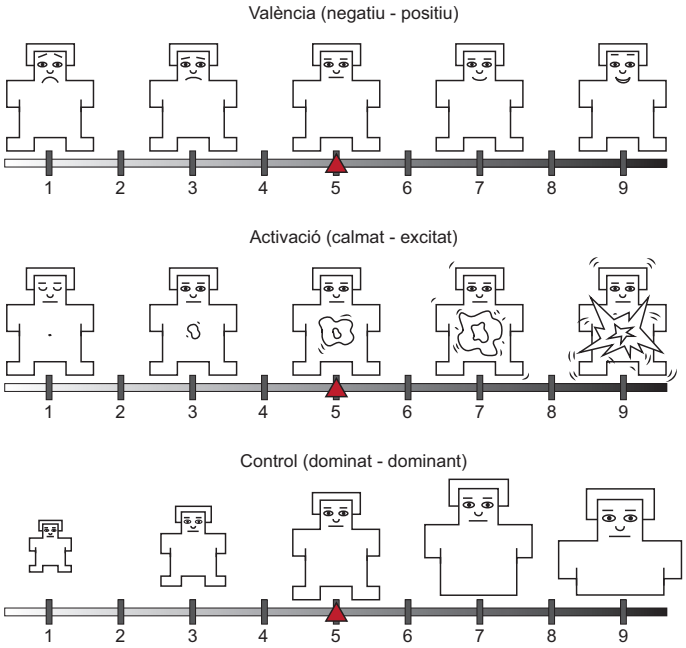


Figura 2.5: *Self-Assessment Manikin* (SAM). Figura adaptada de la seva implementació en l'eina TRUE (Planet *et al.*, 2008) segons la descripció gràfica de Bradley i Lang (1994).

- Canal explícit: destinat a la comunicació del missatge en si mateix.

- Canal implícit: destinat a la comunicació d'informació sobre els agents involucrats en el procés.

Si bé és cert que el contingut explícit d'un missatge pot transmetre informació de l'estat emocional de l'individu moltes vegades és necessari recórrer al canal implícit per obtenir aquesta informació. En certa manera, podem afirmar que tots dos canals interactuen de manera que el canal implícit indica als receptors com han d'interpretar el missatge que es transmet a través del canal explícit (Cowie *et al.*, 2001). Un exemple per il·lustrar aquesta afirmació es pot trobar en els acudits o en els diàlegs sarcàstics, en els quals la informació del canal implícit confereix un significat completament diferent al que es derivaria de la interpretació exclusiva del missatge del canal explícit.

Tal com s'ha explicat en la secció 2.1.2, les emocions desencadenen una sèrie de canvis fisiològics que predisposen a l'acció. La següent secció explica alguns d'aquests canvis a nivell fisiològic general. A continuació es detallaran els diferents elements que formen part de la comunicació humana per donar pas, en les dues seccions posteriors, a l'explicació dels efectes de l'expressió emocional en la veu i com aquests es reflecteixen també en la interpretació musical.

2.3.1 Reaccions fisiològiques relacionades amb l'emoció

Goleman (1995) posa l'accent en la consideració de les emocions com a impulsos per a l'acció i assenyala una sèrie de canvis fisiològics relacionats amb elles descrits com a predisposicions biològiques que afecten, directa o indirectament, a diverses funcions de l'organisme. La següent llista resumeix aquestes predisposicions per a emocions bàsiques que, com es pot comprovar, està molt lligada als postulats de Darwin i la seva teoria evolutiva basada en la supervivència:

- Enuig: augmenta el flux sanguini a les mans, el flux cardíac i la quantitat d'adrenalina en la sang. Això facilita l'agressió manual a l'oponent així com l'escomesa d'accions enèrgiques. Com en el cas de la por, s'activa el sistema nerviós simpàtic[14] preparant l'organisme per a l'acció.
- Por: augmenta el flux sanguini en la musculatura esquelètica llarga (i el disminueix, en conseqüència, del rostre ocasionant pal·lidesa) i la taxa hormonal en sang, la qual cosa facilita la fugida i indueix inquietud i un estat d'alerta. L'atenció es fixa en l'estímul que serveix d'amenaça i es percep certa rigidesa i paràlisi inicial i temporal que facilita l'avaluació de la situació. Es produeix l'activació del sistema nerviós simpàtic.
- Felicitat: s'estimula el centre cerebral encarregat d'inhibir sentiments negatius i estats de preocupació, augmentant l'energia. Això afavoreix un estat de tranquil·litat, repòs i entusiasme per emprendre accions.
- Amor: s'activa el sistema nerviós parasimpàtic[15] que indueix un estat de relaxació i de disminució de l'estrès que afecta de forma generalitzada a tot el cos. La sensació de calma i satisfacció afavoreix la convivència.
- Sorpresa: els ulls s'obren (produint el característic arqueig de celles d'aquesta emoció) per permetre una major entrada de llum en la retina i un augment del camp visual.
- Fàstic: en l'expressió facial associada al fàstic es frunzeix el nas per tancar-lo i evitar una olor desagradable i es torça, al mateix temps, el llavi superior per afavorir l'expulsió d'un aliment tòxic de l'aparell digestiu.
- Tristesa: disminueix l'energia i l'entusiasme per activitats vitals, sobretot les relacionades amb la diversió i el plaer, i el metabolisme corporal es ralenteix. Aquesta

[14] Part del sistema nerviós autònom que manté els mecanismes homeòstatics de l'organisme i el prepara, en cas de necessitat, per emprendre una acció de forma immediata augmentant la freqüència cardíaca o estimulant les glàndules suprarenals, per exemple.

[15] El sistema nerviós parasimpàtic és part del sistema nerviós autònom la funcionalitat del qual és complementària a la del sistema nerviós simpàtic. El seu temps de resposta és inferior a la del seu complementari atès que la seva funció no és la de donar una resposta immediata a un estímul. S'encarrega del restabliment de l'energia corporal.

resposta facilita l'avaluació de la situació que ha provocat aquest estat emocional i la planificació de l'activitat una vegada superada.

Aquestes reaccions fisiològiques repercuteixen també en la musculatura i en els òrgans relacionats amb l'aparell fonador provocant canvis en la veu de l'individu que està experimentant un estat emocional concret. No obstant això, també comporten canvis en altres components de la comunicació. La següent secció detalla quins són els components de la comunicació abans d'aprofundir, en les seccions subsegüents, en l'efecte de l'emoció en la veu i la seva expressió musical.

2.3.2 Elements multimodals de la comunicació

Atès que l'emoció va lligada a una sèrie de canvis fisiològics, aquests canvis influeixen de forma directa i indirecta en la comunicació. Ara bé, el procés comunicatiu s'estableix a diversos nivells i, per tant, l'expressió emocional es manifesta de diferents formes. La figura 2.6, segons Payrató (2005) a partir de Hinde (1972), mostra els components del procés comunicatiu a nivell vocal i no vocal.

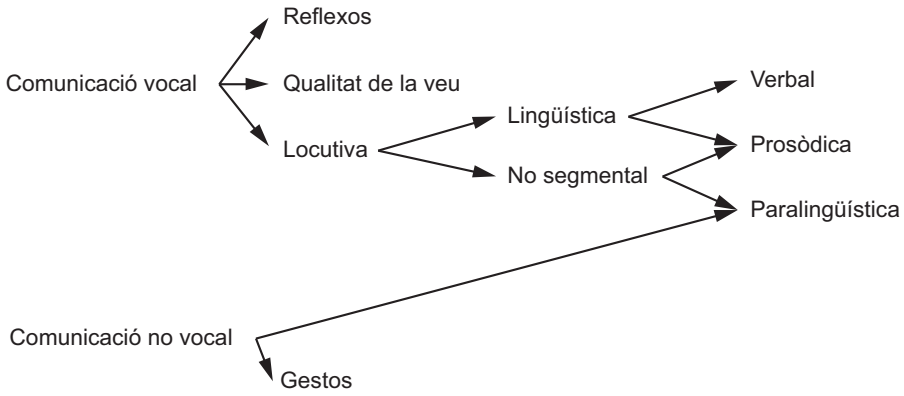


Figura 2.6: Components a nivell vocal i no vocal de la comunicació humana. Adaptat de (Payrató, 2005).

En la figura 2.6 el concepte reflexos fa referència a activitats com, per exemple, tossir i qualitat de veu es refereix a característiques pròpies de l'individu. Així mateix, la prosòdia agrupa el to, la intensitat i la quantitat, tal com es detallarà en la secció 2.3.3. Finalment, la paralingüística inclou els components tant vocals com no vocals que completen la modalitat verbal.

D'altra banda, i considerant els aspectes expressius i receptius de la comunicació, podem distingir diferents modalitats bàsiques en el procés comunicatiu humà. Aquestes modalitats apareixen reflectides en la figura 2.7, segons l'adaptació de Payrató (2005) de l'original de Heinemann (1980).

En la figura 2.7 es destaca la proxèmica com la gestió de l'espai que envolta als in-

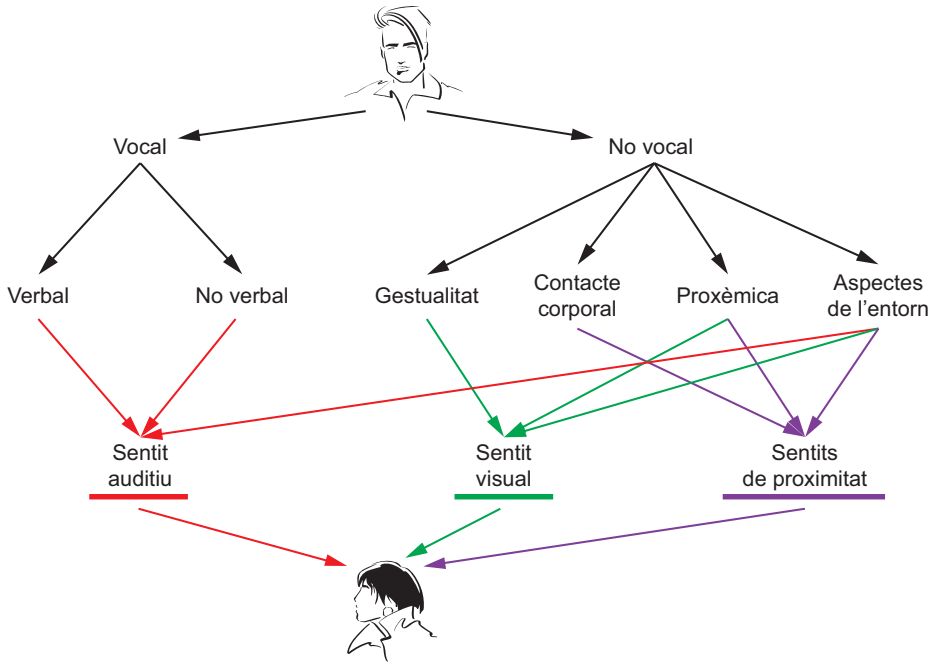


Figura 2.7: Modalitats del procés comunicatiu humà. Adaptat de (Payrató, 2005), segons l'original de Heinemann (1980).

terlocutors i el seu entorn pel que fa a les seves característiques físiques, tals com la seva aparença o el seu vestuari. Els sentits auditiu i visual es consideren "sentits a distància" ja que que proporcionen informació dels objectes des de lluny, i tradicionalment s'han conegut com a "sentits nobles". Com a "sentits de proximitat" es consideren el tacte i el gust. L'olfacte se situaria en un punt intermedi entre ambdues categories i, juntament amb els dos anteriors, formaria part del conjunt conegut tradicionalment com a "sentits inferiors". Ha de considerar-se, no obstant això, que els conceptes que s'estan presentant al llarg d'aquesta secció no sempre poden distingir-se clarament els uns dels altres doncs, sovint, comparteixen certes característiques definitòries o es corresponen amb fenòmens que es donen de forma combinada i fins i tot, en ocasions, de forma inseparable. Per exemple, el gest realitzat amb el dit índex assenyalant un objecte mentre es pronuncia la frase "*Apropa'm això, si us plau*" seria un exemple de submodalitat no verbal sincronitzada amb una submodalitat verbal. Quant a la gestualitat, es poden distingir com a principals fonts de producció de senyals la gesticulació manual, la gesticulació facial, els moviments del cap, els moviments oculars i la postura corporal. La classificació dels gestos més estesa es deu al treball publicat per Ekman i Friesen (1969), en el qual es distingeixen cinc categories d'actes no verbals:

- Emblemes: gestos que poden ser traduïts directament al llenguatge verbal ja que tenen un significat inequívoc i molt lligat a l'àmbit cultural.

- Il·lustradors: complementen el missatge verbal del que depenen.

- Reguladors: destinats a la gestió de la comunicació, facilitant l'intercanvi de torns entre els diferents interlocutors.
- Adaptadors: gestos de contacte amb un mateix, amb objectes o amb altres interlocutors.
- Manifestadors d'estats anímics: expressió d'emocions.

Atès que el procés comunicatiu s'estableix en diversos nivells, tal com s'ha constatat en les línies anteriors, l'expressió emocional es mostra també en aspectes diversos. De les diferents modalitats en les quals pot manifestar-se l'emoció, en la següent secció es posa l'accent en les relacionades amb la parla.

2.3.3 Paràmetres de la parla relacionats amb l'emoció

Malgrat que com s'ha explicat en la secció 2.3.2 la comunicació és un procés multimodal i, per tant, l'expressió emocional pot trobar-se en aquestes múltiples modalitats, la parla per si sola també és capaç de transmetre estats afectius. Així, per exemple, els agents d'una conversa telefònica són capaços de diferenciar les seves emocions tot i no veure les seves expressions facials o les seves gesticulacions corporals. Cal destacar, no obstant això, que tot i que el component afectiu de la parla és principalment de naturalesa no-lèxica, hi ha altres factors en ella, tals com el context o el contingut del missatge, que són també importants en el que a contingut emocional es refereix.

Per entendre com l'emoció pot afectar a la parla s'ha de comprendre com es genera aquesta. La figura 2.8 mostra, de forma esquemàtica, el tracte vocal. De forma molt bàsica, la veu es genera quan l'aire expulsat dels pulmons arriba a la laringe després de travessar la tràquea. En la laringe es troben quatre plecs vocals[16] malgrat que només dos d'ells, els inferiors, són els relacionats amb la producció de la veu. Quan aquests plecs estan oberts l'aire flueix lliurement a través de l'orifici que creen. No obstant això, quan es contreuen i s'ajunten, l'aire impacta contra ells i la fricció els fa vibrar. La freqüència d'aquesta vibració determina la freqüència fonamental del senyal de veu i és, per tant, de la flexibilitat d'aquests plecs que depèn el rang vocal. Al mateix temps, aquesta vibració també genera un espectre d'harmònics[17] que es filtra en el tracte vocal creant-se els diferents sons. A més, es produeix una amplificació del senyal de veu atès que el so produït pels plecs vocals és de baixa intensitat. Així mateix, la fricció de l'aire al seu pas pel tracte vocal crea altres components aperiòdics segons la morfologia que presenti. Això crea, per exemple, els sons sords o els sons consonàntics, deguts a interrupcions totals o parcials del flux d'aire. D'aquesta forma, els sons oclusius, per exemple, es deuen a un tancament del flux de l'aire seguit d'una explosió[18]. L'evolució temporal d'aquest procés genera una

[16] Plecs vocals és el nom anatòmic que reben les comunament conegudes com *cordes vocals*.
[17] L'espectre d'harmònics està format per senyals a freqüències múltiples de la freqüència fonamental.
[18] Per exemple, en unir i separar bruscament els llavis.

ona acústica que es coneix com el senyal de veu. Les múltiples variacions i combinacions dels sons generats són les que permeten la formació de paraules i la transmissió d'un missatge definit.

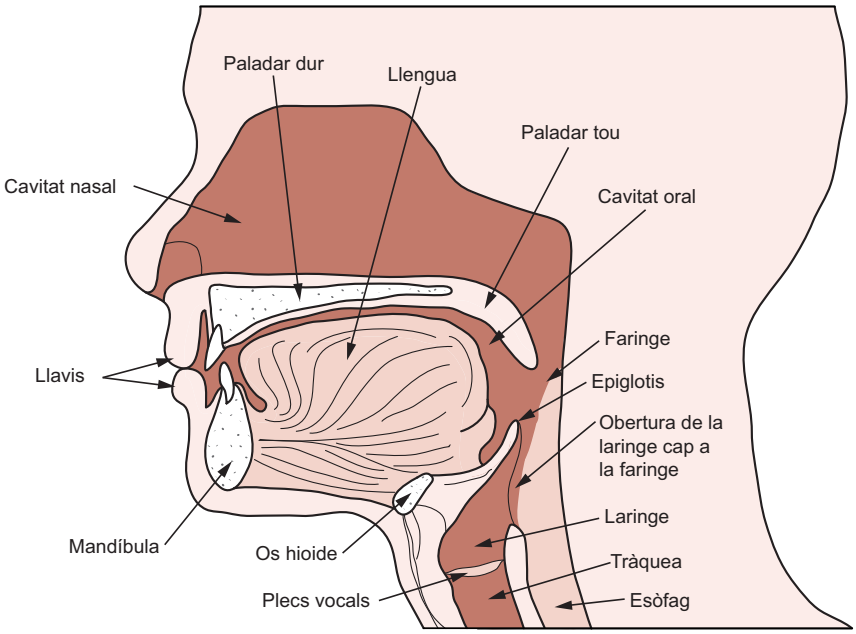


Figura 2.8: Model esquemàtic del tracte vocal.

La prosòdia és la branca de la lingüística que tracta els elements de l'expressió oral i la formació de les paraules. En aquest sentit, els principals trets prosòdics de la parla són: les variacions en la intensitat i en la freqüència fonamental, la durada dels sons i la posició i la durada de les pauses (Llisterri *et al.*, 2004). La funció de la prosòdia és, principalment, lingüística. Així, per exemple, és la que permet la distinció entre una frase enunciativa i una frase interrogativa, sense ser necessària la interpretació del contingut de la frase. No obstant això, la parla també mostra variacions prosòdiques i de timbre amb funcions que no poden considerar-se estrictament lingüístiques. Es distingeixen dues funcions de la prosòdia, a més de la lingüística anteriorment esmentada (Escudero, 2003):

- Paralingüística: complementa el missatge amb una intenció determinada, reflectint una actitud o un estat afectiu del parlant.
- Extralingüística: aporta informació sobre les característiques del parlant, tal com la seva edat, el seu sexe, el seu estatus social, etc.

Les principals propietats acústiques (tant prosòdiques com no prosòdiques) dels sons de la parla que es relacionen amb l'expressivitat vocal es detallen en el següent llistat (Iriondo, 2008), segons la seva relació amb la melodia[19], la intensitat, els aspectes temporals de la parla o el timbre:

[19]La melodia (*pitch*, en anglès) és el fenomen que es relaciona amb la corba de Freqüència fonamental (F0)

- Propietats relacionades amb la melodia:
  - Freqüència fonamental (F0): resultat de la vibració dels plecs vocals. Es defineix com el cicle periòdic del senyal de veu. La seva unitat de mesura habitual és l'hertz (Hz), que mesura el nombre de cicles per segon.
  - Corba de F0 o corba melòdica: és la seqüència de valors de F0 per a un fragment de veu i es relaciona amb la percepció de l'entonació de la parla.
  - *Jitter*: lleugera pertorbació en la F0 deguda a fluctuacions en els temps d'obertura i tancament dels plecs vocals entre dos cicles consecutius.
- Propietats relacionades amb la intensitat:
  - Intensitat: mesura de l'energia del senyal acústic. La seva unitat de mesura habitual és el decibel (dB), que es correspon amb una transformació logarítmica de l'amplitud del senyal per representar millor la percepció humana del so.
  - *Shimmer*: lleugera pertorbació en la intensitat deguda a fluctuacions en l'amplitud entre dos cicles consecutius.
- Propietats relacionades amb els aspectes temporals de la parla:
  - Velocitat de la parla: es pot mesurar a partir de la durada dels segments de la parla o bé com el nombre d'unitats lingüístiques per unitat temporal (per exemple, nombre de paraules per minut o nombre de síl·labes per segon).
  - Pauses: representen el nombre i la durada dels silencis en el senyal de veu[20].
- Propietats relacionades amb el timbre:
  - Energia d'alta freqüència: proporció relativa de l'energia del senyal acústic que sobrepassa una freqüència de tall respecte a l'energia total del senyal.
  - Freqüències dels formants: regions de freqüència que presenten una alta concentració d'energia espectral i que reflecteixen les ressonàncies naturals del tracte vocal. Habitualment es representen per la freqüència central de la regió i el seu ample de banda.
  - Precisió en l'articulació: mesura la desviació de les freqüències dels formants en les vocals des de les freqüències formants neutres (Juslin i Laukka, 2003).

d'un grup fònic (entenent per grup fònic la porció del discurs compresa entre dues pauses) (Garrido, 1991). No ha de confondre's amb l'entonació, atès que aquesta és un fenomen lingüístic relacionat amb la sensació perceptiva produïda per la variació de la F0, l'amplitud i la durada.

[20] Fa referència a les pauses buides (*empty pauses*, en anglès) que es realitzen per respirar. Les pauses buides difereixen de les pauses plenes (*filled pauses*, en anglès) en què en aquestes últimes sí existeix so i es relacionen amb la planificació del discurs (Puigví *et al.*, 1994).

Tant el *jitter* com el *shimmer* no s'associen a propietats prosòdiques, malgrat estar estretament relacionades amb la F0 i la intensitat, respectivament, sinó que se solen associar amb les propietats del timbre. El grup format per les propietats del timbre, el *jitter* i el *shimmer* rep el nom de *Voice Quality* (VoQ)[21] (vegeu la secció 3.3.1.1).

Murray i Arnott (1993) presenten un resum de diversos estudis per investigar la correlació existent entre parla i emoció, concloent que existeixen clars efectes vocals en algunes emocions. La taula 2.2 mostra un resum d'aquests efectes segons la reproducció d'Iriondo (2008). La descripció dels efectes està referida a un estil neutre. No obstant això, la quantificació dels efectes és bastant imprecisa.

Taula 2.2: Resum dels efectes de les emocions en la parla, traduït al català a partir de la reproducció d'Iriondo (2008) de l'original de Murray i Arnott (1993) (en anglès).

	Por	Alegria	Tristesa	Enuig	Fàstic
Velocitat de la parla	Lleugerament més ràpida	Més ràpida o més lenta	Lleugerament més lenta	Molt més ràpida	Molt més lenta
Mitjana de F0	Molt més alta	Més alta	Lleugerament més baixa	Molt més alta	Molt més baixa
Rang de F0	Més ampli	Més ampli	Lleugerament més estret	Més ampli	Lleugerament més ampli
Qualitat de la veu	Panteixant	Estrepitosa	Ressonant	Sonoritat irregular	Sorollosa
Canvis de F0	Abruptes en síl·labes tòniques	Suaus inflexions ascendents	Inflexions descendents	Normal	Amplis, inflexions descendents finals
Articulació	Tibant	Normal	Arrossegada	Precisa	Normal

La figura 2.9 il·lustra a través de quatre exemples, i sense pretendre extreure conclusions quantitatives, com l'emoció influeix en la parla. En la figura es mostra la forma d'ona i l'espectrograma de quatre fragments de veu extrets del corpus BDP-UAB (Iriondo *et al.*, 2009)[22] corresponents als estils neutre (sense emoció), agressiu, alegre i trist. Tal com s'ha exposat fins ara, es pot comprovar com, pel que fa als altres estats emocionals, la parla trista presenta silencis més perllongats i, en general, un nivell d'energia baix. A més, l'energia s'acumula en la part baixa de l'espectrograma el que indica un to menor. L'estil agressiu presenta una forma d'ona més escarpada, en comparació dels estats neutre i alegre, i també variacions més brusques en l'espectrograma. L'estil alegre, en aquest cas, s'assembla bastant a l'agressiu, i difereix del neutre, entre altres trets, en una durada bastant inferior dels silencis entre paraules.

A diferència de Murray i Arnott (1993), existeixen altres treballs que presenten descripcions més específiques com els desenvolupats per Mozziconacci (1998), que quantifica paràmetres de melodia i velocitat en forma de percentatge respecte a l'estat neutre;

[21] Qualitat de la veu, en anglès.

[22] Vegeu la secció 3.4.2 per a una descripció detallada d'aquest corpus.

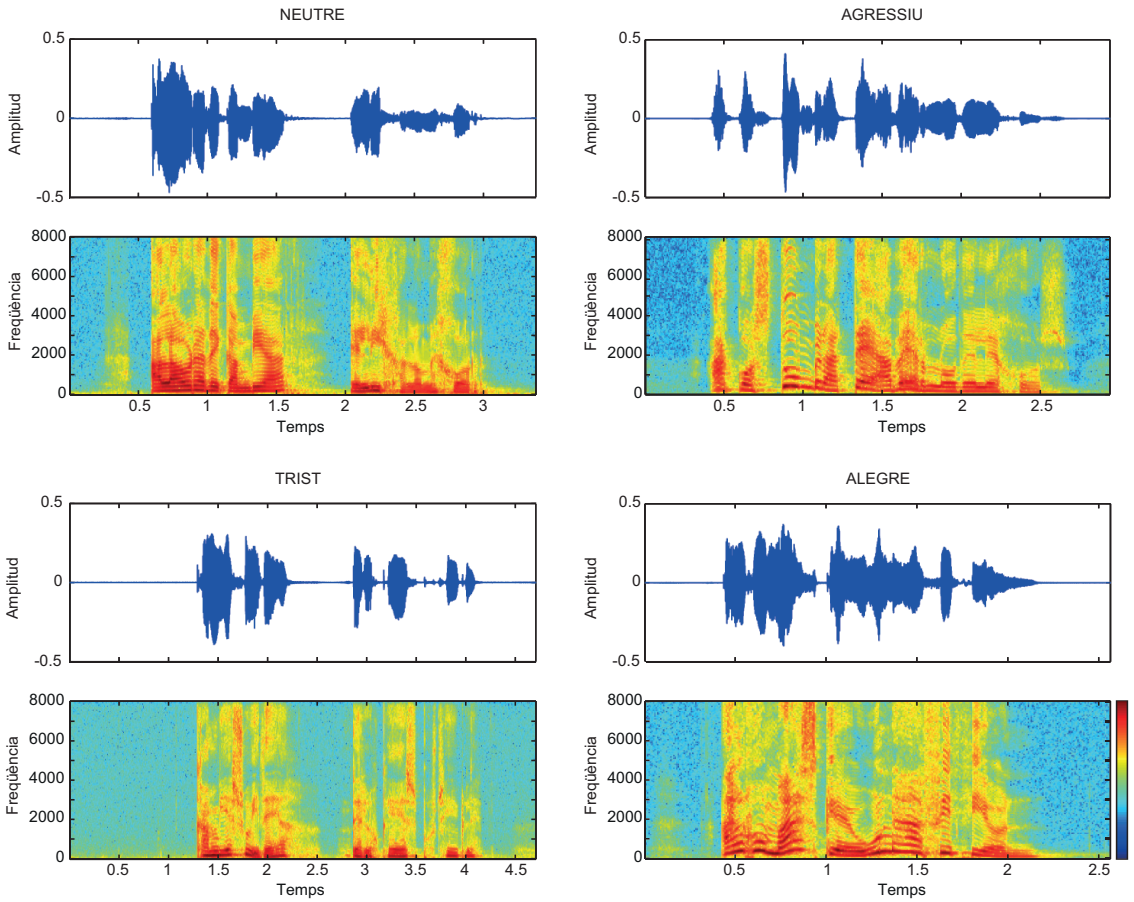


Figura 2.9: Formes d'ona i escpectrogrames de quatre locucions parlades procedents del corpus BDP-UAB. Cada locució es correspon amb un estil neutre, agressiu, alegre i trist (en el sentit de les agulles del rellotge).

Rodríguez *et al.* (1999), que presenten un model per a l'expressió emocional de la parla en castellà amb l'objectiu de millorar la naturalitat de la parla sintètica en sistemes de conversió de text a veu amb requeriments emocionals específics, sent validat posteriorment per Iriondo *et al.* (2000); Cowie *et al.* (2001), que inclouen una anàlisi de 14 estats emocionals caracteritzats per una descripció qualitativa de les característiques de la parla organitzades en cinc categories: acústica, contorn melòdic, to, VoQ i altres, recopilant les conclusions de diversos autors; o Juslin i Laukka (2003), que presenten una anàlisi de 104 estudis relacionats amb l'expressió vocal i 41 estudis sobre la interpretació musical per comprovar si ambdues modalitats transmeten emocions de forma similar, estudiant cinc categories emocionals i comprovant que les emocions de tristesa i enuig són les millor descodificades (també a nivell transcultural) mentre que, a diferència del que succeeix en estudis d'expressivitat facial, l'alegria és la menys identificada entre diferents cultures.

2.3.4 La percepció musical

Malgrat que la música comporta un potent contingut emocional, expressar emocions a través d'ella no és senzill atès que la seva percepció depèn de la cultura, les habilitats de l'intèrpret i de l'edat de l'oïdor (Bartneck, 2000). Scherer i Oshinsky (1977) argumenten que la major part de les variacions emotives de la música es poden crear per modificacions d'amplitud, melodia (nivell, variació i contorn), *tempo*, corba i filtrat, i afirmen, a més, que existeix un paral·lelisme entre aquesta afirmació i el que succeeix en l'expressió emocional en la parla. Juslin (1997) recopila una sèrie de paràmetres musicals relacionats amb cinc emocions i, posteriorment, analitza la similitud entre les propietats acústiques de l'expressió vocal i les de la interpretació musical (Juslin i Laukka, 2003). Les seves conclusions es resumeixen en la taula 2.3 segons la reproducció d'Iriondo (2008).

Taula 2.3: Resum de les propietats acústiques de l'expressió vocal i la interpretació musical de quatre emocions segons la reproducció de Iriondo (2008) de l'original de Juslin i Laukka (2003).

Emoció	Propietats acústiques (expressió vocal/interpretació musical)
Enuig	Velocitat/*tempo* ràpida/-, intensitat/nivell de so forta/-, alta variabilitat d'intensitat/nivell de so, alta energia d'alta freqüència, alt nivell de F0/to, alta variabilitat de F0/to, contorn ascendent de F0/to, ràpid inici de veu/atac
Por	Velocitat/*tempo* ràpida/-, intensitat/nivell de so baixa/- (excepte en pànic), alta variabilitat d'intensitat/nivell de so, baixa energia d'alta freqüència, alt nivell de F0/to, poca variabilitat de F0/to, contorn ascendent de F0/to
Alegria	Velocitat/*tempo* ràpida/-, intensitat/nivell de so mitjana/à a fort, valor mig d'energia d'alta freqüència, alt nivell de F0/to, alta variabilitat de F0/to, contorn ascendent de F0/to, ràpid inici de veu/atac
Tristesa	Velocitat/*tempo* lenta/-, intensitat/nivell de so baixa/-, poca variabilitat d'intensitat/nivell de so, petita energia d'alta freqüència, baix nivell de F0/to, poca variabilitat de F0/to, contorn descendent de F0/to, inici lent de veu/atac

2.4 Resum

En aquest capítol s'han explicat les principals teories de l'emoció indicant la funció de les mateixes en el procés comunicatiu. Comparant les teories de diversos autors s'aprecia un element comú que mostra la "tendència a l'acció" vinculada al concepte "emoció". No obstant això, i a pesar que l'emoció és un pilar fonamental de la comunicació, és difícil aconseguir una única descripció per a aquest concepte i, molt menys, determinar un conjunt específic d'emocions bàsiques. A grans trets, existeixen llistats d'emocions que creen conjunts d'etiquetes categòriques per descriure cadascuna de les emocions que consideren; també existeixen els models *circumplex* que descriuen les relacions entre les emocions de manera anàloga als colors d'una roda cromàtica; i els models multidimensionals que estableixen un espai de dimensionalitat reduïda per definir les emocions humanes.

Una tasca important quan es treballa amb elements que han de ser descrits sobre la base de conceptes emocionals és l'etiquetatge previ dels mateixos. En funció del model escollit per representar les emocions estudiades es pot realitzar un etiquetatge categòric o un etiquetatge dimensional. Si bé el primer pot ser molt senzill, en augmentar el nombre de categories emocionals el procés pot resultar complicat. L'etiquetatge dimensional pretén solucionar aquest inconvenient treballant en un espai continu però, al mateix temps, requereix d'un coneixement previ per part de l'avaluador de certs fonaments teòrics relacionats amb les emocions. No obstant això, algunes eines com SAM pretenen pal·liar aquest inconvenient.

Més enllà dels models escollits per a la definició, descripció i representació de les emocions la veritat, i resulta clarament observable, és que van lligades a reaccions fisiològiques en l'organisme. Sigui quina sigui la funció que exerceixin —la qual cosa, tal com s'ha vist en aquest capítol, varia segons l'autor consultat—, aquestes reaccions fisiològiques incideixen en els diversos aspectes de la comunicació, entenent-la com una activitat multimodal que involucra diversos components i sentits. La parla és un d'aquests components i, a més, és un dels quals que per si sols poden transmetre emocions en un oïdor. De manera similar a la música, la parla presenta una sèrie de variacions en els seus paràmetres que evolucionen en funció de l'estat emocional del parlant.

Capítol 3

Corpus i parametrització

Al llarg d'aquest capítol es descriuen quines són les consideracions generals per crear un corpus de veu amb contingut emocional i quins tipus d'expressions vocals poden considerar-se per al seu disseny, des de la veu natural (la més realista) a l'actuada (la que presenta més facilitats per al control de l'enregistrament). També s'expliquen els diferents paràmetres que poden ser extrets de les locucions dels corpus per a la seva posterior anàlisi mitjançant algorismes d'aprenentatge automàtic, tant a nivell acústic com a nivell lingüístic. Finalment, en l'última secció del capítol, es descriuen els tres corpus que seran analitzats en aquesta tesi i es detallen els paràmetres que s'extreuen dels mateixos.

3.1 Consideracions generals

Gravar un corpus de veu no és una tasca senzilla. Més enllà d'una simple recopilació d'enregistraments de veu, la creació d'un corpus de veu amb contingut emocional està subjecta a diferents consideracions. Douglas-Cowie *et al.* (2003) destaquen, segons cita Iriondo (2008), quatre aspectes importants que s'han de tenir en compte:

- Àmbit. Es tracta de considerar què cobreix el conjunt d'enregistraments quant a nombre de locutors, edat, sexe, idioma, estats emocionals, etc. Aquest aspecte és de crucial rellevància per a la generalització dels estudis posteriors i depèn, en part, de l'aplicació a la qual va orientada el corpus. Així, per exemple, en un corpus de síntesi de veu pot ser suficient un únic locutor mentre que per a un de reconeixement es necessita una major varietat d'enregistraments. Quant a les emocions considerades, escollir emocions plenes o estats emocionals de diversa intensitat (vegeu la secció 2.1.1) depèn de l'autor de cada estudi i no existeix un clar consens sobre aquest tema.
- Naturalitat. La naturalitat dels enregistraments depèn de com es realitzin. Es pot optar per escollir un actor o una actriu que representin estats emocionals concrets o

bé recopilar enregistraments d'emocions espontànies. Per descomptat, l'enregistrament d'emocions actuades suposa un major control de les condicions d'enregistrament però la seva autenticitat pot quedar en dubte (vegeu la secció 3.2).

- Context. Complementa l'estat emocional que el receptor percep. Es diferencien quatre formes bàsiques de context:
  - Semàntic: el contingut emocional es troba en determinades paraules.
  - Estructural: les emocions les determinen patrons específics de pronunciació.
  - Intermodal: l'expressió emocional es manifesta en l'expressió facial, els gestos i la postura.
  - Temporal: les emocions es reconeixen en els canvis acústics que es produeixen en determinats instants de temps.
- Descriptors. Han de representar el contingut lingüístic i l'emocional, cobrint totes les característiques que es relacionen amb l'expressió vocal dels estats emocionals. En funció de la naturalesa de cadascuna d'aquestes característiques es pot optar per una representació qualitativa o quantitativa.

A més, existeix una estreta relació entre els estudis de parla i emoció i les aplicacions en les quals es pretenen utilitzar. Així doncs, els quatre factors abans detallats hauran de ser els adequats per a l'aplicació a la qual es destini el corpus. Schröder (2004) il·lustra el procés d'inferència d'emocions entre els dos agents d'una comunicació, tal com es mostra en la figura 3.1, diferenciant dos tipus d'estudis segons se centrin en l'expressió emocional per part del parlant o en la percepció de les emocions per part de l'oient:

- Estudis que se centren en el parlant: es refereixen, en general, al reconeixement d'emocions del locutor a partir de l'anàlisi del senyal de veu sobre la base d'un conjunt de paràmetres quantificables de la parla. El gran repte, segons Devillers *et al.* (2005), és destriar quins són els paràmetres realment atribuïbles a l'emoció i no a les pròpies característiques de la conversa.
- Estudis que se centren en l'oient: es refereixen al modelatge dels paràmetres de la parla perquè transmeti un determinat estat emocional. Per descomptat, resulta fonamental la descripció que es faci de cadascun d'aquests estats emocionals i és crucial determinar quins són aquests paràmetres abans de procedir al seu estudi.

Així, els corpus destinats a síntesis de veu i, per tant, centrats en l'oient, acostumen a ser corpus extensos, de diverses hores de durada, amb l'enregistrament realitzat per un actor o una actriu professional que llegeixen un conjunt de textos. Aquests textos poden ser neutres (sense contingut emocional en la seva semàntica), els mateixos per a cada emoció, o bé poden ser textos amb contingut emocional. Els primers fan més fàcil la comparació atès que l'única diferència és la component emocional en el senyal de veu,

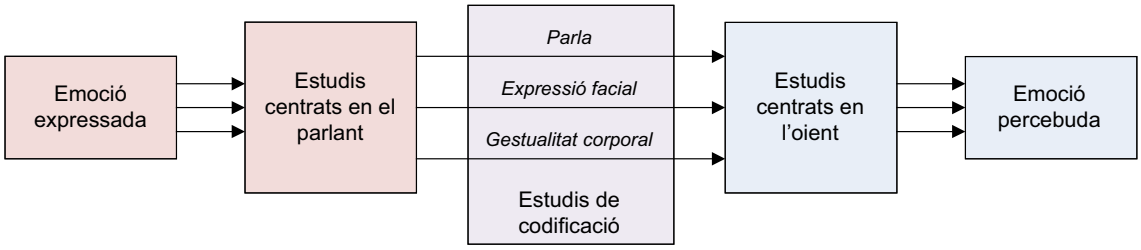


Figura 3.1: Classificació dels estudis sobre parla i emoció segons el focus d'interès. Adaptat d'(Iriondo, 2008).

que és precisament la que s'està analitzant. Per contra, els segons faciliten la tasca de locució dels actors en el moment de l'enregistrament Iriondo (2008).

No obstant això, els corpus destinats a reconeixement emocional i, per tant, centrats en el parlant, acostumen a ser d'una naturalesa més diversa, com pot comprovar-se en la varietat d'estudis de reconeixement afectiu existents (vegeu el capítol 4).

3.2 Mètodes per a la generació de parla emocional

Existeixen diversos mètodes per registrar fragments de parla emocionada. Evidentment, recórrer a actors perquè simulin un estil específic de parla sembla el més senzill però diverses qüestions poden posar en dubte la naturalitat aconseguida mitjançant aquest sistema. Encara que l'ús d'actors és un escenari ideal en tant que ofereix el màxim control del contingut del corpus i de les condicions d'enregistrament, Douglas-Cowie *et al.* (2005) assenyalen que es perd informació rellevant que sí està present en escenaris reals i quotidians. Això pot ocasionar una degradació en els resultats de sistemes de reconeixement automàtic en el moment d'aplicar els models extrets en un escenari real (Grimm *et al.*, 2007). És per això que existeixen diversos mètodes per generar veu emocionada tenint en compte sempre els factors de naturalitat i control de l'enregistrament. Campbell (2000) proposa quatre estratègies per generar parla amb una determinada coloració emocional, les quals també se citen per Schröder (2004) i Steidl (2009), i es descriuen en les seccions següents.

3.2.1 Expressió vocal natural

Les locucions naturals són, en principi, les ideals per recopilar un corpus de veu emocionada. Aquestes locucions poden procedir de fonts diverses com, per exemple, situacions de risc per a pilots d'aviació, entrevistes de ràdio i televisió o cròniques periodístiques (Scherer, 2003). No obstant això, l'ús d'aquest material moltes vegades es veu obstaculitzat per motius de drets d'autor i altres aspectes legals. A més, sorgeixen altres problemes com són l'escassa qualitat d'aquests enregistraments a causa de les distàncies

variables del locutor al micròfon o a les reverberacions, al reduït nombre de locutors, a l'escàs control del contingut i a la dificultat de l'etiquetatge. A tot aquest escenari desfavorable s'afegeix la *Paradoxa de l'Observador* (Steidl, 2009) enunciada per William Labov, la qual indica que "el propòsit de la recerca lingüística ha de ser descobrir com s'expressa la gent quan no està sent observada sistemàticament, a pesar que l'única manera d'obtenir aquesta informació és mitjançant l'observació sistemàtica" (Labov, 1972). Així doncs, s'hauria de proporcionar major llibertat als locutors per realitzar aquests enregistraments, fins i tot deixant fora del seu coneixement el fet d'estar sent gravats, la qual cosa resultaria en una pèrdua encara major del control de l'enregistrament. A més, tal com expressa Steidl (2009), aquest tipus d'enregistrament és altament qüestionable des d'un punt de vista ètic i fins i tot conflictiu a nivell legal.

Alguns corpus d'aquesta categoria són: *The Reading/Leeds Emotional Speech Corpus* (Arnfield *et al.*, 1995); *The Belfast Naturalistic Emotion Database*, descrit i analitzat per Schröder (2004); la base de dades JST/CREST (Campbell, 2002) i el corpus VAM (Grimm *et al.*, 2007).

3.2.2 Expressió vocal induïda

Les tècniques d'inducció d'emocions intenten modificar l'estat emocional del locutor. Gerrards-Hesse *et al.* (1994) analitzen 250 estudis relacionats amb procediments per a la inducció d'estats d'humor[1] i proposa una classificació dels mateixos en cinc categories:

- Procediments basats en la generació lliure d'estats emocionals. No es presenten estímuls als individus sinó que ells mateixos activen els estats afectius mitjançant tècniques hipnòtiques, per exemple, o mitjançant el record de determinades experiències anteriors.
- Procediments basats en la generació guiada d'estats emocionals. Es presenten determinats estímuls als individus i se'ls indica què estat emocional han d'aconseguir, per exemple a través de fragments de pel·lícules o peces musicals.
- Procediments basats en la presentació de material per a la inducció d'emocions. Es presenten determinats estímuls als individus però sense indicar-los l'estat emocional que han d'aconseguir.
- Procediments basats en la presentació de situacions de necessitat emocional. Els individus s'exposen a situacions que generen necessitat d'assoliment o afiliació, com per exemple activitats cognitives avaluades o determinades interaccions socials.
- Procediments per generar estats psicològics amb emocions rellevants. Els individus se sotmeten a tractaments que varien els seus nivells fisiològics com, per exemple,

[1] En anglès, MIPs, acrònim de *Mood Induction Procedures*.

rebent alguna substància tal com adrenalina o un placebo, o bé forçant, a indicació de l'investigador, alguns dels seus músculs facials[2] per provocar una emoció.

No obstant això, molts d'aquests procediments no s'han dissenyat específicament per provocar emocions que afectin al senyal de veu, encara que poden ser adaptats per a tal efecte. Per gravar veu existeixen mètodes específics tals com realitzar una tasca complicada de lletrejar per provocar un estat negatiu, una tasca de càlcul mental per crear una situació d'estrès o un escenari *Mag d'Oz*[3] amb la simulació d'un dispositiu defectuós per provocar una sensació d'empipament (Steidl, 2009).

Els principals inconvenients són que sovint no resulta senzill evocar emocions concretes ja que la resposta individual de cada subjecte a un mateix estímul pot ser diferent i, a més, les emocions induïdes solen ser de baixa intensitat i no emocions plenes. Sense deixar de costat que el contingut de les locucions, per la seva naturalesa espontània, no pot ser controlat. Finalment, induir certes emocions amb la finalitat de recopilar informació científica pot provocar controvèrsies ètiques i morals (Campbell, 2000).

Alguns corpus d'aquesta categoria són els desenvolupats per Bachorowski i Owren (1995); per Fernández i Picard (2003); el corpus FAU Aibo, descrit per Steidl (2009) i el corpus EmoTaboo de Zara *et al.* (2007).

3.2.3 Expressió vocal estimulada

En aquest cas, els actors reben una sèrie de textos amb contingut adequat per a l'emoció que se'ls demana expressar amb l'objectiu d'estimular emocions en la seva veu. Pretén ser un sistema de compromís entre la rigidesa d'un sistema basat en actors (explicat en la següent secció) i els mètodes anteriors, que posseïen un menor control del contingut de les locucions i de les condicions d'enregistrament. No obstant això, per segons quines aplicacions, pot ser contraproduent comptar amb locucions de diferent contingut semàntic per a cada emoció, sobretot en el moment de fer determinades comparacions entre els estils emocionals.

Dos exemples de corpus d'aquesta categoria són la *Belfast Structured Emotion Database* (Douglas-Cowie *et al.*, 2003) i el corpus BDP-UAB (Iriondo *et al.*, 2009).

[2] La hipòtesi de retroalimentació facial (Leventhal, 1980) indica que les expressions facials influeixen en l'estat d'humor.

[3] Dur a terme un determinat exercici en un escenari *Mag d'Oz* vol dir que els subjectes sotmesos a estudi creuen interactuar amb un sistema completament automàtic encara que, en realitat, les respostes del sistema estan controlades per un operador humà que roman ocult durant el mateix.

3.2.4 Expressió vocal actuada

El sistema que permet el major control del contingut fonètic i semàntic de les locucions així com de les condicions d'enregistrament és el consistent a gravar amb actors sobre la base d'un guió preestablert. Permet l'enregistrament d'emocions plenes, prototípiques i de màxima intensitat. Les condicions d'enregistrament poden ser exactes a les especificacions de l'experiment atès que es treballa en estudi, en un entorn completament controlat. Els textos poden ser exactament els mateixos per a cada estil emotiu afavorint la seva posterior comparació, a pesar que això pugui dificultar l'expressió a locutors no especialment entrenats. Posteriorment, i igual que en els procediments anteriors, un test subjectiu pot permetre eliminar els enregistraments que no es corresponguin amb l'estil especificat a l'inici.

No obstant això, aquest sistema torna a plantejar alguns inconvenients. Steidl (2009) recull alguns dels problemes que citen diversos autors. El principal és la falta de naturalitat atès que els actors poden caure en l'exageració de certs aspectes de l'expressió emocional que, en la vida real, tendirien a dissimular-se o passarien desapercebuts. D'altra banda, en les interaccions socials es tendeix a dissimular alguns aspectes emocionals personals mitjançant el control dels paràmetres expressius, aspecte que difícilment es reflectirà en una locució actuada. Al mateix temps, a pesar que un oient pugui percebre l'emoció pretesa, és possible que l'emoció, al no ser sentida, es percebi com a falsa, la qual cosa li provocarà una reacció desagradable. Finalment, les emocions actuades no provenen de reaccions fisiològiques i, per tant, difícilment els enregistraments mostraran els matisos reals de l'estat emocional pretès.

Malgrat l'abans comentat, Busso i Narayanan (2008) assenyalen que el problema no sempre està en l'ús d'actors per realitzar els enregistraments sinó en el mètode emprat i en l'entorn d'enregistrament. Així, unes pautes ben definides i una metodologia explícita poden permetre reduir les diferències observades entre les condicions de laboratori i les aplicacions de la vida real.

Alguns exemples de corpus d'aquesta categoria són: el corpus *Talkapillar* (Beller i Marty, 2006), la *Berlin Database of Emotional Speech* (Burkhardt *et al.*, 2005) i la base de dades desenvolupada per Liberman *et al.* (2002).

3.3 Parametrització

Per tal que els algorismes d'aprenentatge puguin treballar amb els arxius d'àudio d'un corpus és necessari que aquests siguin parametritzats. Mitjançant el procés de parametrització cada arxiu d'àudio passa a estar descrit per una sèrie de valors numèrics i etiquetes addicionals que formen, en conjunt, un únic vector. Aquestes etiquetes addicionals són descriptives de cada locució i la seva definició varia en funció del corpus. La seva funcionalitat, en general, és la d'identificar cadascuna de les locucions mitjançant un codi

i descriure-les emocionalment. Aquesta descripció pot ser una única etiqueta categòrica o les coordenades que l'emoció representa en un espai multidimensional, per exemple. Aquestes etiquetes addicionals s'assignen manualment durant el disseny del corpus. La resta prové d'un procés automàtic o semiautomàtic mitjançant el qual s'extreuen una sèrie de paràmetres numèrics que descriuen diferents característiques de cadascun dels arxius que formen part del corpus.

És important assenyalar que la parametrització es realitza de manera que cada locució, independentment de la seva durada, resulti en un vector de paràmetres de longitud constant. És a dir, es pretén uniformitzar la descripció paramètrica de les locucions sense que importi la durada de les mateixes. El vector resultant de la parametrització de cada arxiu del corpus rep el nom indistint d'instància, exemple o cas. Cadascun dels elements d'aquests vectors rep el nom de paràmetre o atribut. El conjunt de vectors que recull la parametrització de tots i cadascun dels arxius d'àudio i, per tant, la parametrització completa del corpus, rep el nom de conjunt de dades o, segons la seva nomenclatura en anglès, *dataset*. Així doncs, el conjunt de dades resultant de la parametrització d'un corpus contindrà tantes instàncies com a arxius d'àudio formin part del corpus i tants atributs com a paràmetres s'hagin calculat més les etiquetes addicionals que s'incloguin posteriorment.

La definició del vector d'atributs respon al tipus de parametrització que es dugui a terme. En aquesta tesi la parametrització es realitza a dos nivells: a nivell acústic i a nivell lingüístic, els quals es descriuen a continuació.

3.3.1 Parametrització acústica

En el passat, el focus d'estudi dels paràmetres acústics extrets d'un corpus se centrava en la seva anàlisi exclusivament prosòdica, emprant-se conjunts de dades relativament petits i, sovint, incomplets (Cowie *et al.*, 2001). No obstant això, a mesura que els estudis en reconeixement emocional automàtic han anat avançant i la capacitat de càlcul dels dispositius ha augmentat, aquesta parametrització ha anat fent-se més extensa (Schuller *et al.*, 2009).

De forma general, el procés de parametrització acústica d'un arxiu d'àudio consisteix a dividir el senyal acústic en una sèrie de segments que posteriorment són analitzats de manera individual. Aquesta divisió pot realitzar-se a dos nivells diferents que no són excloents:

- Divisió a nivell segmental. Consisteix a dividir cada arxiu d'àudio en segments[4] de longitud constant (Oudeyer, 2003; Planet *et al.*, 2006; Steidl, 2009) mitjançant l'aplicació d'una finestra. Aquests fragments solen durar entre 10 i 20 ms, temps durant el qual es pot considerar que el senyal de veu no presenta grans variacions.

[4]També és comú referir-se a aquests *segments* amb el terme *trames*. En aquesta tesi tots dos termes s'empren de forma indistinta.

- Divisió a nivell fonètic. Consisteix a crear segments del senyal de veu que delimitin els fonemes[5] de la locució (Monzo *et al.*, 2007; Ringeval i Chetouani, 2007; Iriondo *et al.*, 2009). Habitualment s'empren mètodes semiautomàtics per realitzar aquesta tasca: un sistema automàtic realitza una primera segmentació que posteriorment és revisada i ajustada manualment. Aquest mètode sol ser útil per estudiar fragments concrets de la locució de forma aïllada de la resta, com per exemple les vocals o les consonants.

Cada fragment es processa per extreure una sèrie de descriptors que reben el nom de descriptors de baix nivell[6] ja que estan directament relacionats amb el senyal acústic en si mateix. La secció 3.3.1.1 conté una relació dels descriptors més emprats habitualment en treballs de reconeixement emocional i que han estat utilitzats en diferents experiments d'aquesta tesi. Per fer més exhaustiva la parametrització (Oudeyer, 2003) i per obtenir una visió de la dinàmica d'aquests descriptors més enllà dels seus valors instantanis (Steidl, 2009; Bozkurt *et al.*, 2011), es realitza el càlcul la primera i segona derivada discreta dels mateixos. La derivada discreta de la funció $f(n)$ es representa pel símbol $\Delta_n f(n)$ i s'expressa matemàticament com indica l'equació 3.1.

$$\Delta_n f(n) = f(n+1) - f(n) \tag{3.1}$$

A continuació cada arxiu d'àudio és processat per extreure una sèrie de funcionals[7] a partir dels descriptors de baix nivell que s'han calculat en el pas anterior. Busso *et al.* (2009) assenyalen que, a nivell de freqüència, aquests funcionals són més idonis que els paràmetres descriptors de la corba de F0, així com també ho és l'anàlisi a nivell de frase en lloc de segments més breus (com, per exemple, únicament fragments sonors). Igual que en altres treballs, el càlcul de funcionals a nivell de frase es farà extensiu a la resta dels paràmetres analitzats (Litman i Forbes-Riley, 2003; Oudeyer, 2003; Vlasenko *et al.*, 2007; Iriondo *et al.*, 2009; Schuller *et al.*, 2009; Steidl, 2009; Sezgin *et al.*, 2012). Els funcionals més emprats de forma habitual en tasques de reconeixement emocional són: mitjana (primer moment estàndard), desviació estàndard (relacionat amb el segon moment estàndard), asimetria[8] (tercer moment estàndard), curtosis[9] (quart moment estàndard), mitjana, extrems (valors màxim i mínim), posició dels extrems, diferència entre els extrems, quartils, rang interquartílic[10] i coeficients de regressió lineal. Aquesta no és una llista exhaustiva i l'elecció dels funcionals depèn de l'autor. Eyben *et al.* (2009) recullen una llista extensa de funcionals orientats a reconeixement emocional que, a més, poden ser calculats amb

[5] Unitats fonològiques mínimes amb funció distintiva.

[6] De l'anglès, *Low Level Descriptors*.

[7] Funcions que transformen un vector en un valor escalar.

[8] És habitual la seva nomenclatura en anglès, *skewness*. Representa l'asimetria d'una distribució considerant com a eix de simetria una recta vertical que creua el valor mitjà de la mateixa.

[9] És habitual la seva nomenclatura en anglès, *kurtosis*. És una mesura de la forma de la distribució i indica el seu apuntament.

[10] El rang interquartílic es defineix com la diferència entre el tercer i primer quartil.

l'eina OpenEAR[11], de lliure distribució.

Finalment, els funcionals que descriuen cada arxiu d'àudio es concatenen amb les etiquetes addicionals. Cadascun dels vectors resultants forma una instància del conjunt de dades corresponent a la parametrització acústica del corpus. La figura 3.2 il·lustra el procés de creació del conjunt de dades d'un corpus a nivell acústic tal com s'ha descrit en aquesta secció, considerant el càlcul de n descriptors de baix nivell, la seva primera derivada i m funcionals, incloent un codi d'identificació de la locució i una etiqueta descriptiva de l'emoció a nivell categòric.

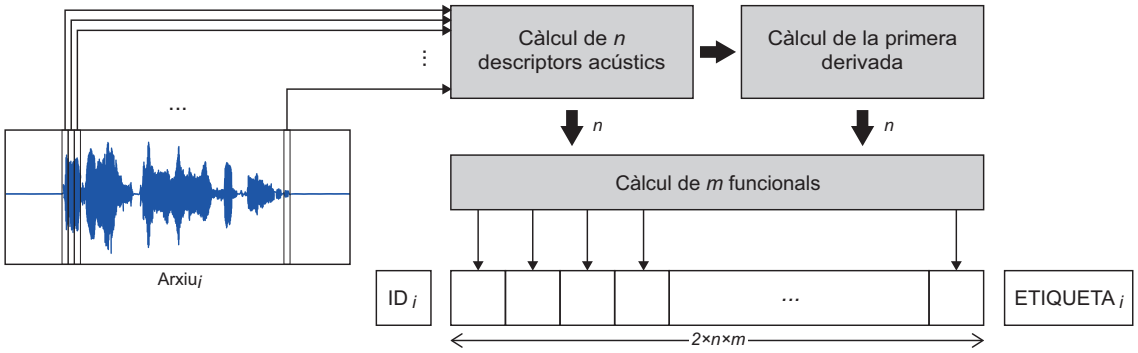


Figura 3.2: Esquema de la parametrització acústica d'un arxiu d'àudio d'un corpus calculant n descriptors de baix nivell, la seva primera derivada i m funcionals.

3.3.1.1 Descriptors de baix nivell

Els paràmetres de baix nivell que sovint es consideren relacionats amb l'expressivitat emocional i que han estat estudiats en diversos experiments de la present tesi es poden agrupar en diferents categories, segons es relacionin amb la prosòdia, la VoQ o l'espectre ampli del senyal. Alguns d'aquests paràmetres es troben explicats des d'una perspectiva centrada en l'expressió i la percepció emocional en la secció 2.3.3.

Paràmetres relacionats amb la prosòdia. La prosòdia caracteritza els trets suprasegmentals percebuts en la parla i per tant es refereix a síl·labes, paraules, frases o locucions senceres en comptes d'únicament a fonemes. La prosòdia es refereix a la percepció acústica dels sons, els quals defineix pel seu to, volum[12] i durada. Segons Gil (2007), el to és la impressió perceptiva que produeix en l'oient la F0 de l'ona sonora; el volum és la força percebuda en els sons influïda tant per la intensitat de l'ona sonora com per la F0, les característiques espectrals i la durada del so; i la durada és la longitud percebuda d'un so, influenciada per canvis en el volum. A pesar que aquestes característiques perceptives no tenen equivalents únics en el senyal de veu existeixen equivalents acústics que s'adapten molt adequadament com la F0, que s'assembla al to percebut, i l'energia del senyal,

[11] http://openart.sourceforge.net
[12] En anglès, *loudness*; en espanyol, *sonía*.

que s'assembla al volum. S'agrupen en tres categories segons la seva relació amb la F0, l'energia o la durada:

- Paràmetres relacionats amb la F0. Poden calcular-se en tots els segments en els quals s'ha fragmentat l'arxiu de veu encara que realment té sentit en aquelles zones que presenten una periodicitat clara, és a dir, en els sons sonors com, per exemple, les vocals. Els fragments que no presenten periodicitat poden ser omesos o ben processats de manera especial. Així, per exemple, el marcador desenvolupat per Alías *et al.* (2006) assigna marques interpolades en base a les zones sonores veïnes. Si no es realitza una segmentació fonètica, Navas *et al.* (2006) proposen l'ús d'un detector d'activitat vocal i un detector de sonoritat per realitzar el càlcul de forma adequada.
- Paràmetres relacionats amb l'energia. Com el seu propi nom indica, mesuren l'energia del fragment acústic analitzat. Es pot mesurar en decibels o pel seu valor quadràtic mitjà[13].
- Paràmetres relacionats amb el ritme. Tenen sentit en la segmentació fonètica de la parla i mesuren la durada dels fonemes de la locució. Malgrat ser rellevants en l'expressió de les emocions (Cowie *et al.*, 2001), sovint no es consideren en alguns estudis per la complexitat que suposa obtenir aquesta informació de manera automàtica (Navas *et al.*, 2006). Habitualment s'empra la mesura estadística *z-score* per analitzar l'estructura temporal de la parla, com en (Schweitzer i Möbius, 2003), segons l'equació 3.2 on μ i σ són la mitjana i la desviació estàndard, respectivament, del segment analitzat i d la durada, en mil·lisegons, del fonema (Iriondo, 2008). Així mateix, en (Iriondo, 2008) també es calcula el nombre de pauses per unitat de temps i el percentatge de temps de silenci respecte a la durada total de la locució amb l'objectiu de representar la freqüència i la durada de les mateixes.

$$z\text{-}score = \frac{d - \mu}{\sigma} \tag{3.2}$$

Paràmetres relacionats amb la VoQ. La VoQ es basa en les pertorbacions en la F0 i l'energia del senyal que es representen per diferents paràmetres. Segons la definició de Laver (1980) recollida en (Monzo, 2010), "La VoQ es concep en un sentit ampli com la característica auditiva de la veu individual d'un interlocutor, i no en el sentit més reduït de la qualitat derivada únicament de l'activitat laríngia. Tant les característiques laríngies i supralaríngies seran vistes com a contribucions a la VoQ." Sobre la base dels paràmetres proposats per Drioli *et al.* (2003), els resultats obtinguts per Monzo *et al.* (2007) i els proposats per Iriondo (2008), es poden destacar els següents:

- *Jitter*. Mesura les variacions a curt termini del període fonamental degudes a fluctuacions en els temps d'obertura i tancament dels plecs vocals d'un cicle al següent,

[13] Mesurat com *rms*, de l'anglès *root pixen square*.

descrivint un soroll que apareix en forma de modulació en freqüència. A pesar que existeixen diverses variants en el càlcul d'aquest paràmetre, la més senzilla i que millor s'adapta a la definició d'aquest paràmetre es descriu en l'equació 3.3.

$$jitter(i) = \frac{|F_0(i+1) - F_0(i)|}{F_0(i)} \tag{3.3}$$

- *Shimmer*. Mesura les variacions a curt termini de l'amplitud de la forma d'ona entre cicles, descrivint un soroll que apareix com una modulació en amplitud. Tot i que, com en el cas del *jitter*, existeixen diverses variants per al seu càlcul, el *shimmer* es pot descriure de forma senzilla mitjançant l'equació 3.4 en funció de l'energia (E) entre cicles consecutius.

$$shimmer(i) = \frac{|E(i+1) - E(i)|}{E(i)} \tag{3.4}$$

- *Harmonic-to-Noise Ratio* (HNR). Es defineix com la relació entre l'energia de la part harmònica i l'energia de la resta del senyal de veu i ve a mesurar la periodicitat del senyal sonor. Aquesta mesura pot veure's afectada pel *jitter* i el *shimmer*. Steidl (2009) descriu el càlcul d'aquest paràmetre segons la funció d'autocorrelació.

- *Glottal-to-Noise Excitation Ratio* (GNE). Quantifica la relació entre l'excitació deguda a oscil·lacions dels plecs vocals respecte l'excitació produïda per soroll turbulent (Michaelis *et al.*, 1997). Similar al HNR en tant que mesura el soroll additiu però difereix d'aquest paràmetre en què GNE es pot considerar pràcticament independent de *jitter* i *shimmer*.

- *Hammamberg Index* (HammI). És la diferència entre els màxims d'energia de les bandes freqüencials de 0 a 2000 Hz i de 2000 a 5000 Hz. L'equació 3.5 defineix el càlcul d'aquest paràmetre en dB on $E_{0...2000}$ és l'energia de la banda freqüencial de 0 a 2000 Hz i $E_{2000...5000}$ és l'energia de la banda freqüencial de 2000 a 5000 Hz.

$$\text{HammI} = 10 \cdot \log \frac{\text{màx}(E_{0...2000})}{\text{màx}(E_{2000...5000})} \tag{3.5}$$

- *Spectral Flatness Measure* (SFM). És la relació entre la mitjana geomètrica i la mitjana aritmètica de la distribució espectral d'energia. L'equació 3.6 defineix el càlcul d'aquest paràmetre en dB, on N és el nombre de mostres espectrals preses en consideració per a freqüències situades entre 0 Hz i la meitat de la freqüència de mostreig de l'arxiu d'àudio i E_i és l'energia de cadascuna d'aquestes freqüències.

$$\text{SFM} = 10 \cdot \log \frac{\sqrt[N]{\prod_{i=1}^{N} E_i}}{\frac{1}{N}\sum_{i=1}^{N} E_i} \tag{3.6}$$

- *Drop-off of spectral energy above 1000Hz* (do1000). És una aproximació lineal del pendent espectral per sobre de 1000 Hz calculada amb el mètode dels mínims quadrats. L'equació 3.7 defineix el càlcul d'aquest paràmetre on N és el nombre de mostres espectrals preses en consideració per a freqüències situades en el marge de 1000 Hz fins a la meitat de la freqüència de mostreig de l'arxiu d'àudio, E_i és l'energia de cadascuna d'aquestes freqüències, $\overline{E}$ és la mitjana de l'energia calculada considerant totes les freqüències, $f1000_i$ són les freqüències superiors a 1000 Hz i $\overline{f1000}$ és el valor mitjà les freqüències superiors a 1000 Hz.

$$\text{do1000} = \frac{\sum_{i=1}^{N}(E_i - \overline{E})(f1000_i - \overline{f1000})}{\sum_{i=1}^{N}(f1000_i - \overline{f1000})^2} \tag{3.7}$$

- *Relative Amount of Energy above 1000 Hz* (pe1000). Mesura la quantitat d'energia relativa en el marge de les freqüències superiors a 1000 Hz respecte a les inferiors. L'equació 3.8 defineix el càlcul d'aquest paràmetre en dB on E_h és l'energia de cadascuna de les freqüències compreses entre 1000 Hz i la meitat del valor de la freqüència de mostreig de l'arxiu d'àudio i E_l és l'energia de les freqüències inferiors a 1000 Hz.

$$\text{pe1000} = 10 \cdot \log \frac{\sum_{h=1}^{H} E_h}{\sum_{l=1}^{L} E_l} \tag{3.8}$$

Paràmetres espectrals. Segons Steidl (2009), si els paràmetres prosòdics es refereixen a la F0, l'energia i la durada, els paràmetres espectrals descriuen les característiques del senyal de veu en el domini de la freqüència estudiant també els harmònics i els formants. Els harmònics són múltiples de la freqüència fonamental i es caracteritzen per la seva freqüència i la seva amplitud. Per la seva banda, els formants són amplificacions de determinades freqüències com a resultat de les ressonàncies del tracte vocal, i es caracteritzen per la seva freqüència, la seva amplitud i el seu ample de banda. Els sons sonors tenen quatre formants com a mínim i, en general, només calen els dos inferiors per destriar les vocals. Altres paràmetres calculen l'energia espectral en diverses bandes de freqüència. Els paràmetres espectrals més emprats en l'àmbit de reconeixement emocional són els *Mel Frequency Cepstral Coefficients* (MFCC)[14], que es consideren un estàndard en la síntesi de veu i en el reconeixement automàtic de la parla. No obstant això, també han mostrat utilitat, per si sols, en l'àmbit del reconeixement afectiu automàtic malgrat suposar un volum de dades reduït (Steidl, 2009).

[14]Els quals, realment, fan referència a la transformada inversa del logaritme de l'espectre, emfatitzant canvis o periodicitat en el mateix

3.3.2 Parametrització lingüística

La parametrització lingüística fa referència a l'extracció d'informació a partir del contingut lèxic de les frases dels corpus. En aquest sentit, aquest tipus de parametrització fa èmfasi en el que es diu i no en com es diu, és a dir, se centra al canal explícit de la comunicació (vegeu la secció 2.3). Aquesta parametrització s'aprofita del fet que les persones tendim a emprar paraules específiques per expressar les nostres emocions en base a les associacions que hem après entre certes paraules i les emocions corresponents (Lee i Narayanan, 2005). No obstant això, a pesar que hi ha intents d'identificar el vocabulari de les emocions a través de descripcions subjectives d'emocions específiques (Plutchik, 1994), aquests només són útils per determinar paraules clau de forma genèrica i no en un entorn de comunicació natural (Lee i Narayanan, 2005). Per aconseguir això últim existeixen paràmetres com els descrits en aquesta secció.

A continuació es descriuen els paràmetres lingüístics analitzats en aquesta tesi. Les descripcions es refereixen en tot moment als elements lingüístics del corpus. Aquests elements lingüístics són, si es treballa a nivell d'unigrames, les paraules que formen cadascuna de les frases del corpus. En el cas de treballar a nivell de bigrames, els elements lingüístics es refereixen a les agrupacions de dues paraules consecutives en forma de parells ordenats. Aquesta descripció pot generalitzar-se per al cas de n-grames. Originàriament, els n-grames van mostrar la seva utilitat en tasques de reconeixement de veu (Nöth *et al.*, 2001) i més tard es van fer extensius als estudis de reconeixement emocional mostrant-se útils en aquest àmbit (Schuller *et al.*, 2005; Polzehl *et al.*, 2009; Steidl, 2009).

3.3.2.1 Prominència emocional i activació

Lee i Narayanan (2005) proposen el concepte de prominència emocional[15] per convertir cadascun dels elements lingüístics d'un corpus en paràmetres relacionats amb les etiquetes emocionals. Assumint que cada element és independent de la resta en una mateixa frase, la prominència d'un element es defineix com la informació mútua entre un element específic i una categoria emocional. En aquest sentit, un element prominentment emocional és un element que apareix més sovint relacionat amb aquesta categoria emocional que amb la resta de categories.

Sigui $W = \{v_1, v_2, ..., v_n\}$ el conjunt dels n elements d'una frase i sigui $I = \{i_1, i_2, ..., i_k\}$ l'univers emocional definit pel conjunt de k categories emocionals. La informació mútua entre l'element v_m i l'emoció i_j es defineix tal com indica l'equació 3.9.

$$i(v_m, i_j) = \log \frac{P(i_j|v_m)}{P(i_j)} \tag{3.9}$$

[15] De l'anglès, *emotional salience*. En aquesta tesi s'empra el terme *prominència emocional* per considerar-ho més descriptiu i adequat en aquest àmbit encara que en psicologia és habitual referir-se a aquest concepte com *saliència*, definit com la destacabilitat d'un estímul o la capacitat per *cridar l'atenció* d'un subjecte.

on $P(i_j|v_m)$ és la probabilitat a posteriori que una frase contenint l'element v_m impliqui l'emoció i_j i $P(i_j)$ és la probabilitat a priori d'aquesta emoció.

La prominència emocional d'un element v_m en relació amb l'univers emocional I (equació 3.10) es denota per $sal(v_m)$ i es defineix com la distància Kullback-Leibler entre la probabilitat a posteriori $P(i_j|v_m)$ d'una emoció i_j donat l'element v_m i la probabilitat a priori d'aquesta emoció $P(i_j)$ (Yildirim *et al.*, 2011).

$$sal(v_m) = \sum_{j=1}^{k} P(i_j|v_m) \log \frac{P(i_j|v_m)}{P(i_j)} \tag{3.10}$$

De tots els elements analitzats només se seleccionen aquells la prominència emocional dels quals supera un determinat llindar, el qual s'escollirà segons s'estipuli en l'experiment en qüestió. A continuació es calculen k paràmetres lingüístics per cada frase del corpus. Aquests paràmetres, anomenats activacions i denotats per a_j, es defineixen segons l'equació 3.11 (Lee i Narayanan, 2005). Així doncs, cada frase comptarà amb tants paràmetres d'activació com etiquetes emocionals es considerin.

$$a_j = \sum_{m=1}^{n} I_m w_{mj} + w_j \tag{3.11}$$

on I_m és 1 si l'element apareix en la llista d'elements que superen el valor llindar de prominència emocional o 0 en cas contrari, w_{mj} és el pes de connexió definit per l'equació 3.12 i w_j és el biaix definit per l'equació 3.13.

$$w_{mj} = i(v_m, i_j) = \log \frac{P(i_j|v_m)}{P(i_j)} \tag{3.12}$$

$$w_j = \log P(i_j) \tag{3.13}$$

La figura 3.3 mostra l'esquema del càlcul dels paràmetres d'activació per a cada categoria emocional a partir dels elements d'anàlisi de cada frase del corpus els quals poden ser, tal com es comentava amb anterioritat, paraules individuals (unigrames) o agrupacions de les mateixes (n-grames, en general).

3.3.2.2 Afinitat prototípica

L'afinitat prototípica és una mesura proposada per Steidl (2009) durant la validació del corpus FAU Aibo. En aquest corpus, a més de la categoria emocional, cada locució porta associat un paràmetre, l'afinitat prototípica, representant l'homogeneïtat de l'etiquetatge dut a terme pels anotadors del corpus. L'afinitat prototípica d'un segment es

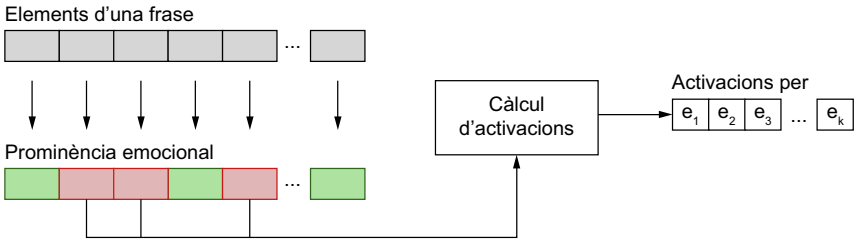


Figura 3.3: Esquema de la parametrització lingüística d'una frase calculant k paràmetres d'activació, un per a categoria emocional. S'han marcat en vermell aquells elements de la frase la prominència emocional dels quals supera un llindar preestablert.

defineix com la relació entre el nombre d'etiquetes categòriques (a nivell de cadascuna de les paraules que formen part del segment en qüestió) que coincideixen amb l'etiqueta assignada finalment a aquest fragment i el nombre total d'etiquetes que ha rebut per part dels anotadors, és a dir, la proporció d'etiquetes categòriques que coincideixen (a nivell de paraula) amb la categoria final del fragment.

Per poder crear una llista de paraules emocionals amb aquesta informació, es calcula l'afinitat prototípica mitjana de cada paraula w del corpus referenciada a aquesta emoció i, tal com s'indica en l'equació 3.14.

$$avg\_prot(w,i) = \frac{\sum_{i=1}^{N} prot_i(w,i)}{N} \tag{3.14}$$

on $prot_i(w,i)$ és l'afinitat prototípica de la paraula w respecte a l'emoció i i N és el nombre de frases amb aquesta etiqueta emocional assignada.

3.3.2.3 Dimensions emocionals

El model bidimensional d'activació i valència (vegeu la secció 2.1.3.3) per mapejar paraules en l'univers emocional ha estat emprat amb èxit en diversos estudis (Stevenson *et al.*, 2007). Per a això és necessari un diccionari afectiu com, per exemple, els diccionaris *Affective Norms for English Words* (ANEW) (Bradley i Lang, 1999; Trilla i Alías, 2009), *Dictionary of Affect in Language* (DAL) (Hofer *et al.*, 2005; Whissell, 2008) o una llista autocompilada de paraules emocionals (Francisco i Gervás, 2006; Osherenko, 2008). No obstant això, existeixen discrepàncies respecte a l'ús d'aquests models atribuïbles al fet que el seu disseny pot veure's influït per la personalitat, el caràcter i el context social dels avaluadors que els dissenyen (Russell *et al.*, 1989). Al mateix temps, l'estudi aïllat de les paraules podria millorar-se amb un estudi de les paraules contigües que puguin modificar els seus significats. Així, per exemple, si un adverbi de negació acompanya a una paraula, els coeficients dimensionals poden girar-se al voltant de l'estat neutre en un model *circumplex*, com en (Trilla i Alías, 2009).

3.3.2.4 Model d'espai vectorial

En aquest model, els elements que apareixen en el corpus defineixen un espai d'alta dimensionalitat en el qual es representen les transcripcions de les frases. Per exemple, Steidl (2009) crea un espai amb la freqüència d'aparició de totes les paraules del corpus. No obstant això, l'alta dimensionalitat d'aquest espai pot ocasionar problemes de dispersió. Malgrat tot, Alías *et al.* (2008) indiquen que aquests problemes es poden esmenar analitzant únicament les paraules avaluades i no el corpus complet.

3.4 Corpus emprats en aquesta tesi

En els experiments de reconeixement automàtic d'emocions explicats al llarg d'aquesta tesi es treballa, fonamentalment, amb tres corpus de veu emocionada. En aquesta secció es descriu cadascun d'aquests corpus i es detalla la parametrització que s'ha realitzat dels mateixos. Aquestes parametritzacions pretenen recollir de forma àmplia els paràmetres explicats en la secció 3.3.

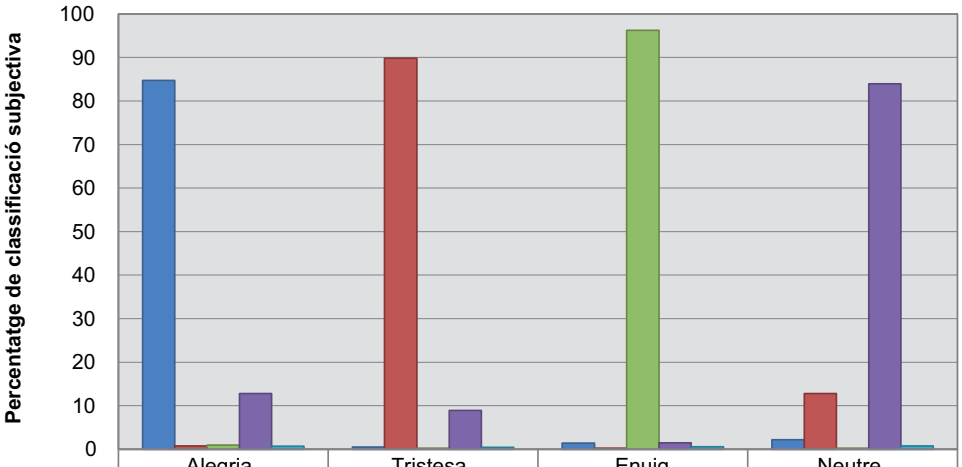
3.4.1 Corpus ESDLA

El corpus ESDLA es concep originalment com un corpus no massa extens i amb una parametrització senzilla destinat a experiments preliminars en l'àmbit del reconeixement afectiu automàtic. A aquest efecte, la seva parametrització es concep sobre la base del treball de Oudeyer (2003), els experiments del qual són adaptats a aquest corpus en una fase inicial d'aquesta tesi, sense que suposin l'eix vertebrador de la mateixa. Tal com es detalla a continuació, malgrat ser un corpus senzill compta amb una anàlisi subjectiva que ofereix una referència per a l'anàlisi dels resultats dels estudis de reconeixement automàtic, així com una parametrització tant a nivell fonètic com a nivell segmental.

3.4.1.1 Descripció

El corpus ESDLA és un corpus d'expressió vocal actuada (vegeu la secció 3.2.4) consistent en la recopilació de les lectures, per part d'un locutor no professional masculí, d'un conjunt de 200 frases en espanyol. Per intentar aconseguir frases semànticament neutres aquestes frases es corresponen amb 50 fragments breus, de caràcter majoritàriament descriptiu, de l'obra *El Señor de los Anillos* (Tolkien, 1954). Cadascun d'aquests 50 fragments són interpretats per simular tres estats emocionals diferents: alegria, enuig i tristesa. També s'inclou una interpretació neutra, sense expressivitat emocional. Els enregistraments estan emmagatzemats en format WAV amb una freqüència de mostreig de 16 KHz i una quantificació de 16 bits/mostra. Les condicions d'enregistrament estaven controlades però l'enregistrament no es va realitzar en un estudi professional.

Una avaluació subjectiva permet la validació de l'expressivitat de la parla actuada des d'un punt de vista d'usuari (Montero *et al.*, 1998a; Hozjan *et al.*, 2002; Navas *et al.*, 2006; Morrison *et al.*, 2007). Així doncs, les 200 frases del corpus se sotmeten a un test subjectiu amb la finalitat d'observar els resultats de classificació obtinguts per avaluadors humans. El test es realitza sobre 35 persones de perfil heterogeni. Entre els avaluadors existeixen persones expertes en processament del senyal de veu (dins del camp de l'enginyeria) i també relacionades amb l'estudi de les emocions (pertanyents al camp de la psicologia), així com no vinculades a cap d'aquests àmbits. El test es duu a terme mitjançant l'eina TRUE. El test consisteix a mostrar a cadascun dels avaluadors, a través de la interfície web, cadascun dels arxius de forma seqüencial i en ordre aleatori, mitjançant un test de resposta forçada al plantejament de la qüestió: *"Quina emoció penses que expressa el locutor en aquesta frase?"* i oferint les quatre possibles opcions de classificació (*"Alegria"*, *"Tristesa"*, *"Enuig"* i *"Neutre"*) més una addicional per seleccionar en cas de dubte (*"NS/NC"*). En una fase prèvia a l'inici del test, a cada usuari se li permet escoltar alguns dels arxius d'àudio com a mostra de cadascuna de les emocions sense que aquesta sigui indicada en cap cas. La finalitat d'aquest pas previ és familiaritzar a l'avaluador amb la veu del locutor oferint-li una referència inicial i intentant, d'aquesta manera, minimitzar els errors de classificació que puguin cometre's a l'inici de la prova.



	Alegria	Tristesa	Enuig	Neutre
Alegría	84,72	0,50	1,44	2,22
Tristeza	0,78	89,83	0,28	12,78
Enfado	1,00	0,28	96,22	0,28
Neutro	12,78	8,94	1,50	83,94
NS/NC	0,72	0,44	0,56	0,78

Figura 3.4: Matriu de confusió de la classificació subjectiva del corpus ESDLA.

La figura 3.4 mostra la matriu de confusió dels resultats del test subjectiu realitzat. La figura representa, en percentatge, la classificació en cadascuna de les categories emocionals dels arxius pertanyents a una mateixa emoció segons les respostes dels avalu-

adors, és a dir, la classificació subjectiva de les frases de cada categoria emocional. Així, per exemple, es pot constatar que l'emoció d'enuig és la reconeguda amb major exactitud, en tant que el 96,22% de les frases etiquetades amb aquesta categoria emocional es reconeixen com a tals i tan sols un 3,78% de les frases de les frases d'aquesta emoció reben un etiquetatge diferent (o no és identificat com a cap de les etiquetes emocionals considerades). La taula 3.1 mostra una anàlisi detallada de la classificació subjectiva detallada per classes, recollint els detalls de la classificació del corpus segons el test subjectiu descrit mostrant els valors de les mesures de precisió, cobertura i mesura F1 per a cadascuna de les emocions estudiades (vegeu la secció 4.3.4.1). En mitjana, el percentatge de casos correctament classificats considerant la totalitat d'avaluadors ascendeix al 88,68%.

Taula 3.1: Detalls, en percentatge, de la classificació subjectiva del corpus ESDLA especificats per classes.

	Alegria	Tristesa	Enuig	Neutre
Precisió	95,31	86,66	98,41	78,33
Cobertura	84,72	89,83	96,22	83,94
F1	89,71	88,22	97,30	81,04

3.4.1.2 Parametrització

El corpus ESDLA s'ha parametritzat acústicament a nivell fonètic i a nivell segmental (Planet *et al.*, 2006). Comptant amb la transcripció fonètica de cadascuna de les locucions del corpus, la segmentació dels arxius en fonemes i l'anàlisi dels mateixos per calcular els descriptors de baix nivell s'ha dut a terme amb el programari ITP[16](Alías i Iriondo, 2002). Aquest programari també ha servit per a l'extracció dels descriptors de baix nivell en el cas de la divisió segmental, per a la qual s'han emprat finestres de 20 ms. La segmentació en fonemes es duu a terme mitjançant l'aplicació de *Hidden Markov Models* (HMM)[17]. El procés de segmentació situa una sèrie de marques temporals en un arxiu d'àudio les quals delimiten cadascun dels fonemes que formen part de l'enregistrament. Aquesta segmentació s'obté de forma automàtica a partir de la transcripció fonètica de les frases, els arxius d'àudio i uns arxius d'entrenament previs. No obstant això, el procés requereix d'una supervisió manual posterior per corregir la posició d'algunes de les marques de segmentació que poguessin no estar correctament situades. La transcripció fonètica de les frases s'obté a partir de les transcripcions de les locucions mitjançant un procés automàtic gràcies al programari específicament dissenyat per a tal finalitat.

La mesura dels silencis no suposa una complicació especial en el cas de la segmentació en fonemes, ja que queden identificats per la pròpia transcripció de la frase i n'hi ha prou amb observar la durada dels fonemes corresponents als silencis. No obstant això, si la frase ha estat dividida en segments de durada constant, cal realitzar un estudi dels

[16]Interfície per al Tractament de la Parla.
[17]Models Ocults de Markov, en anglès.

mateixos per intentar destriar els que es corresponen amb silencis dels que no ho són. El procediment per realitzar aquesta distinció de forma automàtica consisteix a analitzar les primeres trames calculant el valor mitjà d'energia ja que se sap a priori que aquesta part del senyal es correspon amb silenci. De forma heurística, s'escull un valor un 25% superior al valor mitjà abans calculat com a llindar per determinar que aquelles trames que estiguin per sota del mateix són silencis. Addicionalment, per fer més robust el sistema, es corregeixen aquells fragments únics que són considerats com a silencis si estan precedits i seguits per fragments no considerats així. D'igual manera es corregeixen aquells fragments únics que no són considerats silencis i estan precedits i seguits per fragments així considerats.

En el cas de la segmentació en fonemes es consideren 5 paràmetres addicionals: valors mitjà, màxim, mínim i desviació estàndard de les durades dels fonemes i una mesura relativa a la velocitat de la parla sobre la base de la relació entre la durada de cada fonema i la durada mitjana d'aquest fonema al llarg de tot el corpus.

Seguint el treball d'Oudeyer (2003), també es consideren dues versions addicionals dels arxius d'àudio processats mitjançant un filtre passa-altes i un altre passa-baixes. Això permet realitzar una anàlisi més detallat en diferents bandes espectrals podent observar el comportament del senyal en aquestes zones. Els descriptors de baix nivell analitzats són la F0 i l'energia a més de les durades dels silencis. En el cas de la F0 i l'energia, a més, s'analitza la primera derivada per obtenir també una mesura de la variació temporal d'aquests paràmetres.

Els 12 funcionals estudiats en aquest corpus són el valor mitjà, extrems (valor màxim i valor mínim), diferència entre els extrems, desviació estàndard, mitjana, quartils, rang interquartílic, asimetria i curtosis. Per al cas dels silencis es considera un paràmetre addicional: la durada total de silencis en una locució.

Segons les consideracions anteriors, cada element del corpus s'identifica per 188 paràmetres tal com es desglossa a la taula 3.2. Per exemple, per la F0 es consideren 3 seqüències (original, filtrada passa-baixes i filtrada passa-altes), 2 funcions (original i primera derivada) i els 12 funcionals abans detallats.

Taula 3.2: Detall de la parametrització del corpus ESDLA.

	Seqüències	Funcions	Funcionals	Total
F0	3	2	12	72
Energia	3	2	12	72
Silenci	3	-	13	39
Durada	1	1	5	5
Total				188

3.4.2 Corpus BDP-UAB

El corpus BDP-UAB està orientat a la síntesi de veu en espanyol i s'ha desenvolupat amb un doble propòsit: en primer lloc per ser emprat en el modelatge acústic, tant a nivell prosòdic com de VoQ, de la veu emocionada i, en segon lloc, per ser la base de dades d'unitats d'un sintetitzador vocal. Aquest corpus es va desenvolupar pel Grup de Recerca en Processament Multimodal (GPMM) d'Enginyeria i Arquitectura La Salle i el Laboratori d'Anàlisi Instrumental de la Comunicació de la Universitat Autònoma de Barcelona (LAICOM-UAB). Per a una explicació exhaustiva d'aquest corpus pot consultar-se el treball d'Iriondo *et al.* (2009) i el seu desenvolupament complet detallat en (Iriondo, 2008).

3.4.2.1 Descripció

El corpus BDP-UAB és un corpus d'expressió vocal estimulada (vegeu la secció 3.2.3) consistent en la lectura de textos per part d'una locutora femenina professional. El contingut semàntic està dissenyat per facilitar la seva expressió emocional. Els textos van ser seleccionats a partir d'una base de dades d'anuncis extrets de mitjans de premsa escrita, incloent diaris i revistes. Prèviament, aquesta base de dades va ser classificada en diverses categories temàtiques les quals, segons l'estudi previ dut a terme per Montoya (1998), es consideren relacionades amb diferents estils expressius. La locutora va ser prèviament entrenada per adequar el seu estil expressiu a les emocions requerides. Les característiques fonètiques van ser definides pels experts del LAICOM-UAB. L'enregistrament es va dur a terme comptant amb supervisió experta per evitar possibles desviacions de l'estil emocional definit per a cada locució. En total, quatre persones estaven involucrades en el procés d'enregistrament: la locutora professional, un enginyer de so, un expert en comunicació audiovisual i un tècnic de control per a la supervisió. Els estils emocionals recollits en el corpus i les seves categories associades als tipus de publicitat que faciliten la seva interpretació són:

- Neutre: associat a la categoria de noves tecnologies, transmetent maduresa i serietat.
- Alegre: associat a la categoria d'educació, transmetent extraversió.
- Sensual: associat a la categoria de cosmètics, basat en una veu dolça, suau i pausada.
- Agressiu: associat a la categoria d'automòbils, basat en un estil dur.
- Trist: associat a la categoria de viatges, transmetent una sensació de malenconia.

Per aconseguir una selecció fonèticament equilibrada de les frases que formen el corpus, donat el seu origen orientat a la síntesi de veu emocionada, la seva elecció es duu a terme mitjançant un algorisme *greedy* (François i Boëffard, 2002), ometent frases amb

excepcions (paraules estrangeres i amb abreviatures) i evitant frases similars entre si. Pot consultar-se el detall de la distribució fonètica del corpus en (Iriondo *et al.*, 2009)

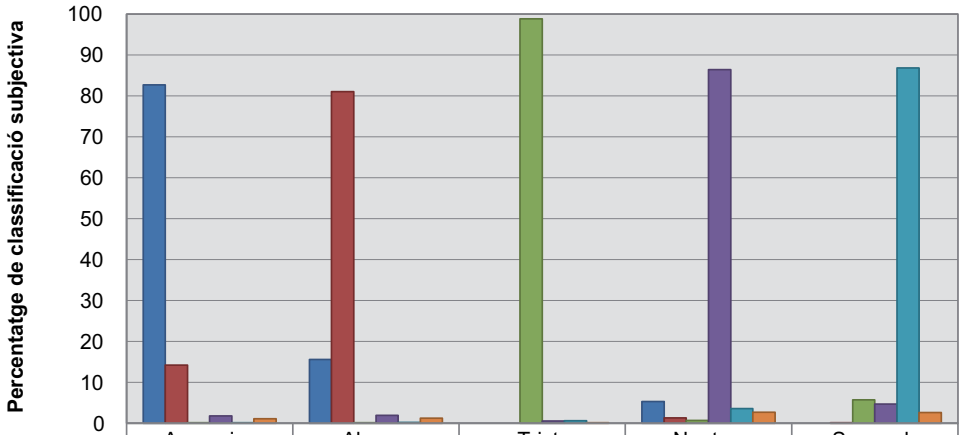
L'enregistrament d'aquest corpus es duu a terme en un estudi d'enregistrament professional preparat acústicament per oferir condicions òptimes d'enregistrament i un alt nivell d'aïllament. La sala d'enregistrament de 20 m^2 es distribueix en la superfície no quadrada d'una habitació de 3,5 metres d'altura. A pesar que l'habitació compta amb un temps de resposta de 0,8 segons la posició del micròfon garanteix l'absència de ressons audibles en l'enregistrament. Aquest micròfon és un micròfon de condensador d'alta qualitat (AKG C-414), amb una resposta plana (2 dB en el marge de 20 Hz a 20 KHz) i amb una relació senyal-soroll de 80 dBA SPL. L'enregistrament digital dels arxius en format WAV de 48 KHz i 24 bits s'emmagatzema en un disc dur emprant el programari Pro Tools 5.1 sobre una plataforma Mac G5 emprant una consola digital Yamaha 02R.

Després d'un procés semiautomàtic de segmentació del corpus i eliminació de les frases gravades incorrectament o no adequades a l'estil requerit, el volum d'informació del mateix és a 4.638 locucions sumant un total de 5 hores i 27 minuts d'enregistrament. 833 locucions (50 minuts) es corresponen a l'estil neutre, 916 locucions (56 minuts) a l'estil feliç, 841 locucions (51 minuts) a l'estil sensual, 1.000 locucions (86 minuts) a l'estil trist i 1.048 locucions (84 minuts) a l'estil agressiu.

Com en el corpus ESDLA (vegeu la secció 3.4.1.1), el corpus BDP-UAB també és sotmès a un procés d'avaluació subjectiva. No obstant això, en aquest cas, donada la grandària del corpus, una avaluació subjectiva exhaustiva podria ser extremadament costosa tant a nivell temporal com de recursos humans per dur-la a terme. En aquest sentit, Iriondo *et al.* (2009) proposen una avaluació subjectiva parcial que permet obtenir una representació de la percepció subjectiva (els resultats de la qual es presenten a continuació) que pot ser emprada, posteriorment, per a una validació automàtica del corpus complet. Per al test subjectiu parcial s'escullen un total de 480 locucions (aproximadament un 10% del total, 96 locucions per cada estil emocional) i 25 voluntaris d'Enginyeria i Arquitectura La Salle com a avaluadors subjectius. Per evitar que cada subjecte hagi de evalúar el total de 480 locucions es creen 4 subconjunts de 120 locucions cadascun. Un parell ordenat de subconjunts és assignat a cada avaluador creant 12 combinacions diferents. La creació de parells ordenats pretén evitar que els subconjunts que s'avaluïn en segon lloc puguin rebre una millor classificació gràcies a l'entrenament rebut en avaluar els que es troben en primer lloc. L'avaluació es duu a terme en la plataforma TRUE mitjançant un test de resposta forçada amb la qüestió *"Quina emoció reconeixes en la veu de la locutora en aquesta frase?"*. Les possibles respostes són els 5 possibles estats emocionals i un d'addicional a seleccionar en cas de dubte (*"NS/NC"*).

La figura 3.5 mostra la matriu de confusió de la classificació subjectiva parcial del corpus BDP-UAB. Es pot afirmar que, en general, tots els estils expressius mostren un alt percentatge d'identificació (87,1% en mitjana). L'estil trist és el millor identificat seguit pel sensual i el neutre. Els principals errors d'identificació es produeixen en l'estil agressiu (el 14,2% de les locucions agressives s'identifiquen com a alegres). A més, l'estil neutre

es confon lleugerament amb la resta d'estils i també hi ha un cert nivell de confusió entre els estils sensual i trist (5,7%). L'opció *"NS/NC"* es va emprar en molt poques ocasions, encara que és més freqüent en els estils neutre i sensual que en la resta.



	Agressiu	Alegre	Trist	Neutre	Sensual
Agresivo	82,70	15,60	0,00	5,30	0,00
Alegre	14,20	81,00	0,00	1,30	0,10
Triste	0,10	0,10	98,80	0,70	5,70
Neutro	1,80	1,90	0,50	86,40	4,70
Sensual	0,10	0,20	0,60	3,60	86,80
NS/NC	1,10	1,20	0,10	2,70	2,60

Classificació real

Figura 3.5: Matriu de confusió de la classificació subjectiva del corpus BDP-UAB. Adaptat d'(Iriondo *et al.*, 2009).

La figura 3.6 mostra dos histogrames que detallen la classificació subjectiva de les locucions del corpus. L'histograma (a) determina el nombre de locucions que són correctament classificades en els percentatges indicats. L'histograma (b) determina el nombre de locucions que reben l'etiqueta *"NS/NC"* en els percentatges també assenyalats.

3.4.2.2 Parametrització

El corpus BDP-UAB està parametritzat acústicament a nivell fonètic. La segmentació en fonemes es va dur a terme mitjançant el programari ITP. L'anàlisi acústica del corpus proporciona paràmetres prosòdics i de VoQ. A nivell prosòdic s'inclou informació relativa a la F0, energia i ritme de la funció original i de les derivades primera i segona.

L'anàlisi de F0 es basa en el marcador fonètic descrit per Alías *et al.* (2006), considerant tres versions diferents de cada locució: completa, excloent silencis i els sons sords, i considerant únicament vocals tòniques. La informació relativa a la sonoritat dels segments així com dels silencis i de les vocals tòniques s'obté a partir de l'etiquetatge fonètic de les locucions. Aquesta mesura es realitza seguint escales lineal i logarítmica.

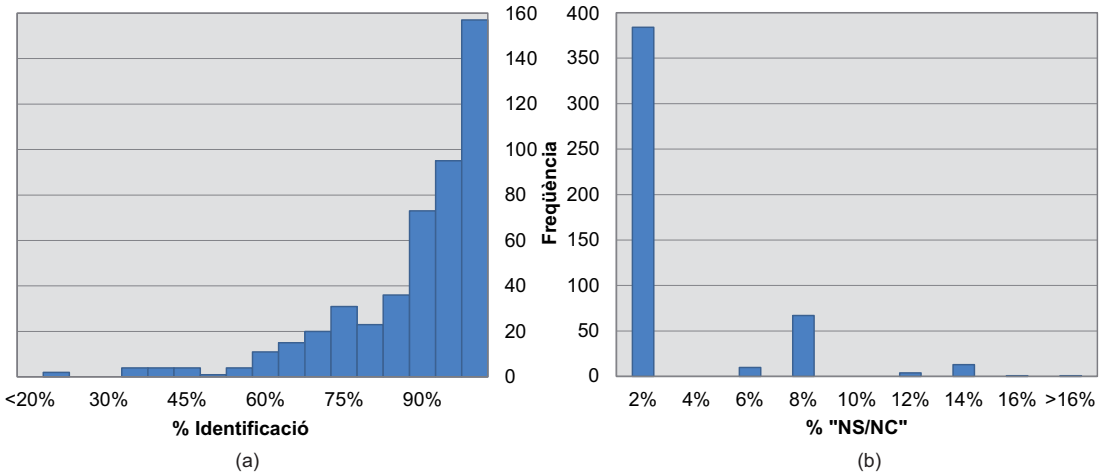


Figura 3.6: Histogrames que reflecteixen l'avaluació de les 480 locucions de la prova subjectiva. L'eix vertical mostra el nombre de locucions que: (a) són correctament classificades en el percentatge indicat en l'eix horitzontal, (b) reben l'etiqueta *"NS/NC"* en el percentatge indicat en l'eix horitzontal. Adaptat d'(Iriondo *et al.*, 2009).

L'anàlisi d'energia es realitza mitjançant finestres rectangulars de 20 ms a intervals de 10 ms, calculant l'energia mitjana en escala lineal i en escala logarítmica (dB). Igual que en el cas de la F0, es consideren les tres seqüències abans descrites.

L'anàlisi del ritme es focalitza en la durada dels fonemes i l'anàlisi dels silencis. Per modelitzar la durada dels fonemes s'empra la mesura *z-score* descrita en la secció 3.3.1.1. Tant la mitjana com la desviació estàndard s'estimen per a cada fonema a partir del corpus complet. Addicionalment es calcula una altra versió de les durades considerant únicament les vocals tòniques. Per analitzar els silencis es calculen el nombre de pauses per segon i el percentatge de temps de silenci pel que fa a la durada total de la locució. Tots dos paràmetres representen la freqüència i la durada de les pauses.

Els paràmetres relatius a la VoQ han estat calculats mitjançant el programari d'anàlisi Praat[18] a partir dels enregistraments d'àudio realitzats i sense necessitat d'emprar dispositius transductors invasius o maquinari addicional d'acord amb la proposta de Drioli *et al.* (2003) i la implementació descrita per Monzo *et al.* (2007). Aquests paràmetres inclouen les mesures de *jitter*, *shimmer*, GNE, HammI i do1000. Aquests 5 paràmetres són els empleats per Monzo *et al.* (2007) per discriminar els 5 estils de parla del corpus BDP-UAB amb uns resultats acceptables, si bé l'estudi conclou indicant la necessitat d'incloure la informació prosòdica per millorar els resultats.

El nombre de funcionals analitzats en aquest corpus ascendeix a 11, incloent el valor mitjà, la variància, els extrems (valor màxim i valor mínim), diferència entre els extrems, asimetria, curtosis, quartils i rang interquartílic.

[18]http://www.praat.org

Segons les consideracions anteriors, cada element del corpus s'identifica per 519 paràmetres tal com es desglossa en la taula 3.3. Per exemple, per la F0 es consideren 2 unitats (lineal i logarítmica), 3 seqüències (completa[19], excloent silencis i sons sords, i considerant únicament vocals tòniques), 3 funcions (original i primera i segona derivades) i els 11 funcionals abans detallats.

Taula 3.3: Detall de la parametrització del corpus BDP-UAB.

	Unitats	Seqüències	Funcions	Funcionals	Total
F0	2	3	3	11	198
Energia	2	3	3	11	198
Durada	1	2	3	11	66
Pauses	2	-	-	-	2
VoQ	5	-	-	11	55
Total					519

3.4.3 Corpus FAU Aibo

Per dur a terme un estudi de reconeixement d'emocions espontànies en un entorn realista és necessari treballar amb un corpus adequat. No obstant això, molts dels corpus existents presenten inconvenients com, entre d'altres, la seva escassa grandària o el reduït nombre de locutors presents en l'enregistrament que impedeix una anàlisi independent del locutor (Steidl, 2009). El corpus FAU Aibo es crea amb la intenció d'adequar-se a l'escenari estudiat intentant evitar els inconvenients anteriors i resulta en un corpus no equilibrat, que inclou dades no prototípiques i locucions d'una baixa intensitat emocional.

L'enregistrament del corpus FAU Aibo es va finançar per la Comunitat Europea en el marc del projecte *Preparing Future Multisensorial Interaction Research* (PF-STAR). Existeixen dues versions (Batliner *et al.*, 2004), una en anglès (duta a terme per l'Escola d'Enginyeria de la Universitat de Birmingham) i una altra en alemany (duta a terme per la Universitat Erlangen-Nürnberg). L'última és la usada en aquesta tesi ja que va ser distribuïda pel comitè organitzador de l'*Interspeech Emotion Challenge 2009* (Schuller *et al.*, 2009).

3.4.3.1 Descripció

El corpus FAU Aibo és un corpus d'expressió vocal natural (vegeu la secció 3.2.1) consistent en la recopilació de 9,2 hores d'enregistraments en alemany corresponents a la interacció de 51 nens (30 nenes i 21 nens) amb el robot Aibo, de Sony, en un escenari de *Mag d'Oz*. Els nens i nenes, d'edats compreses entre els 10 i els 13 anys d'edat, pertanyien a dues escoles diferents de Erlangen (Alemanya), l'*Ohm-Gymnasium* i la *Montessori-Schule*. Les

[19] Per seqüència completa s'entén aquella que considera tots els segments.

sessions d'enregistrament es van dur a terme en aules separades assistides per tres persones. Es va emprar l'equip sense fils Shure UT 14/20 UHF amb el micròfon WH20TQG en combinació amb l'equip d'enregistrament DAT Tascam DÓNA-P1, a una freqüència de mostreig de 48 KHz i quantificació de 16 bits. D'acord amb l'estàndard de reconeixement automàtic de veu, els enregistraments es van remostrejar a 16 KHz. Addicionalment, encara que sense interès per al present treball, es van realitzar enregistraments en vídeo de les sessions amb lent gran angular, a 25 fotogrames per segon, en sistema PAL. Aquests enregistraments de vídeo contenen també una pista d'àudio amb les mateixes locucions emmagatzemades a través del micròfon sense fils però amb un major nivell de soroll i reverberació.

Els experiments duts a terme sota l'entorn de *Mag d'Oz* consistien a demanar als nens i nenes que guiessin el robot mitjançant ordres de veu per realitzar diverses tasques en el menor temps possible (consulteu el treball de Steidl (2009) per a major informació). Malgrat les instruccions, el robot responia a una seqüència programada d'accions de forma que no sempre realitzava les tasques dictades. A diferència d'altres experiments duts a terme en entorns de *Mag d'Oz*, en aquest cas els participants no van haver de pretendre estar interessats en la tasca duta a terme sinó que realment van gaudir realitzant-la (segons els investigadors que van dur a terme l'experiment (Steidl, 2009)) i no van saber, en cap moment, que el robot estava sent controlat remotament.

Els enregistraments realitzats en aquest escenari es divideixen en un total de 18.216 fragments[20]. Per permetre una anàlisi posterior independent del locutor els fragments d'ambdues escoles s'agrupen per separat per crear dos conjunts diferents i independents, amb enregistraments de nens i nenes diferents. Així doncs, els enregistraments de l'escola *Ohm-Gymnasium* (Ohm) conformen un grup de 9.959 locucions corresponents a 13 nens i 13 nenes, mentre que els enregistraments de l'escola *Montessori-Schule* (Mont) conformen un grup de 8.257 locucions corresponents a 8 nens i 17 nenes. Schuller *et al.* (2009) escull el conjunt de locucions (Ohm) com a conjunt d'entrenament dels algorismes de reconeixement afectiu i el conjunt (Mont) com a conjunt de prova.

El corpus es divideix originalment en 11 categories si bé la versió emprada en aquest treball, d'acord amb la distribució de Schuller *et al.* (2009), redueix la divisió original a 5 categories afectives sobre la base de l'etiquetatge rebut per un grup de 5 anotadors segons el procediment detallat per Steidl (2009):

- Enuig (etiquetat amb la lletra A, de l'anglès *Anger*), que inclou els estats d'enuig (en el sentit de molest), irascible (com a pas previ a l'enuig) i reprensor.
- Emfàtic (etiquetat amb la lletra I), estat caracteritzat per un estil de parla molt articulat però sense emoció.
- Neutre (etiquetat amb la lletra N).

[20]Els fragments en els que es divideix el corpus FAU Aibo són segments d'àudio amb significat complet a nivell sintàctic i semàntic. Es defineixen manualment sobre la base de criteris sintàctics i prosòdics (Schuller *et al.*, 2009).

- Positiu (etiquetat amb la lletra P), incloent un estil maternal (similar al parla dirigida als nens i nenes però, en aquest cas, del nen o la nena dirigits al robot) i alegre (quan el nen o la nena gaudeixen d'una situació concreta).

- Resta (etiquetat amb la lletra R), una classe que inclou altres estats com els de sorpresa (en un sentit positiu), avorriment (mancat d'interès per la interacció amb el robot) i impotència (dubitatiu, una locució plena pauses i amb poca fluïdessa).

La figura 3.7 mostra com es distribueixen els fragments del corpus FAU Aibo en les 5 categories anteriors. Com es pot veure, la distribució és poc equilibrada. Així, per exemple, la classe majoritària (N) consisteix en 10.967 locucions (60,21% del total) mentre que la classe minoritària (P) consisteix en només 889 locucions (4,88% del total).

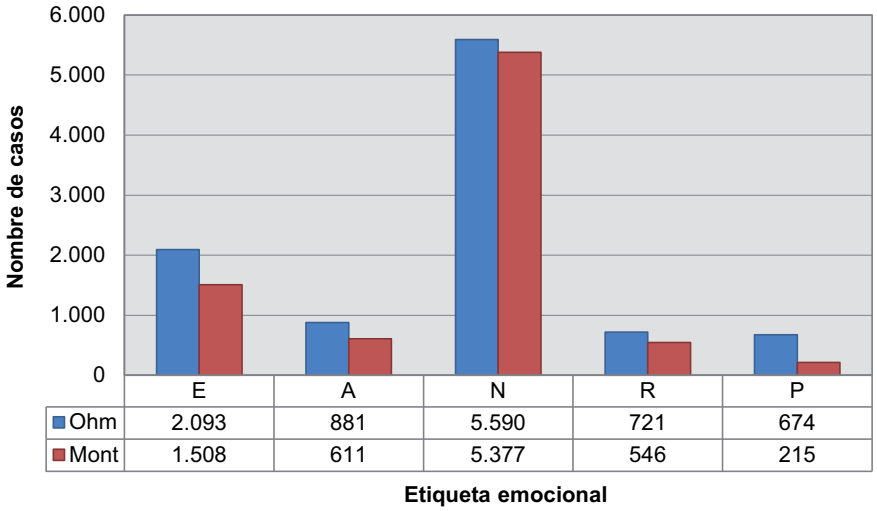


Figura 3.7: Distribució de classes del corpus FAU Aibo.

El corpus FAU Aibo compta amb un paràmetre d'afinitat prototípica (vegeu la secció 3.3.2.2) calculat sobre la base de 5 anotadors involucrats en el procés de validació. Aquest paràmetre pot prendre valors entre 0 i 1 i, per al cas d'aquest corpus, els valors inferiors a 0,25 s'assumeixen a 0, segons es desprèn de l'observació del mateix. Atès que la classe R està formada per diverses emocions dispars la major part de les seves locucions s'associen a un valor d'afinitat prototípica igual a 0 (concretament el 90% dels fragments que reben aquesta etiqueta). La resta de casos d'aquesta mateixa categoria tenen un valor d'afinitat prototípica molt baix. Això pot observar-se en la figura 3.8, la qual mostra l'índex d'ocurrència, en percentatge, de cada etiqueta categòrica dins de 10 intervals equiespaiats de prototipicalitat. Les altres 4 classes es distribueixen al llarg dels diferents valors d'aquest paràmetre però només uns pocs casos (corresponents a la classe N en la major part de les ocasions) aconsegueixen un valor d'afinitat prototípica igual a 1.

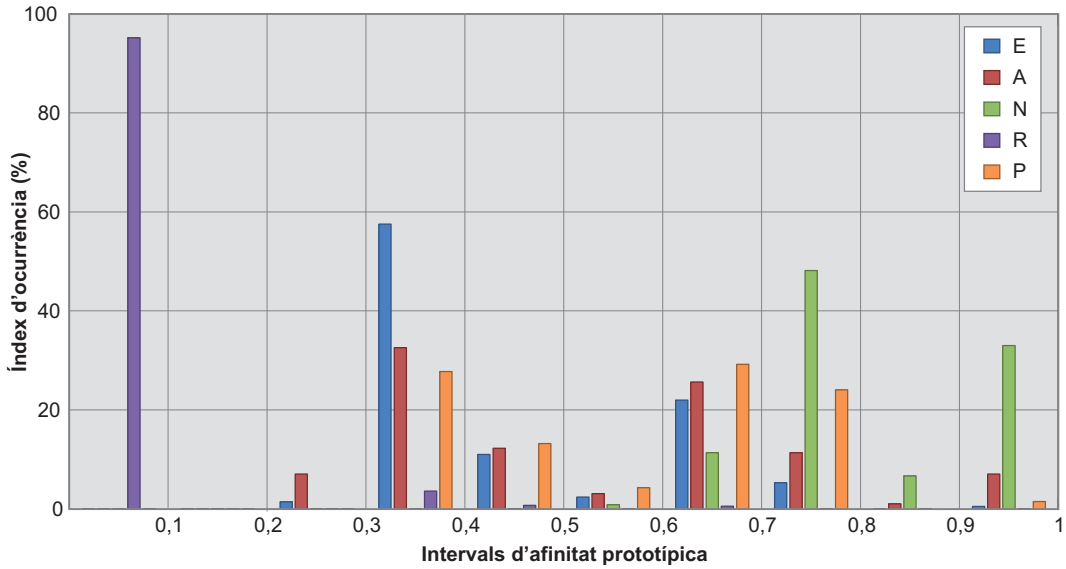


Figura 3.8: Índex d'ocurrència, en percentatge, de cada etiqueta categòrica dins de cada interval d'afinitat prototípica. El marge d'afinitat prototípica de 0 a 1 de l'eix horitzontal apareix dividit en 10 intervals equiespaiats. Atès que els valors d'afinitat prototípica inferiors a 0,25 s'assumeixen com 0, no hi ha ocurrències en l'interval de 0,1 a 0,2 ja que es desplacen a l'interval anterior.

3.4.3.2 Parametrització

El corpus FAU Aibo està parametritzat a nivell acústic i a nivell lingüístic.

La parametrització acústica del corpus FAU Aibo està realitzada a nivell segmental. Aquesta parametrització es va realitzar mitjançant el programari openSmile inclòs en el paquet openEAR (Eyben *et al.*, 2009).

A nivell acústic s'extreuen 16 descriptors de baix nivell. Aquests descriptors són: la ràtio de creuaments per zero[21] analitzada en el domini temporal del senyal de veu, l'energia quadràtica mitjana de la finestra, la F0 normalitzada a 500 Hz, HNR i 12 coeficients MFCC. A més dels paràmetres originals es calcula la primera derivada dels mateixos.

D'aquests descriptors es calculen 12 funcionals: valor mitjà, desviació estàndard, asimetria, curtosis, extrems, diferència entre els extrems, posició dels extrems i dos coeficients de regressió lineal (*offset* i pendent) i l'error quadràtic mitjà associat.

Segons les consideracions anteriors, cada element del corpus s'identifica per 384 paràmetres acústics tal com es desglossa en la taula 3.4. Per exemple, per la F0 es consideren 2 funcions (original i primera derivada) i els 12 funcionals abans detallats.

La parametrització lingüística es basa en la transcripció de les locucions del corpus proporcionada pels seus desenvolupadors i es fonamenta en el concepte de prominència

[21] De l'anglès, *Zero-Crossing Rate* (ZCR).

Taula 3.4: Detall de la parametrització del corpus FAU Aibo.

	Unitats	Funcions	Funcionals	Total
F0	1	2	12	24
Energia	1	2	12	24
ZCR	1	2	12	24
HNR	1	2	12	24
MFCC	12	2	12	288
Total				384

emocional i el càlcul de paràmetres d'activació detallat en la secció 3.3.2.1. L'anàlisi es realitza a nivell de paraula i a nivell morfològic tal com es detalla a continuació. Cal destacar la reduïda longitud de les frases del corpus tal com s'il·lustra en la figura 3.9, la qual mostra un histograma de la longitud de les locucions i la seva distribució, en percentatge, en el conjunt d'entrenament.

A nivell de paraula, en aquest treball s'ha realitzat una anàlisi considerant-les individualment (unigrames) i en agrupacions de parells ordenats (bigrames). No s'ha considerat l'anàlisi a nivell de paraula en forma de trigrames ja que els models podrien no estimar-se adequadament (Steidl, 2009).

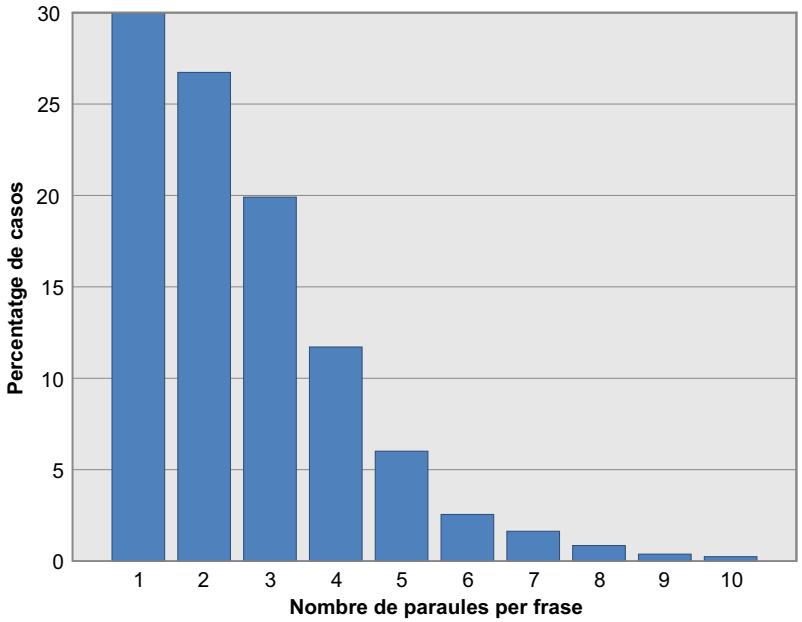


Figura 3.9: Histograma de la longitud de les locucions i la seva distribució, en percentatge, en el conjunt d'entrenament.

Així mateix, la parametrització lingüística també té en consideració la categoria morfològica de cada paraula dins de la locució en la qual apareix mitjançant l'etiquetatge POS[22] de les locucions del corpus. Aquest etiquetatge es duu a terme mitjançant la

[22] De l'anglès, *Part-Of-Speech* (POS) *tagging*.

plataforma MARY TTS (Pammi *et al.*, 2010). Posteriorment es realitzen agrupacions en trigrames d'aquestes etiquetes POS, creant grups ordenats de tres etiquetes tal com proposen, de forma similar, Zhang *et al.* (2006).

En els casos de bigrames i trigrames es consideren delimitadors a l'inici i al final de les locucions que permeten determinar la posició de les paraules inicials i finals. El procés de creació d'aquestes agrupacions pot veure's il·lustrat en la figura 3.10.

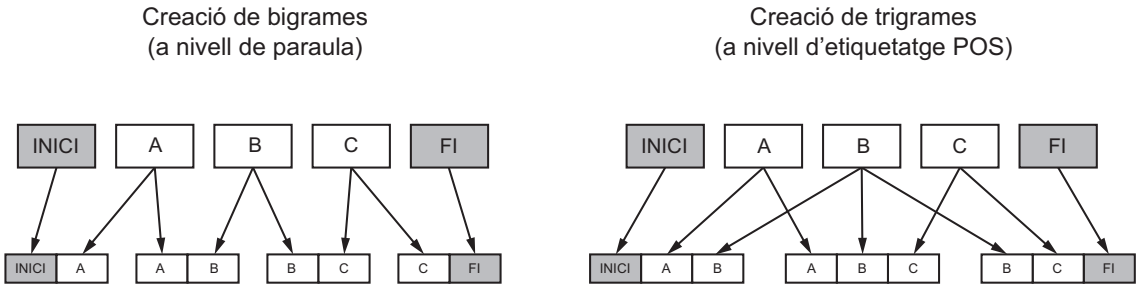


Figura 3.10: Procés de creació dels bigrames i trigrames per als elements de les locucions del corpus. En aquesta figura es mostra, com a exemple, una locució de 3 elements (paraules o etiquetes POS, segons correspongui). Noti's l'addició de les etiquetes delimitadores a l'inici i al final de la locució.

En tots els casos (unigrames, bigrames i trigrames) la parametrització consisteix en l'extracció dels 5 paràmetres d'activació descrits en la secció 3.3.2.1, un per a cadascuna de les 5 categories emocionals d'aquest corpus. Per mantenir la independència entre els conjunts d'entrenament (Ohm) i de prova (Mont), la llista d'elements emocionals empleada per al seu càlcul (aquells que superen un determinat valor llindar) es calcula considerant únicament el conjunt d'entrenament. Així mateix, els paràmetres d'activació es calculen en tots dos conjunts però considerant només els valors de prominència emocional i probabilitats a priori del conjunt d'entrenament. Per determinar el llindar anterior s'analitzen les paraules més freqüents, és a dir, les que més apareixen el conjunt d'entrenament. Per a això s'analitza l'índex d'ocurrència de cada paraula definint com a freqüents aquelles l'índex de les quals se situa per sobre del tercer quartil. La mitjana dels valors de prominència emocional d'aquestes paraules defineixen el valor llindar que, per a aquest corpus és de 0,3.

El total de paràmetres lingüístics segons la parametrització aquí detallada és de 15 paràmetres (5 paràmetres d'activació per a cadascun dels conjunts de unigrames i bigrames a nivell de paraula i de trigrames a nivell POS).

3.5 Resum

En aquest capítol s'han tractat les consideracions generals per a la creació de corpus de veu amb contingut emocional, detallant quatre tipus d'expressió vocal: natural, induïda, estimulada i actuada. Si bé la natural és la més realista, l'actuada permet un major control de les condicions d'enregistrament i del contingut del corpus, sent la induïda i l'estimulada alternatives a mig camí entre les anteriors.

Així mateix, s'han descrit diferents paràmetres que poden ser extrets de forma automàtica o semiautomàtica dels corpus de veu a dos nivells: acústic i lingüístic, segons es consideri el senyal de veu de forma aïllada o bé s'analitzi el contingut semàntic de les locucions gravades.

Finalment s'han detallat els tres corpus estudiats en aquesta tesi, presentant una descripció dels mateixos i explicant els paràmetres que, de cadascun, s'han extret per a la seva anàlisi en els experiments desenvolupats als següents capítols.

Capítol 4

Estat de la qüestió

Les darreres dècades han estat testimoni del creixement dels estudis de reconeixement afectiu automàtic. L'augment de la potència de càlcul dels ordinadors han permès analitzar corpus més extensos i realitzar treballs basats en conjunts de paràmetres cada vegada més grans. Els principals objectes d'anàlisi han estat el senyal de veu, l'expressivitat facial i els senyals fisiològics, essent les dues primeres modalitats les més fàcils d'enregistrar mitjançant sensors no invasius.

En aquest capítol s'analitzen alguns dels treballs més rellevants en l'àmbit del reconeixement afectiu basat en el senyal de veu, objecte d'aquesta tesi, sense deixar d'observar alguns estudis importants realitzats sobre la base d'altres modalitats i la seva anàlisi conjunta mitjançant tècniques de fusió. Es fa un recorregut per les primeres aproximacions al reconeixement automàtic fins a arribar a les tendències en les quals es basen els estudis més recents, assenyalant també quines són les principals aplicacions d'aquestes investigacions en àmbits, sovint, molt diferents entre si.

4.1 Antecedents del reconeixement emocional

El reconeixement afectiu automàtic ha cridat l'atenció de nombrosos investigadors en les darreres dècades (Zeng *et al.*, 2009). El treball de Suwa *et al.* (1978), basat en l'anàlisi de l'expressivitat facial, és un dels primers referents en el reconeixement automàtic d'emocions. De fet, l'emotivitat facial ha estat sempre un referent en les anàlisis de les emocions donada la claredat de la seva expressió tant en animals com en humans, mostrant-se a més com un tret intercultural (Darwin, 1872). No obstant això, no només l'expressivitat facial ha estat objecte d'anàlisi en el passat. Treballs com els de Williams i Stevens (1972) i Scherer (1979) són precursors en l'establiment d'unes referències sobre com les emocions s'expressen a través de la parla.

El 1989, l'eina *The Affect Editor* (Cahn, 1989) es converteix en una mostra de com es

pot realitzar una anàlisi emocional a partir de la síntesi de veu amb coloració afectiva. L'eina és una aplicació de síntesi de veu que implementa un model acústic de la parla i genera les instruccions per produir l'efecte indicat mitjançant la modificació de quatre categories de paràmetres. Les categories es refereixen al to, al ritme, a la VoQ i a l'articulació. El model acústic considera tots aquests paràmetres independents entre si. En els test perceptius duts a terme sobre les locucions creades mitjançant aquesta eina es va documentar una taxa de reconeixement d'emocions bàsiques del 78,7%. No obstant això, el sintetitzador de veu (al que s'adapta la parametrització del model acústic), i la incompleta descripció dels fenòmens acústics i perceptius limiten l'autenticitat de les emocions sintetitzades.

Intentat evitar els inconvenients del sistema anterior, Rodríguez *et al.* (1999) presenten un model de l'expressió emocional en espanyol, posteriorment validat per Iriondo *et al.* (2000). L'estudi neix de la hipòtesi que considera que la veu és susceptible a les alteracions fisiològiques produïdes en l'organisme en experimentar-se una emoció. Estudia set emocions bàsiques: alegria, desig, enuig, por, sorpresa, tristesa i fàstic. En una anàlisi acústica posterior s'observen els paràmetres referents a la F0, pressió sonora i ritme i s'estudia la representació gràfica de les frases en forma de oscil·lograma, corba de F0 i corba de pressió sonora. D'aquest estudi acústic es presenten models vàlids per a sis de les set emocions considerades, sent fàstic la que queda fora de la modelització proposada. La validació d'aquest model per Iriondo *et al.* (2000) consisteix en la seva aplicació a un convertidor text-veu capaç de modificar els paràmetres abans citats.

No obstant això, tot i que el treball de Iriondo *et al.* (2000) suposa un avanç important respecte al considerat inicialment per Cahn (1989)[1], l'anàlisi afectiva és realitzada manualment i no de forma automàtica. En el plantejament d'un model de reconeixement adaptat a un subjecte de manera individual és inviable considerar la repetició d'un estudi com l'abans indicat. Encara en el cas de pretendre un reconeixement genèric, readaptar el model anterior resultaria molt costós. És per això que el plantejament lògic consisteix a automatitzar el procés mitjançant tècniques d'aprenentatge artificial com les proposades per Litman i Forbes-Riley (2003) i Oudeyer (2003).

El projecte PHYSTA (Kollias i Piat, 1999), desenvolupat entre els anys 1998 i 2001, és una mostra de l'interès pel desenvolupament d'un sistema artificial de descodificació emocional mitjançant l'ús de tècniques d'intel·ligència artificial i xarxes neuronals. Alguns dels treballs desenvolupats en el marc d'aquest projecte introdueixen la consideració de les emocions com a elements multidimensionals. Actualment el camp de l'estudi emocional basat en el senyal de veu segueix sent àmpliament investigat tal com ho demostra la xarxa d'excel·lència finançada per la Unió Europea HUMAINE[2].

Donades les facilitats per al processament dels senyals audiovisuals, tant per a l'extracció de paràmetres com per a la seva posterior anàlisi mitjançant tècniques d'aprenentatge artificial, en la darrera dècada han augmentat els treballs vinculats al reconeixement

[1] En tant que es basa en un model acústic precís partint de l'exhaustiva anàlisi realitzada per Rodríguez *et al.* (1999).

[2] http://emotion-research.net

afectiu automàtic. El treball d'Oudeyer (2003) és el primer treball a gran escala que empra tècniques de mineria de dades per analitzar un conjunt extens de paràmetres vocals segons diversos algorismes d'aprenentatge. Existeixen diversos estudis que realitzen una recopilació i una anàlisi de treballs duts a terme dins d'aquest àmbit com els realitzats per Cowie *et al.* (2001) i Pantic i Rothkrantz (2003), centrats en l'anàlisi de l'expressió facial i vocal, i els més recents realitzats per Ververidis i Kotropoulos (2006) i Zeng *et al.* (2009), analitzant també l'expressió facial i vocal a excepció de l'últim que se centra en l'expressió vocal, únicament. Existeixen, no obstant això, grans diferències entre els estudis de reconeixement que empren corpus actuats i els que empren corpus espontanis, així com també dificultats a l'hora de comparar els diversos tipus de treballs realitzats donada la multitud de mètriques, entorns i aplicacions en les quals es fonamenten els diversos investigadors (Schuller *et al.*, 2011).

4.2 Aplicacions del reconeixement emocional automàtic

En general, la inclusió d'intel·ligència emocional en eines d'intel·ligència artificial pot millorar els processos d'interacció persona-màquina (Picard *et al.*, 2001; Cowie *et al.*, 2001). No obstant això, aquesta afirmació és massa general. Es poden trobar casos concrets en els quals els estudis de reconeixement afectiu automàtic (o, en general, els intents de reproduir el canal implícit de la comunicació humana en sistemes automàtics) tenen un objectiu específic molt bé definit. Cowie *et al.* (2001) descriuen diversos camps d'aplicació enumerant-los dels d'àmbit més general als d'àmbit més concret: convergència, interacció entre els canals de comunicació, assistència a l'avaluació, desambiguació, producció, tutories, desviament, alerta i entreteniment. Si bé molts estudis es duen a terme de forma general, una part dels mateixos pot situar-se dins dels camps anteriors.

Convergència i interacció entre els canals de comunicació. La detecció de l'estat afectiu de l'usuari permet a una aplicació convergir en una sèrie de paràmetres vocals, aspecte que en una comunicació entre humans mostra simpatia i que, de no produir-se, denota indiferència o falta d'empatia. D'altra banda, el canal implícit indica com ha de considerar-se el missatge del canal explícit, atès que unes mateixes paraules, per exemple, poden considerar-se de forma seriosa o com un acudit segons la manera en la qual es pronuncien. Una aplicació ha d'interpretar el canal implícit per poder gestionar correctament una conversa.

En aquest àmbit, l'objectiu del qual és afavorir la comunicació fluïda entre la persona i la màquina, se circumscriuen treballs com els de Cañamero i Fredslund (2001); Lee i Narayanan (2005); Maria i Zitar (2007); de Melo *et al.* (2010) o el descrit per Pous i Ceccaroni (2010) en el marc del projecte *INterfaces de RElación entre el entorno y las personas con DIScapacidad* (INREDIS)[3].

[3]Projecte CEN-20072011 emmarcat en la iniciativa INGENIO 2010 del govern espanyol.

Assistència a l'avaluació humana. Una de les aplicacions del reconeixement afectiu automàtic és la valoració o l'enjudiciament d'una persona, per exemple en l'assistència mèdica en la detecció de l'esquizofrènia (McGilloway *et al.*, 2003; Rapcan *et al.*, 2010) o en la detecció de mentides (Gadallah *et al.*, 1999; Enos, 2009; Mihalcea i Strapparava, 2009). No obstant això, existeix controvèrsia sobre els sistemes comercials destinats a aquesta darrera aplicació (Eriksson i Lacerda, 2007).

Desambiguació. Atès que l'estat afectiu modifica els trets prosòdics de la veu i que aquests també es veuen afectats pel moment específic en el qual es troba una conversa (per exemple, en una fase inicial, de negociació o d'acord final), una aplicació ha d'interpretar adequadament aquests senyals per reconèixer el missatge subjacent de la conversa. En l'àmbit d'una negociació, per exemple, no és el mateix negociar amb un agent enfadat que amb un alegre i aquest comportament pot reflectir-se també en la interacció persona-màquina (Broekens *et al.*, 2010; Ochs i Prendinger, 2010; de Melo *et al.*, 2011).

Producció. El reconeixement afectiu pot emprar-se per detectar i aprendre les subtileses de la parla emocionalment acolorida per a, posteriorment, aplicar-les en sistemes de síntesi de veu expressiva (Schröder *et al.*, 2001; Přibil i Přibilová, 2009; Barra, 2011), o bé per dotar de l'expressivitat adequada a les locucions generades per sistemes de conversió de text a veu (Bhutekar i Chandak, 2012).

Tutors automàtics. En aplicacions de formació i educació (sobretot en les no presencials) pot resultar interessant disposar de tutors automàtics que guiïn als alumnes al llarg del seu aprenentatge. En aquest tipus d'aplicacions, la identificació del seu estat afectiu permetria determinar el nivell de comprensió de les matèries i detectar si els exemples proposats són ben rebuts o, per contra, resulten confusos, avorrits o repetitius (Litman i Forbes-Riley, 2003, 2006; Hollingsed i Ward, 2007; Litman *et al.*, 2012). D'aquesta forma el sistema podria adaptar els continguts de manera personalitzada a cada alumne en funció del seu estat.

Desviament. En aplicacions que realitzen tasques de funcionariat (assistents personals, proveïdors d'informació, etc.) el reconeixement afectiu pot derivar les peticions dels usuaris que escapen al domini de l'aplicació a altres aplicacions o empleats humans preparats per manejar-les, o proporcionar noves respostes adequades a la situació detectada. Aquestes aplicacions són d'interès, per exemple, en centres de trucades (Morrison *et al.*, 2007; Gupta i Rajput, 2007), o com un mòdul addicional als sistemes de tutoria automàtica (Kapoor *et al.*, 2007).

Alerta. En ocasions és necessari alertar de l'estat afectiu d'un individu l'atenció del qual està posada en altres circumstàncies, com per exemple un usuari d'un servei d'informació

que necessita ajuda urgent (com el servei desenvolupat en el marc del projecte INREDIS), o bé és interessant detectar l'estat afectiu d'un individu com a mesura de seguretat (Bullington, 2005; Clavel *et al.*, 2008; Wang i Fang, 2008).

Entreteniment. El sector de l'entreteniment seria, probablement, la principal aplicació comercial del reconeixement afectiu automàtic, podent-se emprar en videojocs que reconeguin l'estat emocional del jugador (Yildirim *et al.*, 2011), que detectin si el jugador s'està avorrint (Giakoumis *et al.*, 2011), en "jocs seriosos" amb aplicació, per exemple, en tractaments psicològics (Kostoulas *et al.*, 2012), personatges de mons virtuals (Gratch i Marsella, 2001), mascotes virtuals que adaptin la seva activitat a l'estat anímic del seu amo mostrant comportaments més complexos i emulant empatia (Picard, 2000) o robots que interpretin l'estat emocional del seu propietari (Oudeyer, 2003).

4.3 Conceptes previs

Com s'ha vist en la secció 4.1 i com es veurà a continuació amb més detall, el reconeixement afectiu automàtic pot ser abordat des de diferents perspectives. Això afavoreix múltiples plantejaments per a un mateix tema que, sovint, dificulten la comparativa entre els diferents treballs duts a terme en aquest àmbit de recerca.

Les dues arquitectures predominants en el reconeixement afectiu automàtic són el processament dinàmic dels descriptors de baix nivell i el processament estàtic usant informació suprasegmental a través de l'extracció de descriptors estadístics d'aquests mateixos descriptors de baix nivell (Schuller *et al.*, 2009). Aquesta tesi se centra en el processament estàtic, la qual cosa suposa l'ús de diversos algorismes d'aprenentatge (tècniques de classificació per a estudis categòrics o de regressió per a estudis dimensionals, habitualment) els quals han de ser analitzats de forma adequada. A continuació es detallen alguns aspectes comuns en la major part dels estudis d'aquest àmbit.

4.3.1 Principals modalitats analitzades en el reconeixement afectiu automàtic i procediments per a la seva fusió

Existeixen estudis de reconeixement afectiu automàtic basats en diverses modalitats si bé les tres principals són el senyal de veu, l'expressió facial i els senyals fisiològics. Tal com s'ha detallat, a nivell vocal es poden considerar anàlisis prosòdics, de VoQ i, fins i tot, a nivell lèxic, extraient informació del contingut del missatge expressat mitjançant la parla. Aquestes són modalitats pròpies de la parla i el seu estudi, malgrat estar centrat únicament en el senyal de veu, pot considerar-se multimodal i emprar tècniques de fusió pròpies d'aquest àmbit.

De forma general, la multimodalitat entesa a nivell d'expressió a través de senyals

diferents és una forma natural de comunicació (Oviatt, 2002; Sebe *et al.*, 2005). Així doncs, un sistema podria reconèixer la parla i els gestos de l'usuari per poder interactuar amb ell a través d'avatars virtuals capaços de sintetitzar veu emocionada i expressar, de forma visual, estats afectius concrets. Oviatt (2002) defineix una sèrie de conceptes bàsics relacionats amb la multimodalitat entre els quals destaquen la diferenciació entre modes d'entrada actius (aquells que són emprats intencionadament per l'usuari per dirigir una ordre explícita a la interfície) i els passius (els relacionats amb la conducta i les accions que l'usuari desenvolupa de forma natural i que són reconeguts —idealment mitjançant sensors no invasius— per la interfície, com per exemple l'expressió facial o la gesticulació manual), el concepte d'interfícies multimodals mixtes (que incorporen almenys un mode d'entrada actiu i un altre passiu), les interfícies en cascada (que processen dos o més modes de l'usuari de forma seqüencial en el temps), i el concepte de desambiguació mútua (la que permet discriminar la informació en sistemes que compten amb una modalitat propensa a l'error fent ús d'altres modalitats més robustes). La figura 4.1 mostra l'esquema d'un sistema bimodal (Sebe *et al.*, 2005), si bé podria completar-se afegint-li altres modalitats com la tàctil o la relacionada amb senyals fisiològics.

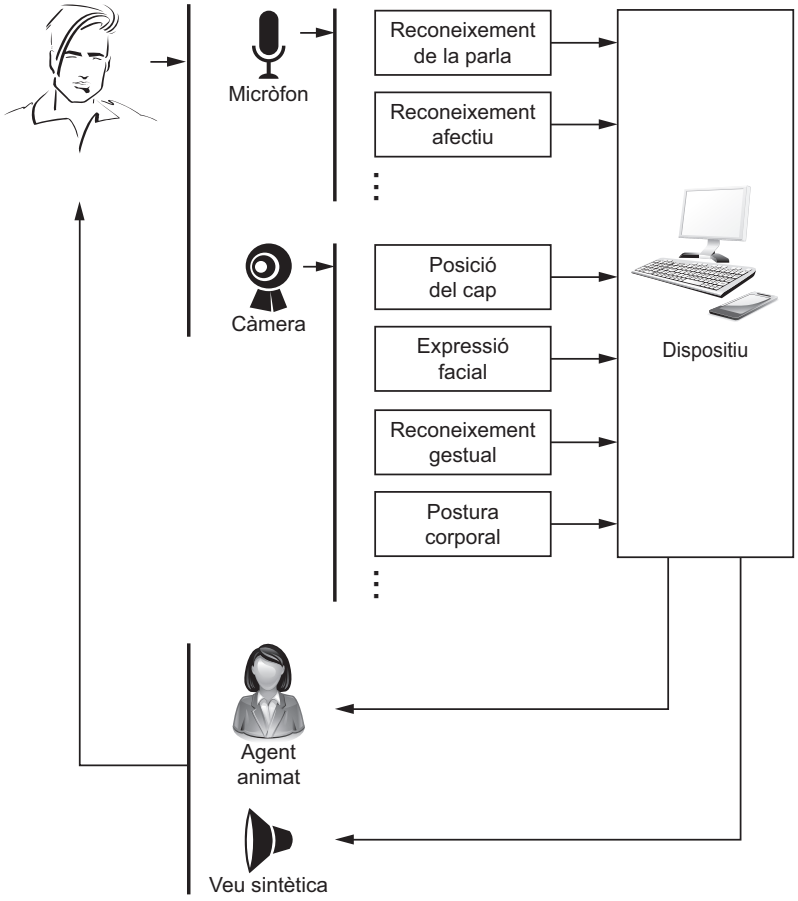


Figura 4.1: Esquema d'interacció persona-màquina bimodal. Adaptat de (Sebe *et al.*, 2005).

L'anàlisi dels diversos senyals d'entrada és el que s'entén per fusió multimodal[4]. Ja sigui l'anàlisi de senyals procedents de diversos sensors o bé d'aspectes derivats d'un mateix sensor, existeixen diverses formes de realitzar aquesta fusió. Zeng *et al.* (2009) descriuen els tres tipus de fusió més representatius en aquests treballs:

- Fusió a nivell de paràmetres. També coneguda com a fusió primerenca[5]. Consisteix a fusionar els paràmetres abans que aquests siguin processats pels algorismes d'aprenentatge en una etapa posterior.
- Fusió a nivell de decisió. També coneguda com a fusió tardana[6] o a nivell de classificadors. Consisteix a processar independentment cada modalitat i després fusionar els resultats en l'etapa final. El més comú és emprar una estratègia de *stacking* (vegeu també la secció 4.3.3.4). El mètode de *stacking* més simple és una votació que pot ser ponderada o no. Altres tècniques més complexes intenten aprendre una sèrie de regles per part d'un classificador específic (que rep el nom de classificador de nivell 1), que millorin la classificació de cada classificador individual (classificadors de nivell 0) (Witten i Frank, 2005).
- Fusió híbrida. Intenta combinar els beneficis dels dos tipus previs de fusió si bé és un model menys explorat que els anteriors.

Tal com s'ha explicat, aquestes tècniques de fusió no són exclusives per a dades procedents de diferents sensors i poden emprar-se per fusionar informació procedent de diferents anàlisis d'una mateixa modalitat àmplia com, per exemple, la parla, tal com es veurà en les següents seccions. A més, tècniques com el *stacking* permeten també fusionar resultats de diversos classificadors que analitzen un mateix conjunt de dades però sobre la base de diferents plantejaments o mitjançant tècniques diferents. En un cas com aquest no es pretén fusionar la informació de diversos conjunts de dades sinó aprofitar la redundància dels classificadors independents per aconseguir major robustesa en el resultat a partir de la seva combinació (Bezdek *et al.*, 1999).

4.3.2 Selecció de paràmetres

Els corpus, sovint, compten amb parametritzacions exhaustives que suposen un gran volum de dades. Aquest volum es veu incrementat, a més, si s'incorpora l'estudi de diverses modalitats. D'altra banda, a un algorisme de classificació li pot resultar difícil processar un volum molt gran de dades de forma eficient i més si existeix redundància. En moltes ocasions, un nombre elevat de paràmetres porta associat un subconjunt de dades que podrien considerar-se irrellevants i que podrien ser eliminades sense que es degradés

[4]Per contra, la fissió multimodal fa referència a la unió de diversos elements a la sortida del sistema (veu sintètica, gràfics, etc.).

[5]De l'anglès, *early fusion*.

[6]De l'anglès, *late fusion*.

la taxa de classificació o arribant, fins i tot, a millorar-la (Witten i Frank, 2005). Existeixen tècniques que, de forma automàtica, poden reduir la dimensionalitat d'un conjunt de dades seleccionant els paràmetres més rellevants. Aquestes tècniques, de forma general, poden seguir dos enfocaments (Langley, 1994; Guyon i Elisseeff, 2003): mètodes de filtre[7] i mètodes d'embolcall[8].

- Mètodes de filtre. Es basen únicament en les característiques de les dades. Els paràmetres se seleccionen abans d'iniciar-se el procés d'aprenentatge mitjançant un algorisme específic. Aquesta selecció és independent del classificador posterior. Un exemple de mètode de filtre és el que segueix el criteri de mínima redundància i màxima rellevància (*Minimal-Redundancy Maximal-Relevance* (mRMR)) (Peng *et al.*, 2005), que té com objectiu seleccionar paràmetres mútuament excloents la rellevància dels quals sigui la més gran possible envers l'etiqueta categòrica de cada cas.
- Mètodes d'embolcall. Es creen subconjunts de paràmetres mitjançant un procés de cerca que després són avaluats mitjançant un classificador específic. En aquest cas la complexitat computacional és més elevada que en els mètodes de filtre. Un exemple de mètode d'embolcall són els algorismes de cerca voraç[9].

Molts mètodes de selecció de paràmetres en general i els de embolcall en particular porten associats una fase de cerca en l'espai de paràmetres. Aquesta fase crea subconjunts de paràmetres candidats a ser els que millor prediuen la classe final. És per això que es fa necessària una altra fase, la d'avaluació, per mesurar la capacitat predictiva d'aquests subconjunts. Aquesta fase d'avaluació pot emprar el mateix classificador que s'emprarà posteriorment durant l'explotació del sistema o bé un algorisme específic (Witten i Frank, 2005).

Un dels mètodes de cerca més habituals és el mètode de cerca voraç. Aquests mètodes són iteratius i poden ser incrementals o decrementals, depenent de la direcció que es prengui per recórrer l'espai de paràmetres. Així doncs, en cada iteració es realitza un canvi local en el subconjunt actual afegint o eliminant un paràmetre. En la cerca incremental el procés s'inicia sense paràmetres i s'afegeix un en cada pas mentre que en la cerca decremental el procés s'inicia amb el conjunt de dades complet i s'elimina un en cada iteració. La fase d'avaluació té lloc en cada pas en el qual s'afegeix o s'elimina, temptativament, un paràmetre, analitzant els subconjunts resultants. La fase d'avaluació genera una mesura del rendiment previst dels subconjunts. En cada pas s'escull el millor i el procés continua. El procés s'atura quan l'addició o l'eliminació d'un paràmetre, en els casos de cerca incremental i decremental, respectivament, no milloren el rendiment dels subconjunts. També pot considerar-se una estratègia mixta que combini tots dos mètodes de cerca creant una de tipus bidireccional com la proposada per Iriondo *et al.* (2007b), en la qual es defineix una estratègia de selecció d'atributs basada en una cerca voraç de p passos cap endavant

[7] De l'anglès, *filter methods*.
[8] De l'anglès, *wrapper methods*.
[9] De l'anglès, *greedy search*.

(fase de procés incremental) i q passos cap enrere (fase de procés decremental). Aquesta estratègia es representa per l'acrònim pFW-qBW[10], éssent p i q dos nombres enters positius que verifiquen que $p > q$.

Un altre mètode de cerca emprat habitualment, i que proporciona bons resultats en experiments de reconeixement afectiu (Oudeyer, 2003; Planet *et al.*, 2006), és l'algorisme genètic que segueix l'esquema clàssic descrit per Goldberg (1989). En la fase d'avaluació vinculada a aquest mètode de cerca es pot emprar, novament, qualsevol algorisme de classificació si bé pot resultar convenient utilitzar un algorisme que no sigui el que finalment s'empri com a classificador en el sistema de reconeixement. Emprant un classificador diferent es pot crear un conjunt de dades òptim sobre la base d'alguna propietat determinada i que sigui, per tant, més genèric, intentant així evitar que es creï un vincle massa estret amb el classificador escollit. En aquest sentit, l'algorisme descrit per Hall (1998) permet avaluar els diferents individus de les poblacions generades per l'algorisme genètic[11] afavorint a aquells els paràmetres que estan altament correlats amb la classe i presenten, alhora, una baixa correlació creuada. Aquest és el mètode escollit en els experiments de la present tesi en els quals s'aplica aquest algorisme genètic.

4.3.3 Algorismes de classificació

Com s'ha comentat a l'inici d'aquesta secció, en aquesta tesi s'adopta el processament estàtic dels descriptors de baix nivell. En els propers capítols es proposen algunes estructures de classificació complexes basades en algorismes més senzills els quals també s'analitzen de forma independent. A continuació s'introdueixen alguns d'aquests algorismes de classificació, considerant aquells que mostren una major adequació a l'estudi del reconeixement afectiu automàtic segons els treballs realitzats per diversos investigadors en l'estat de la qüestió tal com es recull en la secció 4.4. Les explicacions aquí presentades no pretenen ser exhaustives sinó que la seva intenció és servir com a referència i es remet al lector a les cites bibliogràfiques per obtenir una visió completa d'aquests algorismes de classificació.

4.3.3.1 Arbres, regles i taules de decisió

Un arbre de decisió funciona avaluant els paràmetres pels quals es defineix un cas a classificar en cadascun dels nodes que formen el model. Els casos que, partint de l'*arrel*, arriben a una determinada *fulla* reben la classificació que aquesta indica. L'algorisme es pot expressar recursivament i consisteix a seleccionar un paràmetre a partir del qual fer

[10] FW fa referència a la fase de procés incremental (de l'anglès, *forward*) i BW fa referència a la fase de procés decremental (de l'anglès, *backward*).

[11] En la terminologia de la computació evolutiva una població és un conjunt d'individus que evoluciona segons regles similars a l'evolució biològica. Aplicada en l'àmbit de la selecció de paràmetres, un individu és un subconjunt concret del conjunt de paràmetres original.

una subdivisió de les dades obtenint tants subconjunts com a valors possibles pugui prendre aquest. El procés es repeteix iterativament per a cada divisió fins que tots els casos d'una mateixa branca reben la mateixa classificació o s'aconsegueix un llindar predefinit. L'algorisme J4.8 (Witten i Frank, 2005) és un algorisme de creació d'arbres de decisió basat en entropia que adapta l'algorisme C4.5 (Quinlan, 1993). C4.5 és, a la vegada, el resultat d'una sèrie de millores de l'algorisme ID3 (Quinlan, 1986) que, per determinar els paràmetres que defineixen cadascun dels nodes, utilitza els conceptes d'informació i guany d'informació tal com es descriu a continuació.

Sigui S un conjunt de casos en el qual es defineixen dues classes, P i N, siguin p i n els respectius nombres de casos de les classes anteriors, A el paràmetre proposat per definir un node de l'arbre, v el nombre de subconjunts de S que es formen en el node definit per aquest paràmetre, i p_i i n_i el nombre de casos del subconjunt S_i ($\forall i \in [1, v]$). El guany d'informació del paràmetre A es denota per $G(A)$ i es defineix per l'equació 4.1. El paràmetre que maximitza l'expressió de $G(A)$ definirà el node[12].

$$G(A) = I(p, n) - I(A) \tag{4.1}$$

On:

$$I(p, n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \tag{4.2}$$

$$I(A) = \sum_{i=1}^{v}\frac{p_i + n_i}{p+n}I(p_i, n_i) \tag{4.3}$$

Les regles de decisió poden considerar-se, en part, similars als arbres abans descrits (Witten i Frank, 2005) i, sovint, es creen a partir d'ells. Un algorisme comunament emprat és l'algorisme PART (Frank i Witten, 1998) el qual crea regles a partir d'arbres de decisió podats. Mitjançant aquesta poda s'aconsegueix reduir el nombre de regles creades, per la qual cosa presenta un bon rendiment tot i que el nombre de casos d'entrenament pugui ser elevat. Consisteix en la creació d'una regla que considera la fulla d'un arbre de decisió que cobreix el màxim nombre de casos. A continuació s'eliminen tots aquells que ja són abastats per aquesta regla i es repeteix el procés amb els casos restants fins que s'ha iterat amb tots els disponibles.

Per la seva banda, les taules de decisió solen ser un mètode rudimentari per representar la informació d'un algorisme d'aprenentatge i ho fan representant-la tal com són les dades d'entrada. No és trivial, no obstant això, determinar què paràmetres són irrellevants per poder resumir aquestes taules. Així, per exemple, Kohavi (1995) descriu un algorisme de creació de taules de decisió que empra un mètode d'embolcall (vegeu

[12] Per tal que l'algorisme J4.8 pugui tractar paràmetres numèrics s'adopta una solució que permet divisions binàries en els nodes (Witten i Frank, 2005).

la secció 4.3.2) per determinar el millor conjunt de paràmetres que defineix la taula de decisió.

4.3.3.2 Aprenentatge basat en casos

Els classificadors basats en casos[13] emmagatzemen els casos del conjunt de dades d'entrenament convenientment processats (Aha *et al.*, 1991). Per classificar una nova mostra s'empra una funció de distància que avalua què cas o casos del conjunt d'entrenament són els més propers a ella. Per això sol referir-se a ells com a algorismes *k-Nearest Neighbour* (KNN) o dels *k* veïns més propers. De forma general, la nova mostra es classifica amb l'etiqueta majoritària dels *k* casos més propers, encara que també és habitual considerar el valor concret de $k = 1$ assignant a la nova mostra l'etiqueta del cas més proper.

La principal diferència entre els algorismes d'aquest tipus és la funció de distància escollida. Així, per exemple, és habitual l'ús de la distància euclidiana o de la distància Manhattan per mesurar la proximitat entre els casos, encara que poden considerar-se unes altres com en el cas de l'algorisme K* (Cleary i Trigg, 1995), que usa una funció de distància basada en entropia.

4.3.3.3 Classificadors probabilístics i basats en funcions

Naïve-Bayes és un classificador probabilístic que parteix de la premissa que cada parell paràmetre-valor d'un mateix cas és independent de la resta[14]. A cada parell paràmetre-valor se li assigna una probabilitat de pertinença a una classe. Per a això es divideix el nombre de casos de cada classe en els quals apareix aquest parell entre el nombre de casos que pertanyen a aquesta classe. Per classificar un cas nou es calcula la probabilitat de pertinença d'aquest cas a cada classe, classificant-ho en la classe on aquesta probabilitat sigui major adoptant un criteri d'estimació *Maximum A Posteriori* (MAP). Aquesta probabilitat de pertinença es calcula com el producte de la probabilitat de pertinença a cada classe de cadascun dels parells paràmetre-valor que defineixen el cas a classificar.

L'algorisme Naïve-Bayes es basa en la regla de Bayes de probabilitat condicional expressada en l'equació 4.4, la qual relaciona la probabilitat de la hipòtesi H donada l'evidència E a través de l'expressió $p(H|E)$ —en la qual la hipòtesi H és la classificació d'un cas en una classe concreta i l'evidència E és la descripció d'aquest cas definit per $E = (E_1, E_2, \ldots, E_n)$, on E_i és cadascun dels parells paràmetre-valor— amb la probabilitat de l'evidència donada la hipòtesi $p(E|H)$ i la probabilitat a priori de la hipòtesi $p(H)$, sent aquestes dues últimes probabilitats fàcilment calculables a partir l'observació dels casos ja coneguts i etiquetats tal com s'ha descrit en el paràgraf anterior i com es descriu en l'equació 4.5[15].

[13] De l'anglès, *instance based*.

[14] D'aquí el nom del classificador: *naïve* (de l'anglès, ingenu).

[15] El valor de $p(E)$ no és rellevant atès que es podrà ignorar en normalitzar els resultats.

$$p(H|E) = \frac{p(E|H)p(H)}{p(E)} \tag{4.4}$$

$$p(E|H) = p(E_1|H)p(E_2|H)\dots p(E_n|H) \tag{4.5}$$

En el cas que algun paràmetre sigui numèric se li assumeix una distribució de probabilitat normal, tal com s'expressa en l'equació 4.6, considerant que el paràmetre P_i adopti el valor x, on μ i σ són la mitjana i la desviació estàndard de la distribució, respectivament[16].

$$p(P_i = x|H) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{4.6}$$

Un altre mètode comunament emprat en experiments de classificació afectiva són els classificadors SVM. De forma resumida, es poden descriure com a esquemes de classificació que amplien les característiques dels models lineals permetent realitzar la distinció entre classes que presenten límits de decisió no lineals. Per a això es transformen les dades originals mapejant-les de forma no lineal en un nou espai de dimensió superior. En aquest nou espai es construeix un model lineal que sigui capaç de representar un límit de decisió no lineal en l'espai original. Aquest model lineal rep el nom de hiperplà de marge màxim i permet realitzar la separació màxima entre les dades pertanyents a dues classes diferents. Les mostres més properes al hiperplà reben el nom de vectors de suport.

Assumint que la classe pugui prendre únicament els valors $\{-1, 1\}$ l'expressió del hiperplà de marge màxim és la descrita per l'equació 4.7, on y_i és la classificació de la mostra d'entrenament a_i i a és la mostra a classificar. Ambdues mostres estan expressades com a vectors i multiplicades segons el producte escalar. Els vectors a_i són els vectors de suport. Els termes α_i i b són paràmetres a determinar per l'algorisme d'entrenament abordant un problema d'optimització quadràtica amb restriccions, el qual pot ser realitzat de forma eficient si s'adopten algorismes específics per a aquest tipus d'esquemes d'aprenentatge com, per exemple, el d'optimització seqüencial mínima (Platt, 1998). $K(x, y)$ és la funció de transformació i rep el nom de *kernel*. La seva finalitat és transformar les dades per facilitar la seva separació per l'hiperplà de marge màxim. Pot ser de diversos tipus sent els polinòmics els més habituals.

$$x = b + \sum_{i=1}^{n} \alpha_i y_i K(a_i \cdot a) \tag{4.7}$$

Altres sistemes de classificació habituals són les xarxes neuronals i, en menor mesura, els sistemes de regressió logística.

[16] No obstant això, aquesta assumpció no és correcta en la majoria dels casos i es comprova una substancial millora en els resultats de classificació si els paràmetres numèrics es discretitzen prèviament.

4.3.3.4 Metamètodes

Els metamètodes són esquemes de classificació que combinen les decisions de diferents models per assignar una única classe a cada cas classificat. Els tècniques generals són *bagging*, *boosting* i *stacking*.

La tècnica *bagging*[17] intenta neutralitzar la inestabilitat d'un algorisme de classificació i consisteix a crear, de forma aleatòria a partir del conjunt de dades d'entrenament, subconjunts amb els quals es creen diferents models mitjançant la seva avaluació amb un classificador específic. Cada model es valora igual que els altres i la classificació final per a una mostra consisteix a escollir la més freqüent de les predites per cada model.

La tècnica *boosting* intenta millorar la precisió d'un model d'aprenentatge. L'algorisme descrit per Freund i Schapire (1996) és també iteratiu però en aquesta ocasió s'analitza el conjunt de dades complet i cada model construït té en compte l'error dels models anteriors. Per a això s'assignen pesos iguals a tots els casos del conjunt d'entrenament i, en cada iteració, els pesos s'actualitzen segons l'error de classificació del model en funció de si el cas es classifica correctament o incorrectament. Finalment, per realitzar la predicció, els models són combinats usant un vot ponderat. El pes de la predicció de cada model per a una mostra es calcula en funció de l'error de classificació e segons l'equació 4.8.

$$e = -\log \frac{e}{1-e} \tag{4.8}$$

La tècnica *stacking*[18] difereix de les dues anteriors en què, en general, no combina models creats amb el mateix algorisme de classificació sinó que s'aplica a models construïts a partir de classificadors de nivell 0 diferents. Existeixen diversos mètodes per realitzar aquesta combinació. Els dos més empleats són la tècnica de vot majoritari, en la qual la classe assignada a cada cas es correspon amb la que la major part dels classificadors assigna de manera individual, o emprant un classificador addicional de nivell 1 que actua d'àrbitre. Aquest classificador s'entrena de forma supervisada amb els resultats de classificació individuals dels altres classificadors per, posteriorment, assignar una classe a cada nou cas.

4.3.4 Mesures del rendiment dels algorismes de reconeixement

Els resultats dels algorismes d'aprenentatge responsables de realitzar la classificació d'un conjunt de casos, la qual cosa suposa l'assignació una etiqueta categòrica concreta a partir d'un conjunt de casos prèviament analitzats, poden ser mesurats conforme a diverses mètriques. Les més habituals en estudis de reconeixement afectiu són les que es detallen a continuació.

[17] Abreviatura de *bootstrap aggregating*.
[18] Abreviatura de *stacked generalization*.

4.3.4.1 Mètriques per categoria

Hi ha tres mètriques principals per mesurar el rendiment d'un classificador estudiant per separat cada categoria emocional: precisió, cobertura i mesura F1. Aquestes mètriques sovint solen ser una dada informativa addicional a altres mesures com les taxes de classificació globals, com succeeix en els treballs de Litman i Forbes-Riley (2003), Grimm *et al.* (2007) i Busso *et al.* (2009).

Precisió. La precisió mesura quina quantitat d'informació de la que retorna el classificador és correcta, expressada com el percentatge de casos d'una categoria concreta que són correctament classificats respecte al total de casos que reben (correctament o no) aquesta mateixa etiqueta categòrica. L'equació 4.9 mostra el càlcul de la precisió de la categoria c on VP_c representen els casos d'aquesta categoria correctament classificats (veritables positius) i FP_c representen els casos d'altres categories que, incorrectament, reben aquesta etiqueta categòrica (falsos positius).

$$precisió_c = \frac{VP_c}{VP_c + FP_c} \tag{4.9}$$

Cobertura. La cobertura mesura quanta informació rellevant ha extret el classificador, expressada com el percentatge de casos correctament classificats en una categoria respecte al total de casos que realment pertanyen a aquesta categoria. L'equació 4.10 mostra el càlcul de la cobertura de la categoria c on, en aquest cas, FN_c representa el nombre de casos d'aquesta categoria que reben, incorrectament, una etiqueta categòrica diferent (falsos negatius).

$$cobertura_c = \frac{VP_c}{VP_c + FN_c} \tag{4.10}$$

Mesura F1. La mesura F1 intenta caracteritzar el rendiment d'un classificador mitjançant un únic valor. Per a això es calcula la mitjana harmònica dels paràmetres anteriors segons l'equació 4.11. L'equació 4.12 mostra l'expressió genèrica de l'equació anterior, segons es desitgi atorgar un major pes a la mesura de la precisió o a la de la cobertura, per a valors del paràmetre β dins dels nombres reals positius. Així, un valor de $\beta = 2$ suposa que la cobertura es pondera amb el doble de pes que la precisió mentre que un valor de $\beta = 0,5$ suposa que la precisió es pondera amb el doble de pes que la cobertura.

$$\mathrm{F1}_c = \frac{2 \times precisió_c \times cobertura_c}{precisió_c + cobertura_c} \tag{4.11}$$

$$\mathrm{F}\beta_c = \frac{(1+\beta^2) \times (precisió_c \times cobertura_c)}{\beta^2 \times precisió_c + cobertura_c} \tag{4.12}$$

4.3.4.2 Mètriques globals

Les mètriques globals considerades en aquesta tesi són la taxa de classificació ponderada i la taxa de classificació no ponderada. Ambdues valoren, de forma general, el rendiment d'un classificador sobre la base del nombre de casos correctament classificats, si bé presenten definicions diferents.

Taxa de classificació ponderada — *Weighted Average Recall* (WAR). La taxa de classificació ponderada, coneguda pel seu acrònim en anglès com WAR, es calcula com la proporció de casos correctament classificats respecte el nombre de casos totals, tal com es mostra en l'equació 4.13. Determina, bàsicament, el percentatge de casos que es classifiquen correctament. En aquest cas no es té en compte l'anàlisi detallada per classe ja que els veritables positius (VP) i els falsos negatius (FN) fan referència a la totalitat dels casos del conjunt de dades a classificar.

$$\mathrm{WAR} = \frac{VP}{VP+FN} = \frac{número\ de\ casos\ correctament\ classificats}{número\ total\ de\ casos} \tag{4.13}$$

En aquesta mètrica es valora el volum de casos correctament classificats sense considerar si procedeixen de classes majoritàries o minoritàries en el cas de conjunts de dades no equilibrats. És la mètrica més emprada en els treballs de reconeixement afectiu, com els de Polzin i Waibel (2000), Oudeyer (2003), Litman i Forbes-Riley (2003), Hassan i Damper (2009) i, en general, de tots aquells que analitzen corpus equilibrats.

Taxa de classificació no ponderada — *Unweighted Average Recall* (UAR). La taxa de classificació no ponderada, coneguda pel seu acrònim en anglès com UAR, a diferència de la WAR sí té en consideració el nombre de casos que formen part cada classe de manera que pot resultar més útil per mesurar el rendiment d'un classificador que treballa amb un conjunt de dades poc equilibrat. L'equació 4.14 mostra el càlcul d'aquesta mesura, on $|C|$ és el nombre de classes.

$$\mathrm{UAR} = \frac{\sum_{c=1}^{|C|} cobertura_c}{|C|} \tag{4.14}$$

Cal assenyalar que per al cas d'un conjunt de dades equilibrat, és a dir, amb el mateix nombre o un nombre molt similar de casos per a cada classe, les mesures WAR i UAR

són idèntiques. Això no és així si el conjunt de dades no està equilibrat. La mesura UAR, en aquest cas, calcula el rendiment del classificador valorant per igual totes les classes, ja siguin majoritàries o minoritàries.

Els treballs que empren la mesura UAR estan associats habitualment a estudis de corpus no equilibrats com, per exemple, els de Lee *et al.* (2009), Kockmann *et al.* (2009) i Vogt i André (2009), tots ells associats al corpus FAU Aibo.

Diferenciació entre ambdues taxes de classificació. De forma general, un valor elevat de WAR indica que un classificador és capaç de classificar correctament un gran nombre de casos, sense importar la classe a la qual pertanyin. D'altra banda, un valor elevat de UAR indica que un classificador és capaç de classificar correctament un gran nombre de casos de cada classe. Un valor elevat per a ambdues mesures indica que un classificador és capaç de classificar un gran nombre de casos i que, al mateix temps, ho fa de forma òptima per a cada classe. Si el valor de UAR és elevat però no ho és el valor de WAR, és senyal que el classificador classifica de forma òptima les classes minoritàries però no les majoritàries. Per contra, si el valor de WAR és elevat però no ho és el valor de UAR, és senyal que el classificador classifica de forma òptima les classes majoritàries però no les minoritàries.

L'elecció d'una o una altra mètrica per comparar els rendiments dels classificadors dependrà de l'aplicació i de la valoració que es faci de cada classe, no podent-se considerar una millor que una altra de forma general.

4.3.5 Mètodes d'avaluació dels algorismes de reconeixement

L'avaluació dels sistemes de classificació és un punt delicat doncs un disseny incorrecte pot implicar un resultat massa allunyat del rendiment que el sistema realment presentaria en el seu entorn d'explotació. Així, per exemple, un sistema avaluat amb les mateixes dades que s'han emprat per al seu entrenament resultarà en una taxa de classificació possiblement superior a la que s'hagués obtingut emprant dades diferents. Existeixen diverses tècniques estadístiques destinades a avaluar de forma adequada els sistemes de classificació, de les quals a continuació es detallen les més emprades en els treballs de reconeixement afectiu automàtic.

La validació creuada de diverses iteracions[19] consisteix a dividir el conjunt de dades en k subconjunts. En un procés iteratiu, cadascun d'aquests subconjunts es reserva per provar el sistema de classificació que s'entrena amb els $k-1$ subconjunts restants. El procés es repeteix k vegades i el rendiment final s'obté com la mitjana dels rendiments intermedis. Aquesta tècnica s'empra en estudis com els de Litman i Forbes-Riley (2003) i Oudeyer (2003). Un cas extrem d'aquesta tècnica, la validació creuada deixant un fora[20],

[19] De l'anglès, *k-fold cross-validation*
[20] De l'anglès, *leave-one-out cross-validation*

consisteix a considerar el valor de k igual al nombre d'elements del conjunt de dades de manera que en cada iteració només una dada s'empra per provar el classificador entrenat amb el conjunt complet excepte la dada reservada per a prova. Per exemple, Chen *et al.* (1998) i Luengo *et al.* (2005) empren aquesta tècnica. En l'extrem oposat es troba el cas de considerar el valor de $k = 2$, el qual crea només dos subconjunts de dades emprant un d'ells per a entrenament i l'altre per a prova. En aquest cas, la divisió no necessàriament crea dos subconjunts homogenis, com per exemple els treballs derivats a partir del de Schuller *et al.* (2009).

En sistemes de reconeixement en els que es consideren dades de diferents locutors sovint s'empra una variant del mètode de validació creuada en el qual es creen tantes subdivisions com a locutors, creant un sistema de validació creuada independent del locutor[21]. D'aquesta forma es garanteix que el conjunt de dades de prova no conté informació dels locutors del conjunt de dades d'entrenament. Així, si un conjunt de dades conté informació de x locutors, els casos d'un d'ells s'usa per provar el sistema de classificació i els casos dels $x - 1$ locutors restants s'empren per entrenar-ho prèviament. Aquest procés es repeteix per a tots els locutors. Aquesta és la tècnica emprada, per exemple, per Ruvolo *et al.* (2008), Lee *et al.* (2009) i Kotti *et al.* (2010). En estudis de reconeixement afectiu aquest sistema es pot portar un pas més enllà i crear subconjunts considerant també les categories emocionals, la qual cosa suposaria un esquema de validació creuada independent dels subjectes i de les emocions (Truong i van Leeuwen, 2007).

4.4 Estudis de reconeixement emocional basats en el senyal de veu

La veu és una modalitat important en la comunicació humana i porta associada informació afectiva tant al canal explícit (referent a l'àmbit lingüístic) com a l'implícit (referent a l'àmbit paralingüístic) (Zeng *et al.*, 2009). A nivell lingüístic, la informació emocional pot detectar-se en les pròpies paraules, per la qual cosa es poden crear diccionaris afectius (Plutchik, 1980; Whissell, 1989). No obstant això alguns estudis determinen que la informació lingüística no és sempre un mitjà fiable per determinar els estats afectius (Ambady i Rosenthal, 1992) sent, a més, altament depenent de l'idioma i donat l'ús subjectiu que pot fer-se del vocabulari (Furnas *et al.*, 1987). Quant al nivell paralingüístic, no existeix un conjunt concret de paràmetres que permeti realitzar una discriminació efectiva entre diversos estats afectius (Zeng *et al.*, 2009), si bé algunes emocions bàsiques poden destriar-se sobre la base de paràmetres prosòdics (Juslin i Scherer, 2005) i algunes no bàsiques poden interpretar-se a partir d'elements no lingüístics (com, per exemple, riures o badalls) (Russell *et al.*, 2003). No obstant això, en un entorn realista, la informació acústica per si sola podria no ser suficient per realitzar un reconeixement afectiu automàtic de forma eficaç (Batliner *et al.*, 2003).

[21] De l'anglès, *Leave One Speaker Out* (LOSO).

Molts dels estudis de reconeixement afectiu automàtic es basen en enregistraments realitzats per actors el que, sovint, no afavoreix l'anàlisi de la correlació existent entre els elements paralingüístics i el contingut lingüístic (Zeng *et al.*, 2009). És per això que es realitzen cada vegada més estudis basats en enregistraments naturals (com a centres de trucades, escenaris de *Mag d'Oz* i sistemes de diàleg en general) mitjançant l'ús de diversos algorismes de classificació (Koolagudi i Rao, 2012). Les dues seccions següents aprofundeixen en l'estat de la qüestió d'ambdues perspectives.

4.4.1 Reconeixement emocional d'emocions actuades

En la major part dels estudis de reconeixement afectiu automàtic basats en corpus actuats, les taxes de reconeixement acostumen a ser molt elevades, al voltant del 90% de WAR en els estudis més recents, si bé es tracta d'estudis depenents del locutor i amb enregistraments d'emocions plenes en la majoria de les ocasions. A continuació es detallen alguns treballs en els quals s'aprecia l'evolució en el plantejament dels experiments de reconeixement afectiu així com també dels resultats que s'obtenen, on es pot observar, com a principal element diferenciador dels més recents respecte als més antics, que es consideren conjunts de paràmetres molt més extensos i es realitzen proves amb un ventall més gran d'algorismes d'aprenentatge.

El treball de Banse i Scherer (1996) pretén mostrar com els humans poden determinar estats emocionals a partir, únicament, de l'expressió vocal, sense pretendre realitzar un experiment de reconeixement afectiu automàtic tal com s'entén en l'actualitat. El seu experiment parteix de l'anàlisi d'un corpus d'expressió vocal actuada gravat per 12 actors alemanys pronunciant textos sense significat i representant un total de 14 estats afectius. Considerant el seu disseny, el corpus està format per 1.344 locucions (agrupades posteriorment en 280 conjunts) parametritzades mitjançant 29 paràmetres acústics. En l'estudi es presenta una comparació entre el reconeixement emocional realitzat per un grup d'avaluadors humans i dos procediments de classificació estadística. La taxa de reconeixement per part dels jutges humans se situa al voltant del 50%, així com de manera similar també ho fan els procediments estadístics, si bé no existeix un mapatge entre el reconeixement humà i l'estadístic quan s'analitzen les classes emocionals per separat.

Polzin i Waibel (2000) realitzen un experiment de reconeixement afectiu automàtic basat en un corpus creat a partir de frases extretes de pel·lícules en anglès relatives a 3 estats emocionals: tristesa, empipament i estat neutre. La parametrització es realitza a nivell verbal (calculant la probabilitat d'una paraula donades les paraules anteriors) i a nivell oral (8 paràmetres relatius a la prosòdia i a la VoQ i 32 coeficients cepstrals seleccionats automàticament d'un conjunt major). L'experiment es configura sobre la base de dos subconjunts de dades, un d'entrenament i un altre de prova. L'estudi compara una classificació subjectiva (70% de taxa de reconeixement) i la realitzada automàticament segons els tipus de paràmetres de forma independent (60,4% segons la informació prosòdica i de VoQ, 63,9% segons la informació cepstral i 46,7% segons la informació verbal).

McGilloway *et al.* (2000) presenten un estudi de reconeixement afectiu automàtic en el qual s'extreuen descriptors de baix nivell i funcionals que van més enllà dels conjunts estàndard (incloent, per exemple, distribucions d'intensitat i longituds de segments sil·làbics). Aquests descriptors són majoritàriament prosòdics i són extrets mitjançant el sistema *Automatic Statistical Summary of Elementary Speech Structures* (ASSESS). També, a diferència d'altres estudis, és dels primers a emprar tècniques de mineria de dades per analitzar els conjunts de dades. El conjunt de tècniques emprades es redueix a 3 algorismes analitzats mitjançant una validació creuada de 10 iteracions i la parametrització és de 32 paràmetres. El corpus utilitzat és d'expressió vocal estimulada recollint locucions de 40 voluntaris llegint textos emocionalment acolorits i procedents de retalls de premsa per representar 5 emocions bàsiques: enuig, alegria, por, tristesa i estat neutre, amb un total de 197 frases. La taxa de classificació oscil·la entre el 21% i el 55%.

Breazeal (2002) empra un classificador de diverses etapes basat en *Gaussian-Mixture Models* (GMM). El corpus són locucions gravades per dues locutores representant 5 estats afectius relacionats amb la interacció amb un robot (aprovació, demanda d'atenció, prohibició, confort i estat neutre), amb un total de 726 locucions (145 per classe, aproximadament) de les quals s'extreuen 12 paràmetres prosòdics. L'experiment es planteja mitjançant una validació creuada de 100 iteracions encara que per a l'avaluació final del sistema es crea un nou conjunt de dades de 371 locucions que s'empra únicament com a conjunt de prova després de l'entrenament del model de reconeixement. La taxa de reconeixement ascendeix a un 81,9%.

Oudeyer (2003) realitza un experiment pioner, a gran escala, de reconeixement automàtic emprant un corpus gravat per 6 actors professionals pronunciant frases d'ús diari, tals com *"Hola"* i *"Com estàs?"* (el que suposa un plantejament més natural que el del treball presentat per McGilloway *et al.* (2000)), en 4 estats emocionals diferents: alegria, tristesa, enuig i estat neutre. En total es consideren 200 frases per locutor i per emoció, la qual cosa resulta en un corpus de 4.800 frases que, parametritzades acústicament, creen un conjunt de dades de 200 paràmetres per frase, sent un dels primers estudis a comptar amb un corpus de tals dimensions i que empra, posteriorment, una àmplia diversitat de tècniques de mineria de dades per a la seva anàlisi mitjançant l'eina WEKA (Witten i Frank, 2005)[22]. L'experiment es configura mitjançant una validació creuada de 10 iteracions i analitza un total de 19 algorismes d'aprenentatge. Els resultats, considerant la totalitat dels paràmetres, oscil·len entre un 81% i un 95,7% de taxa de reconeixement. Al mateix temps, l'estudi proposa una reducció de paràmetres mitjançant un algorisme genètic com el descrit en la secció 4.3.2. Amb el conjunt de dades reduït a 15 paràmetres els resultats són similars si bé el conjunt de dades emprat és més de 10 vegades més petit que l'original. Així mateix, el treball planteja un senzill algorisme que permet l'entrenament del sistema de forma interactiva i incremental implantant-ho en un robot de tipus Aibo, de Sony, o Papero, de NEC. Gràcies a aquest algorisme, el robot està contínuament a l'es-

[22]WEKA és una llibreria que recull implementacions en Java d'un gran nombre d'algorismes d'aprenentatge desenvolupada per la Universitat de Waikato, Nova Zelanda. El seu nom és l'acrònim de *Waikato Environment for Knowledge Analysis*.

colta de les indicacions d'un usuari les quals classifica de forma automàtica i mostra el resultat d'aquesta classificació mitjançant un sistema de LEDs. L'usuari ha d'assenyalar mitjançant indicacions de veu si la classificació ha estat realitzada amb èxit o no, la qual cosa proporciona una realimentación al robot que realitza un nou entrenament amb la nova locució.

Luengo *et al.* (2005) proposen un estudi basat en un corpus en basc gravat per una actriu de doblatge representant 6 emocions bàsiques: enuig, por, sorpresa, fàstic, alegria i tristesa. Una avaluació subjectiva prèvia mostra una taxa de reconeixement superior al 70% per a les diferents emocions excepte la de fàstic, que també és difícil de reconèixer en altres idiomes (Rodríguez *et al.*, 1999; Burkhardt i Sendlmeier, 2000). Del corpus s'extreuen 86 paràmetres prosòdics i de VoQ. L'experiment consisteix en tres sistemes de classificació, dos d'ells basats en GMM i un altre basat en SVM. Per avaluar aquests models s'empra una validació creuada de 5 models deixant un fora. La taxa de reconeixement aconsegueix el 98,4%, sent del 92,3% estudiant únicament paràmetres prosòdics.

Barra *et al.* (2007) realitzen un estudi sobre dos corpus actuats, un en espanyol: el corpus *Spanish Emotional Speech* (SES) (Montero *et al.*, 1998b) consistent en 30 paraules, 15 frases curtes i 4 paràgrafs, i un altre en alemany: la *Berlin Database of Emotional Speech* (EMODB) (Burkhardt *et al.*, 2005) consistent en 800 frases curtes. El primer representa 3 emocions bàsiques (tristesa, alegria i enuig) i el segon 7. La parametrització de tots dos corpus consisteix en l'extracció de coeficients MFCC. El model de classificació es basa en GMM amb un baix nombre de distribucions. El corpus SES obté una taxa de reconeixement del 89,98% (en un test subjectiu aquesta taxa és del 90,20%), emprant part del corpus per a entrenament i una altra part per a prova, mentre que el corpus EMODB s'avalua mitjançant una validació creuada independent del locutor obtenint una taxa de reconeixement (incorporant tècniques de normalització de les dades) del 51,33% (en un test subjectiu aquesta taxa és del 86,08%).

Hassan i Damper (2009), analitzant el corpus EMODB, obtenen una taxa de reconeixement de fins al 90% considerant 4 emocions bàsiques de les 7 de les quals consta el corpus i de fins al 70,8% considerant les 7 emocions del corpus complet, emprant una validació creuada de 10 iteracions, un algorisme de tipus KNN i la selecció automàtica prèvia dels paràmetres acústics més rellevants mitjançant un algorisme de cerca voraç incremental. Emprant diferents configuracions i corpus, altres estudis obtenen resultats elevats com (Shami i Verhelst, 2007), amb una taxa del 80,7% sobre el corpus EMODB, i (Shaukat i Chen, 2008), amb una taxa del 89,7% sobre el corpus *Serbian emotional speech database* (Jovičić *et al.*, 2004).

Com es pot observar, la disparitat dels corpus emprats (diverses grandàries i diferents emocions considerades), la varietat de parametritzacions, els diferents mètodes de classificació i els diferents mètodes d'avaluació dels mateixos, fan complicada la comparativa entre els resultats obtinguts. Aquest fet també es veurà reflectit en els estudis relacionats amb emocions espontànies.

4.4.2 Reconeixement emocional d'emocions espontànies

Tal com indiquen Zeng *et al.* (2009), cada vegada és més freqüent realitzar estudis basats en enregistraments naturals de veu. No obstant això, en les dades recopilades d'una interacció natural els estats afectius són, sovint, més subtils i les emocions plenes rarament es manifesten. Per aquest motiu, és habitual que els estudis se centrin en el reconeixement d'estats afectius més amplis com, per exemple, positius i negatius (Litman i Forbes-Riley, 2004; Lee i Narayanan, 2005; Neiberg *et al.*, 2006), per simplificar el problema. Per exemple, Ang *et al.* (2002) recullen 6 estats afectius però finalment només treballen amb 2: negatiu i altres. A continuació es detallen alguns treballs emmarcats en aquesta línia de recerca dels quals es desprèn que el reconeixement d'emocions espontànies és més complex donada la pròpia naturalesa de les dades i, per tant, les taxes de reconeixement són bastant inferiors a les d'estudis d'emocions actuades, fins i tot en els treballs més recents. Cal assenyalar que, excepte en els casos assenyalats, les taxes de reconeixement es corresponen amb la mitjana ponderada (WAR).

Slaney i McRoberts (1998) duen a terme un dels primers estudis que analitza emocions no actuades en un context natural, si no el primer a fer-ho (Batliner *et al.*, 2011). En aquest treball s'estudia un corpus de veu format per enregistraments de 12 pares (6 mares i 6 pares) dirigint-se als seus fills d'entre 10 i 18 mesos d'edat en un entorn lúdic. Les locucions s'agrupen en 3 categories: aprovació (212 locucions), atenció (149 locucions) i prohibició (148 locucions). Així mateix, les locucions s'avaluen subjectivament per determinar el grau de pertinença a cada categoria. Les locucions s'analitzen íntegrament i fragmentant-les en tres terços, extraient descriptors de baix nivell relatius a F0, informació cepstral i energia. Els funcionals calculats varien tant en nombre com en tipus. La parametrització final és de només 8 paràmetres. L'algorisme de classificació es basa en GMM avaluat mitjançant una validació creuada de 100 iteracions i s'empra un algorisme de cerca voraç incremental per determinar el nombre òptim de paràmetres. La taxa de reconeixement se situa al voltant del 65%, sent millor la identificació de les categories en el cas de les dones que la dels homes.

Batliner *et al.* (2000) realitzen un estudi a partir de 20 diàlegs recopilats a través d'un sistema de *Mag d'Oz* simulant un sistema de funcionament erroni per estimular empipament en els usuaris. Les categories en les quals s'agrupen les locucions són: emocional i no emocional. La parametrització recull 27 paràmetres acústics i s'inclou informació lingüística a l'efecte de comparació. S'estudien tres tipus algorismes de classificació: anàlisi de discriminant lineal, arbres de regressió i un perceptró multicapa. Els dos primers algorismes s'avaluen mitjançant validació creuada i el tercer mitjançant dos subconjunts diferents per a entrenament i prova. La taxa de classificació oscil·la entre el 63% i el 71%.

El treball de Lee *et al.* (2001) se centra en el reconeixement afectiu automàtic basat en l'anàlisi acústica d'un corpus que recull locucions d'usuaris reals d'un centre de trucades. El nombre total de locucions ascendeix a 7.200. Les categories en les quals s'agrupen són negatives (incloent enuig i frustració) i no negatives (incloent felicitat i delit,

a pesar que la majoria són neutres segons una avaluació subjectiva). La parametrització recull paràmetres relatius a F0 i energia, reduïts posteriorment mitjançant una anàlisi de components principals i un algorisme de cerca voraç incremental. La classificació analitza classificadors de discriminant lineal i de tipus KNN avaluats mitjançant un procés de validació creuada deixant un fora. En el millor dels casos, l'error de classificació se situa entre el 20% i el 25%, sent millor el reconeixement en el cas de les dones que dels homes. Posteriorment, Lee i Narayanan (2005) presenten un dels primers treballs en els quals no només s'estudia el contingut acústic de les locucions d'un corpus (derivat del corpus de l'estudi anterior) sinó que combina 3 tipus d'informació: acústica, lèxica i referent al discurs. El conjunt de paràmetres acústics s'optimitza mitjançant l'anàlisi de diversos subconjunts i l'anàlisi de components principals, i procedeix de la informació de 21 paràmetres relatius a F0, energia, durada i freqüències dels formants. La informació lèxica s'extreu a partir del concepte de prominència emocional (vegeu la secció 3.3.2.1). La informació de discurs procedeix de la categorització del tipus de resposta dels usuaris (per exemple, reformulació o repetició, entre altres). La fusió es realitza a nivell de classificadors, després de classificar independentment cada conjunt de paràmetres, mitjançant una senzilla mitjana. L'estudi torna a analitzar classificadors de discriminant lineal i de tipus KNN. En aquesta ocasió, no obstant això, els algorismes s'avaluen mitjançant una validació creuada de 10 iteracions, distingint entre locutors masculins i femenins. L'estudi marca l'error de classificació que, en el millor dels casos, és del 7,95% per al cas les locutores i del 10,55% per al cas dels locutors.

Litman i Forbes-Riley (2003) realitzen un estudi centrat en sistemes de tutoria al món docent. S'analitza un corpus de 202 locucions de 2 homes i 2 dones interactuant, a través d'una interfície web, amb un tutor humà. L'anàlisi es duu a terme a nivell acústic (33 paràmetres) i s'incorporen 3 identificadors de context a cada locució (incloent un identificador de l'usuari i el seu sexe). Es consideren 3 categories emocionals: positiu (15 locucions), neutre (145 locucions) i negatiu (42 locucions). S'estudien diferents agrupacions de paràmetres amb 4 algorismes de classificació: un arbre de decisió J4.8, un de tipus KNN, un de màxima versemblança (ZeroR) com a referència i un algorisme de *boosting* amb l'arbre J4.8 com a algorisme feble. L'avaluació es realitza mitjançant una validació creuada de 10 iteracions. El millor resultat obtingut considera els 3 identificadors de context a més dels paràmetres acústics i ofereix una taxa de reconeixement del 80,53%.

Truong i Raaijmakers (2008) empren un algorisme de *boosting* per crear dos classificadors, un d'acústic i un altre lingüístic, avaluant la tasca de classificació en les dimensions emocionals d'activació i valència, i també a nivell categòric. El corpus recull locucions de 17 homes i 11 dones jugant a un videojoc de tipus *first-person shooter*. Els paràmetres acústics fan referència a la F0 i la intensitat mentre que els lingüístics es refereixen a n-grames a nivell de paraula i a la velocitat de la parla. Distingeixen dues categories afectives: positiu i negatiu, i treballen a nivell individual i mitjançant la fusió de tots dos paràmetres a nivell de paràmetres. En la dimensió d'activació els classificadors acústics funcionen millor, mentre que en la d'avaluació obtenen millors resultats els lingüístics. La fusió a nivell de paràmetres millora lleugerament els resultats, encara que no en tots els casos.

A nivell categòric la taxa de classificació està entre el 60% i el 70%. A nivell dimensional se situa entre el 40% i el 55%. La classificació lingüística realitzada a partir de l'etiquetatge manual funciona millor que la realitzada de forma automàtica a partir d'un sistema *Automatic Speech Recogniser* (ASR), encara que la diferència no és notable a nivell dimensional. Posteriorment, Osherenko *et al.* (2009) també realitzen un estudi de fusió acústica i lingüística, en aquesta ocasió a nivell de paràmetres i també a nivell de decisió sense trobar que un esquema sigui millor que l'altre ni aconseguir millorar, mitjançant la fusió, els resultats individuals de classificació.

Donada la mateixa disparitat que ja es produïa en el cas dels estudis d'emocions actuades quant als corpus, les emocions considerades (tant en nombre com en tipus), els mètodes d'avaluació i la, freqüentment, escassa documentació relativa als mitjans emprats per dur a terme reduccions de dimensionalitat, Schuller *et al.* (2009), a través de l'*Emotion Challenge* dins de la conferència *INTERSPEECH* de l'any 2009, proposen un entorn de treball amb un corpus específic i un mètode d'avaluació concret perquè els treballs que ho segueixin puguin ser comparats, i ho fan en forma de "desafiament". L'*Emotion Challenge 2009* planteja tres desafiaments diferents: un desafiament obert, en el qual es poden emprar qualssevol paràmetres i algorismes; un de classificació, emprant uns paràmetres predeterminats; i un de parametrització, en el qual els autors han de proporcionar els paràmetres que considerin més rellevants. El corpus proposat és el corpus FAU Aibo (vegeu la secció 3.4.3 per a una descripció del corpus i els detalls de la seva parametrització). Compta amb més locutors que els estudis previs, així com més frases. Considera 2 categories emocionals però també un etiquetatge de 5. La mesura de rendiment dels classificadors s'estableix com la mitjana no ponderada de reconeixement (UAR) donada la distribució no uniforme de les categories del corpus. Així mateix, estableix una divisió de les dades del corpus definint un subconjunt específic per a entrenament i un altre per a prova, forçant la total independència entre les dades de tots dos conjunts a nivell, fins i tot, de locutors, aspecte no contemplat en els estudis anteriorment citats. A continuació es detallen alguns dels treballs emmarcats en aquest context els resultats del qual es recullen en el sumari de l'*Emotion Challenge 2009* publicat en (Schuller *et al.*, 2011).

Schuller *et al.* (2009) presenten uns resultats de referència al mateix temps que defineixen l'*Emotion Challenge 2009* i especifiquen les bases de participació. Emprant un classificador SVM i el processament previ dels paràmetres originals del corpus, aquests resultats ofereixen un valor de UAR del 67,7% per al problema de 2 categories i 38,2% per al problema de 5 categories[23]. Noti's com es redueix la taxa de classificació en el cas de treballar amb 5 categories respecte a la resta d'estudis. La disminució de la taxa de reconeixement en estudis independents del locutor amb un nombre de classes relativament elevat és un efecte observable sovint en els treballs de reconeixement afectiu (Kwon *et al.*, 2003; Austermann *et al.*, 2005).

El treball de Lee *et al.* (2009) es basa en la teoria d'avaluació (Lazarus, 2001) que determina que el reconeixement emocional es fonamenta en processos conscients i incons-

[23] Aquest últim resultat s'emprarà com a valor de referència per als experiments detallats en el capítol 6.

cients de diverses etapes en la qual el subjecte, en cadascuna d'elles, avalua la situació, reacciona i la torna a avaluar. Per això proposa una estructura jeràrquica per al problema de 5 categories afectives en la qual la tasca de classificació més senzilla se situa en la part superior mentre que la decisió entre les classes més ambigües es deixa per al final. Cada classificador és binari i empra regressió logística bayesiana. La taxa de classificació és del 41,6%.

Vogt i André (2009) analitzen l'entorn de reconeixement emocional a temps real EmoVoice (Vogt *et al.*, 2008) amb el corpus FAU Aibo. Per al problema de 2 categories no superen el resultat de referència i la taxa de reconeixement arriba al 66,4% mentre que per al problema de 5 categories la taxa és del 39,4%. Barra *et al.* (2009) milloren la taxa de reconeixement en el cas del problema de 2 categories però sense superar tampoc el resultat de referència, arribant al 67,1%, si bé per al problema de 5 categories ho iguala amb un resultat del 38,2%, emprant l'anàlisi de paràmetres MFCC, el logaritme de l'energia i la F0 i la primera i segona derivades amb una xarxa bayesiana dinàmica. Luengo *et al.* (2009) també empren paràmetres cepstrals (en aquest cas *Log-Frequency Power Coefficients* (LFPC)), considerant la primera i segona derivades, que són classificats mitjançant GMM. Un nou conjunt de 10 paràmetres prosòdics es classifiquen mitjançant SVM. Les taxes de reconeixement són del 67,2% per al problema de 2 categories i 41,4% pel de 5.

Polzehl *et al.* (2009) realitza una anàlisi centrada en el problema de 2 categories i classifica de forma independent un conjunt de paràmetres acústics i un altre de paràmetres lingüístics fusionant els resultats posteriorment. La taxa de reconeixement és molt propera als resultats de referència i ascendeix al 67,6%.

Els millors resultats de l'*Emotion Challenge 2009* van ser els obtinguts per Dumouchel *et al.* (2009) i Kockmann *et al.* (2009). Dumouchel *et al.* (2009) obtenen una taxa del 70,29% en el problema de 2 categories mitjançant l'ús de tres classificadors GMM fusionats a nivell de decisió, emprant paràmetres MFCC i prosòdics. La taxa de Kockmann *et al.* (2009) per al problema de 5 categories és del 41,65% mitjançant l'anàlisi de paràmetres MFCC i classificació GMM.

Schuller *et al.* (2011) realitzen una fusió a partir dels resultats dels millors esquemes de classificació presentats a l'*Emotion Challenge 2009* mitjançant un mètode de vot majoritari. Per al problema de 2 categories la millor taxa de classificació s'obté fusionant els 5 millors classificadors (Dumouchel *et al.*, 2009; Vlasenko i Wendemuth, 2009; Kockmann *et al.*, 2009; Bozkurt *et al.*, 2009; Polzehl *et al.*, 2009) obtenint un resultat del 71,16% mentre que per al problema de 5 categories el millor resultat s'obté mitjançant la fusió dels 7 millors classificadors (Kockmann *et al.*, 2009; Bozkurt *et al.*, 2009; Lee *et al.*, 2009; Vlasenko i Wendemuth, 2009; Luengo *et al.*, 2009; Planet *et al.*, 2009; Dumouchel *et al.*, 2009) ascendint al 44,01%. En tots dos casos, el millor esquema individual de classificació sempre es veu millorat en fusionar-ho, almenys, amb altres dos classificadors.

4.5 Altres aproximacions al reconeixement emocional

Existeixen altres aproximacions al reconeixement afectiu automàtic que estudien modalitats diferents a la parla o bé realitzen un processament multimodal en el que la parla és només una part del procés de reconeixement. No obstant això, no és l'objectiu d'aquesta tesi l'aprofundir en aquests estudis atès que el present treball se centra en el reconeixement afectiu automàtic a partir del senyal de veu.

Existeixen estudis basats en l'expressió facial dels estats afectius com per exemple els treballs de Bartlett *et al.* (2003), Cohn *et al.* (2004), Ioannou *et al.* (2005) i Valstar *et al.* (2007), centrats en l'expressió facial d'emocions espontànies. Les taxes de reconeixement per a aquesta modalitat se solen situar al voltant del 80% (Zeng *et al.*, 2009), si bé existeixen diferències notables entre els treballs que analitzen imatges estàtiques i els que estudien seqüències de vídeo. També existeixen aproximacions multimodals com la de Busso *et al.* (2004), que fusionen a nivell de paràmetres i de decisió les modalitats acústica i d'expressió facial, la de Fragopanagos i Taylor (2005), que combinen l'anàlisi facial, prosòdic i lèxic o la de Zeng *et al.* (2007) a nivell facial i acústic.

Altres estudis analitzen els senyals fisiològics captats mitjançant mètodes invasius i no invasius per detectar l'emoció dels usuaris, com Picard *et al.* (2001), Nasoz *et al.* (2004), Li i Chen (2006), Lin *et al.* (2010) i Gouizi *et al.* (2011). Les taxes de reconeixement a partir de senyals fisiològics varien d'un treball a un altre però habitualment se situen al voltant del 70%. El treball de Kim i André (2006) no es basa exclusivament en els senyals fisiològics sinó que inclou, a més, la seva fusió amb l'anàlisi acústica del discurs mitjançant un esquema a nivell de paràmetres i un altre híbrid, en el qual la sortida anterior serveix d'entrada a un esquema de fusió a nivell de decisió.

L'anàlisi dels resultats d'algunes de les aproximacions multimodals mostra valors elevats de taxes de reconeixement, del 75% fins al 98% (Zeng *et al.*, 2009). No obstant això, les aproximacions al reconeixement afectiu automàtic multimodal no són alienes als inconvenients citats per als estudis de reconeixement basats en el senyal de veu, tals com la varietat de corpus, emocions i mesures d'avaluació que dificulten la comparativa entre els diversos treballs.

4.6 Resum

En aquest capítol s'han recopilat alguns dels estudis relacionats amb el reconeixement afectiu automàtic, sobretot els relacionats amb el senyal de veu atès que és l'objecte de la present tesi. Després de recórrer els antecedents a partir dels primers treballs realitzats en aquest àmbit s'han tractat les diferents aplicacions del reconeixement afectiu per posar de manifest l'interès de l'objecte d'estudi. A continuació, s'han mostrat quines són les tendències més representatives en els treballs de reconeixement a partir del senyal de veu més recents a través de l'anàlisi d'alguns dels estudis més rellevants. Així, es com-

prova l'auge dels treballs que realitzen processos de parametrització cada vegada més exhaustius dels corpus emprant, posteriorment, diverses tècniques d'aprenentatge artificial per a la seva anàlisi. Així mateix, s'han posat de manifest les diferències existents entre els treballs basats en emocions actuades i emocions espontànies, així com la dificultat d'estudiar un conjunt relativament gran d'emocions concretes enfront d'analitzar un conjunt més petit que consideri etiquetes emocionals més àmplies. També s'aprecia l'evolució que, al llarg del temps, ha experimentat el procés de parametrització, incorporant als paràmetres prosòdics inicials, paràmetres relatius a la VoQ o al contingut lèxic de les locucions dels corpus.

L'anàlisi d'aquests treballs previs posen de manifest la dificultat en la comparativa entre els mateixos, donada la diversitat de corpus, categories emocionals, processos de parametrització i mètodes d'avaluació emprats per tots ells.

També s'ha fet un breu incís en les anàlisis d'altres modalitats i s'han introduït els tres tipus principals de fusió multimodal, útils per intentar millorar les taxes de reconeixement a partir de l'anàlisi de dades heterogènies, procedents de diversos sensors, o bé per fusionar les dades extretes de les anàlisis d'una mateixa modalitat però realitzades des de diferents perspectives.

Capítol 5

Primeres aproximacions al reconeixement afectiu en corpus actuats

Aquest capítol se centra en el reconeixement afectiu automàtic realitzat sobre corpus gravats per actors, incloent tant els corpus d'expressió vocal actuada com els de expressió vocal estimulada. En tots els casos es fa referència a corpus que han estat gravats en condicions ideals i amb el màxim control tant de les condicions d'enregistrament com de contingut.

Els continguts d'aquest capítol no són el nucli de la tesi en si i es refereixen als experiments realitzats com a pas previ als que es realitzaran al voltant de la línia principal de recerca, els quals s'aborden al capítol següent. El seu objectiu és establir una base d'experimentació i poder obtenir uns resultats que serveixin per comparar-los posteriorment, observant les diferències entre tots dos escenaris.

El capítol es divideix en dues seccions. Ambdues fan referència al reconeixement afectiu automàtic en corpus de veu emocionada gravada per actors però difereixen en l'objectiu final. La primera secció planteja un reconeixement afectiu bàsic la finalitat del qual és maximitzar el nombre de coincidències entre l'estil afectiu assignat a cada locució i l'estil determinat pel sistema automàtic. No obstant això, aquest procediment dóna per fet que l'estil afectiu assignat a les locucions és el correcte, sense considerar que tal vegada l'actor no va encertar a proporcionar a la locució l'estil pretès. Determinar quines locucions són correctes o incorrectes, des del punt de vista de correspondre's, o no, amb l'estil pretès, suposa un procés de validació posterior a l'enregistrament realitzat per un conjunt d'avaluadors humans, és a dir, subjectius. La segona secció d'aqueest capítol exposa la manera d'automatitzar aquest procés realitzant un reconeixement afectiu automàtic maximitzant el nombre de coincidències per a aquelles locucions que també serien correctament identificades de forma subjectiva. D'igual manera, les locucions que de for-

ma subjectiva serien rebutjades, el sistema automàtic hauria de poder identificar-les com a incorrectes rebutjant-les també atès que no s'ajusten a l'estil pretès.

5.1 Reconeixement afectiu bàsic

Aquesta secció recull una sèrie d'experiments denominats bàsics ja que plantegen un reconeixement senzill i directe. L'entorn de treball se centra al voltant de corpus actuats, convenientment etiquetats i abastant emocions clarament definides. L'objectiu dels experiments és maximitzar la WAR, és a dir, classificar correctament el major nombre de casos del total existent ja que els corpus s'han dissenyat de forma equilibrada. Així doncs, es considera reconeixement afectiu bàsic atès que segueix les pautes de la majoria dels treballs previs realitzats dins del marc del reconeixement emocional, tals com els citats en la secció 4.4.1. Aquests experiments previs i l'anàlisi dels seus resultats marcaran la base dels treballs posteriors detallats en les seccions següents.

L'estructura genèrica dels experiments de reconeixement afectiu bàsic es pot observar en la figura 5.1. Els experiments es divideixen en dues parts independents. La primera d'elles consisteix a analitzar el corpus de forma subjectiva, recopilant les avaluacions realitzades per un grup de persones en escoltar els diferents elements del corpus. La segona part consisteix a realitzar una classificació automàtica d'aquests mateixos elements emprant tècniques d'aprenentatge artificial. Finalment, es comparen els resultats d'ambdues parts. Aquesta estructura bàsica és la que se segueix en les següents seccions.

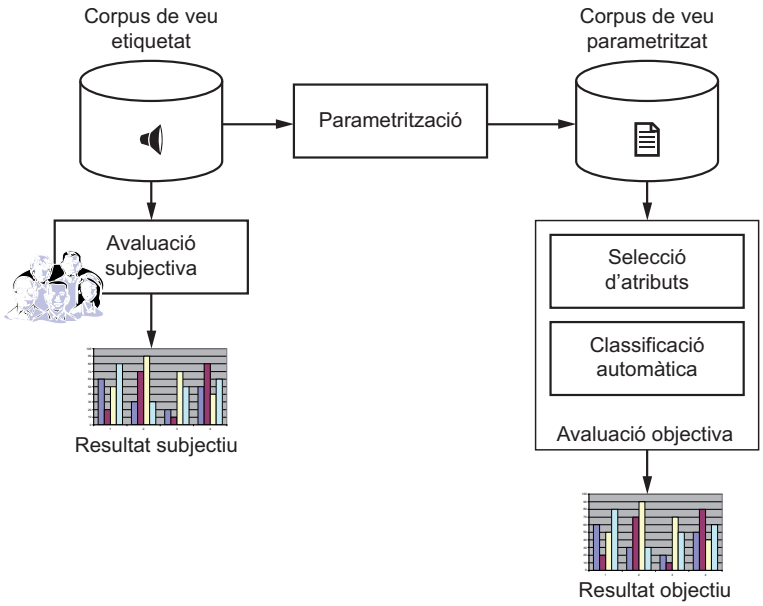


Figura 5.1: Esquema d'un experiment genèric de reconeixement afectiu bàsic.

A continuació es presenten dos estudis de reconeixement afectiu bàsic, tal com es recullen en (Planet *et al.*, 2006) i (Iriondo *et al.*, 2007c). El primer d'ells és un estudi

preliminar basat en el corpus de veu actuada ESDLA. El segon presenta una anàlisi més detallada, basada en el corpus veu estimulada BDP-UAB, gravat en millors condicions tècniques que l'anterior i comptant amb un major volum de dades.

5.1.1 Corpus de veu actuada

Per realitzar l'experiment de reconeixement afectiu bàsic sobre un corpus de veu actuada s'empra el corpus ESDLA, estudiant únicament els paràmetres analitzats a nivell segmental mitjançant l'aplicació de finestres de durada constant, és a dir, 183 paràmetres relatius a F0, energia i silencis, segons es detalla en la secció 3.4.1.2. En (Planet *et al.*, 2006) es recullen els detalls de l'experiment aquí explicat.

5.1.1.1 Metodologia

En l'estudi es tenen en consideració conjunts de dades que inclouen la informació referent als quatre estats emocionals del corpus i també conjunts dels que s'elimina tota referència a l'estat neutre, per comprovar l'impacte d'un estat sense contingut afectiu en el sistema automàtic de classificació. Referent al nombre de paràmetres analitzats, els algorismes d'aprenentatge s'entrenen tant amb la totalitat dels paràmetres com amb subconjunts dels mateixos per intentar fitar el problema i evitar els efectes del sobreentrenament. Aquests subconjunts es creen sobre la base del coneixement previ dels paràmetres que són considerats més rellevants per la bibliografia en general i també seleccionant manualment els que se suposa, a priori, que poden ser els més rellevants a nivell subjectiu. La figura 5.2 mostra la definició d'aquests subconjunts. Així doncs, *data2* es crea després d'eliminar la primera derivada del conjunt complet *data1*. *data3* no considera les versions filtrades del conjunt complet, i *data4* tampoc ho fa a partir del subconjunt *data2*. *data5* incorpora, a diferència de *data4*, els paràmetres relatius a l'energia d'aquestes versions filtrades. *data6* redueix el nombre de funcionals a 5 (mitjana, desviació estàndard, asimetria, curtosis i mitjana) excepte pels paràmetres relatius als silencis que incorporen, a més, la durada total dels mateixos. D'altra banda, es crea de forma automàtica un altre subconjunt (*data7*) format pels paràmetres que un algorisme genètic com el descrit en la secció 4.3.2 identifica com els més rellevants. Mitjançant aquest procés automàtic de selecció d'atributs, per al conjunt de tots els estats emocionals el nombre de paràmetres es redueix a 37 (reducció del 79,78% del total), mentre que per al conjunt que omet l'estat neutre s'obtenen 19 paràmetres (reducció del 89,62% del total).

L'experiment avalua 8 algorismes de classificació i 2 metaclassificadors. Els algorismes de classificació, segons implementacions de la llibreria WEKA, són: l'arbre de decisió basat en entropia J4.8 (Witten i Frank, 2005), l'algorisme de creació de llistes de decisió PART (Frank i Witten, 1998), dos algorismes KNN (Aha *et al.*, 1991), basats en distància euclidiana i seleccionant un únic veí de forma predeterminada (IB1) i una versió en la

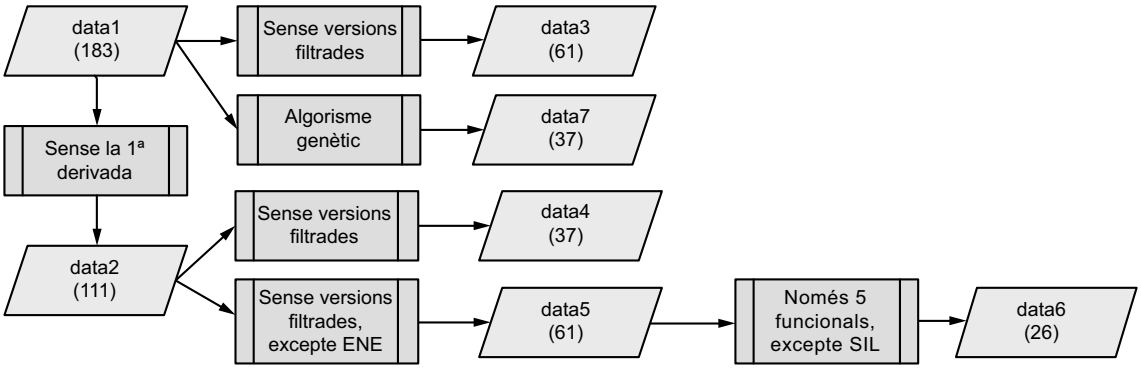


Figura 5.2: Definició dels conjunts de dades creats per a l'experiment de reconeixement afectiu bàsic sobre el corpus ESDLA. Entre parèntesis s'indica el nombre de paràmetres que conté cada conjunt de dades. ENE fa referència als paràmetres d'energia i SIL als de silenci.

qual el nombre de veïns propers (k) es determina automàticament de forma òptima[1] (IBk), l'algorisme Naïve-Bayes (John i Langley, 1995), amb discretització supervisada basada en l'algorisme de Fayyad i Irani (1993), un algorisme de taules de decisió (DT) (Kohavi, 1995), un algorisme SVM de kernel polinòmic de segon grau, amb un entrenament mitjançant l'algorisme d'optimització seqüencial mínima (Platt, 1998) i utilitzant discriminació per parells (Hastie i Tibshirani, 1998) per poder realitzar un experiment de classificació amb un nombre de classes superior a dues. L'algorisme de màxima versemblança ZeroR[2] s'escull com a línia de base per a aquest experiment tal com també fa Litman i Forbes-Riley (2003) en un experiment de similars característiques[3]. Els dos metaclassificadors són de tipus *boosting* (algorisme AdaBoost (Freund i Schapire, 1996)) i *bagging* (Breiman, 1996) en combinació amb els algorismes J4.8, PART i DT (AB/J4.8, AB/PART, AB/DT i Bag/J4.8, respectivament). En tots els casos es mesura la WAR[4] avaluada mitjançant una validació creuada de 10 iteracions.

5.1.1.2 Resultats

La figura 5.3 mostra els valors de WAR aconseguits per cada algorisme per a cadascun dels conjunts de dades avaluats considerant el corpus complet, amb les 4 categories emocionals. L'algorisme de referència ZeroR, que el seu resultat no s'inclou en la gràfica per simplificar la visualització del resultat, és del 25% atès que el corpus està format per 4 classes amb un nombre idèntic de casos. S'observa que el millor resultat (89,5%) s'obté

[1] El procés automàtic consisteix en un procés iteratiu realitzat mitjançant una validació creuada deixant un fora.

[2] L'algorisme ZeroR prediu el valor mitjà si l'etiqueta de classificació és numèrica, o la moda (el valor més freqüent) si l'etiqueta de classificació és, com en aquest cas, categòrica.

[3] No obstant això ZeroR és un algorisme molt senzill i, malgrat ser emprat sovint com a referència en l'experimentació amb WEKA, tan sols s'emprarà en els experiments d'aquest capítol pel seu caràcter preliminar.

[4] Considerant que el corpus ESDLA és un corpus equilibrat, amb el mateix nombre de locucions per a cada emoció, els valors de les mesures WAR i UAR són iguals.

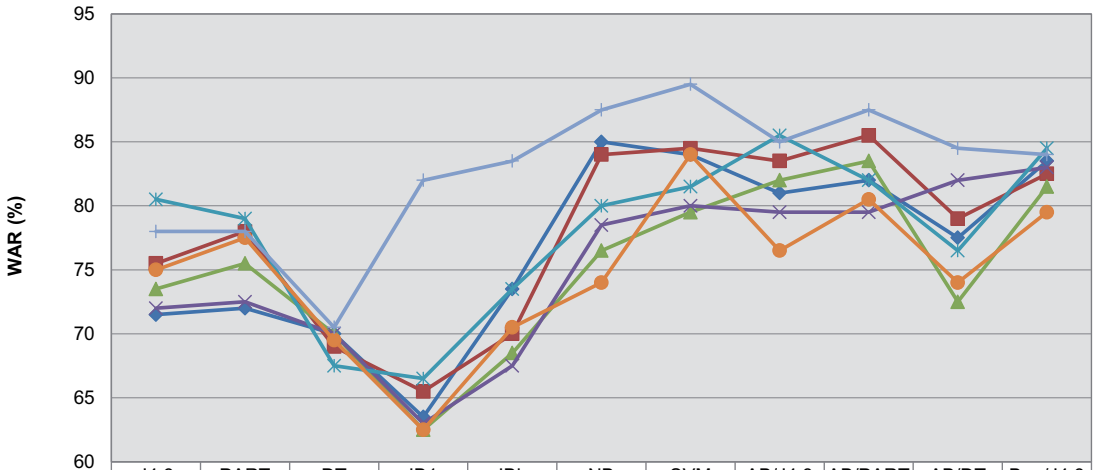
mitjançant l'algorisme SVM en combinació amb el conjunt de dades *data7*, obtenint una millora absoluta del 64,5% respecte a l'algorisme de referència. Aquest resultat és també superior en un 0,82% a la taxa de classificació obtinguda en el test subjectiu. Al mateix temps, supera en un 2% els resultats obtinguts pels dos següents millors classificadors que són l'esquema de *boosting* en combinació amb l'algorisme PART i l'algorisme Naïve-Bayes amb discretització supervisada prèvia i considerant el mateix conjunt de dades.

En la figura 5.3 es pot observar també com milloren els algorismes J4.8, PART i DT quan són vinculats a un esquema de *boosting* i com es produeix també aquesta millora en l'esquema *bagging* observant-se, a més, major homogeneïtat entre els resultats dels diferents conjunts de dades. Així mateix, la reducció de dades millora els resultats de classificació. Això pot observar-se en comprovar que el conjunt de dades *data7* obté unes taxes de classificació superiors en la majoria dels casos. Si es pren com a referència el conjunt *data1*, format per la totalitat de paràmetres, el conjunt *data7* aconsegueix una millora absoluta del 5,5% (6,55% relativa) en l'algorisme SVM tot i treballar amb el 20,22% del total de paràmetres. Aquesta millora és encara més destacable en el cas de l'algorisme IB1, on aconsegueix un valor del 18,5% de forma absoluta i del 29,13% de forma relativa. Similar és el comportament de l'algorisme IBk, el qual obté un resultat òptim per a un valor de $k = 14$, si bé el seu resultat millora l'obtingut pel IB1. Tots dos algorismes proporcionen uns resultats bons malgrat no ser massa complexos ni computacionalment costosos, la qual cosa també succeeix en el cas de l'algorisme Naïve-Bayes que, malgrat la seva senzillesa, obté una millora absoluta del 62,5% respecte a l'algorisme de referència, situant-se al nivell dels metaclassificadors que suposen un cost computacional més elevat.

En referència als conjunts de dades, en general la no consideració de les versions filtrades per part de *data3* i *data4* empitjora els resultats obtinguts per *data1* i *data2*. El subconjunt *data5* millora els resultats obtinguts per *data4* en incorporar els paràmetres d'energia extrets de les versions filtrades però *data7*, obtingut a través d'un algorisme genètic a partir del conjunt complet, és el conjunt de menor grandària que optimitza les taxes de classificació.

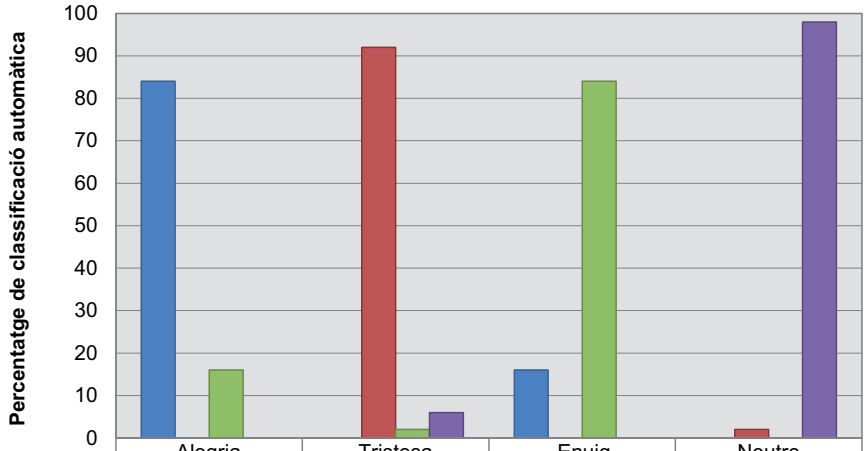
La figura 5.4 mostra la matriu de confusió corresponent a l'algorisme SVM i el conjunt de dades *data7*. De la matriu de confusió es desprèn que alguns dels exemples de la classe alegria es classifiquen com a enuig i viceversa, la classe tristesa és, encara que en poques ocasions, classificada com a neutre mentre que la classe neutre es classifica gairebé perfectament. Aquest fet, que és més palès encara en altres classificadors, contrasta amb els resultats obtinguts en l'avaluació subjectiva (vegeu la figura 3.4).

En (Planet *et al.*, 2006) també es recullen els resultats del mateix experiment quan s'eliminen les referències a l'estat neutre. Efectivament, i tal com es recull en altres estudis de l'estat de la qüestió, la disminució del nombre de classes desemboca en una millora de les taxes de classificació. En aquest cas, però, el millor resultat s'obté mitjançant l'algorisme AB/DT i un conjunt de dades procedent, igual que en el cas anterior, d'un algorisme genètic. La taxa de reconeixement arriba al 90,67% (una millora absoluta del 57,34% respecte a l'algorisme de referència, el resultat del qual és del 33,33% en el cas de



	J4.8	PART	DT	IB1	IBk	NB	SVM	AB/J4.8	AB/PART	AB/DT	Bag/J4.8
data1	71,50	72,00	70,00	63,50	73,50	85,00	84,00	81,00	82,00	77,50	83,50
data2	75,50	78,00	69,00	65,50	70,00	84,00	84,50	83,50	85,50	79,00	82,50
data3	73,50	75,50	70,00	62,50	68,50	76,50	79,50	82,00	83,50	72,50	81,50
data4	72,00	72,50	70,00	63,00	67,50	78,50	80,00	79,50	79,50	82,00	83,00
data5	80,50	79,00	67,50	66,50	73,50	80,00	81,50	85,50	82,00	76,50	84,50
data6	75,00	77,50	69,50	62,50	70,50	74,00	84,00	76,50	80,50	74,00	79,50
data7	78,00	78,00	70,50	82,00	83,50	87,50	89,50	85,00	87,50	84,50	84,00

Figura 5.3: Resultats WAR de l'experiment de reconeixement afectiu bàsic en el corpus ESDLA considerant els 4 estats emocionals. El resultat de l'algorisme de referència ZeroR és 25%.



	Alegria	Tristesa	Enuig	Neutre
Alegria	84,00	0,00	16,00	0,00
Tristesa	0,00	92,00	0,00	2,00
Enuig	16,00	2,00	84,00	0,00
Neutre	0,00	6,00	0,00	98,00

Classificació real

Figura 5.4: Matriu de confusió dels resultats de classificació de l'algorisme SVM i el conjunt de dades *data7* en l'experiment de reconeixement afectiu bàsic amb el corpus ESDLA considerant els 4 estats emocionals.

3 classes). Novament, l'algorisme Naïve-Bayes, amb un cost computacional molt inferior al de l'algorisme AB/DT, obté un bon resultat (el segon millor, en aquest cas) amb una taxa de classificació del 89,33%, que representa una millora absoluta del 56% respecte a l'algorisme de referència.

Les diferències citades entre els resultats de la classificació subjectiva (vegeu la secció 3.4.1.1) i l'automàtica poden atribuir-se a diferències entre els criteris dels avaluadors subjectius i els representats pels paràmetres dels conjunts de dades, segons les observacions realitzades pels propis avaluadors tal com es recull en (Planet *et al.*, 2006). Els participants del test subjectiu coincideixen a afirmar que l'emoció enuig ha estat la més fàcil d'identificar, assenyalant la major dificultat a diferenciar neutre i tristesa en primer lloc, i alegria i neutre, en alguns casos, la qual cosa apareix reflectit en la figura 3.4. Atès que el context no és d'utilitat per prendre la decisió entre una classe o una altra, els participants assenyalen com a aspectes diferenciadors el to de la veu i el volum, així com les seves variacions respectives, i la velocitat de la parla. Alguns participants assenyalen també com a decisius els "cops de veu", terme que podria interpretar-se com a exageracions de les síl·labes accentuades, ja siguin relatives al seu volum o a la seva durada en comparació de la resta d'elles, aspectes no contemplats en la parametrització emprada.

5.1.2 Corpus de veu estimulada

A diferència del treball exposat en la secció anterior, el presentat en (Iriondo *et al.*, 2007c) desenvolupa un plantejament similar però basat en el corpus BDP-UAB que, a diferència del corpus ESDLA, és de veu estimulada. Així mateix, malgrat ser un corpus bastant equilibrat, no ho és en la mateixa mesura que el corpus anterior i entre les categories emocionals s'inclou l'estat sensual, que no es correspon amb una emoció bàsica. El treball recollit en (Iriondo *et al.*, 2007c) només té en consideració els paràmetres prosòdics del corpus, els quals suposen un total de 464 paràmetres per frase.

5.1.2.1 Metodologia

El conjunt de dades original, obtingut a partir d'una segmentació fonètica, es divideix en subconjunts per reduir la seva dimensió. La figura 5.5 mostra les diferents estratègies seguides per aconseguir aquest objectiu. El primer criteri consisteix a ometre la segona derivada (de *data1* a *data2*) per valorar la rellevància d'aquesta funció. D'altra banda, i atès que experiments previs van mostrar un millor rendiment dels paràmetres en format logarítmic, dos conjunts de dades es creen a partir de les versions logarítmiques d'energia i F0 (*data1L* i *data2L*). Cadascun d'aquests subconjunts es divideixen novament en dos considerant tots els fonemes o tan sols les vocals tòniques. A més, es redueix la dimensió dels conjunts de dades inicials amb l'algorisme genètic ja emprat en (Planet *et al.*, 2006) (*data1G* i *data2G*). També es creen dos subconjunts similars als proposats per Navas *et al.* (2006) per observar la rellevància dels paràmetres temporals (*data1N* i *data1NG*).

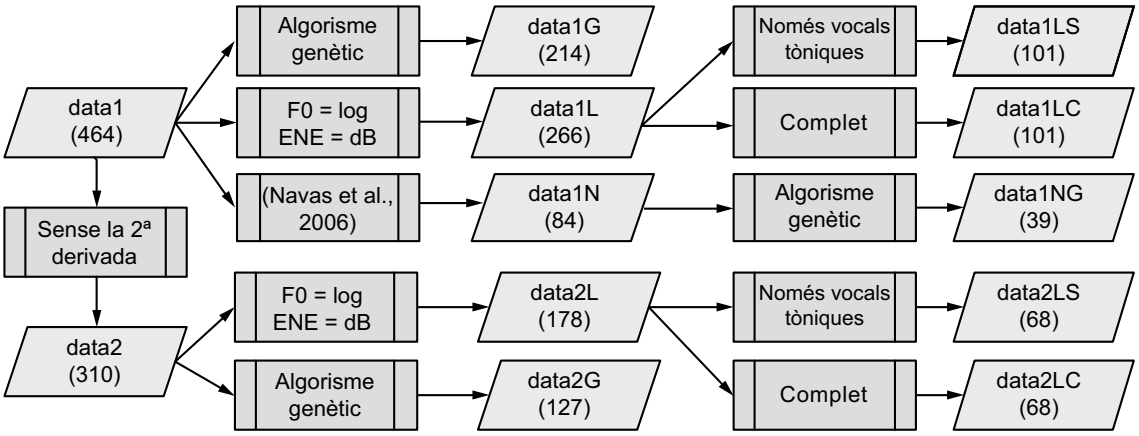


Figura 5.5: Definició dels conjunts de dades creats per a l'experiment de reconeixement afectiu bàsic sobre el corpus BDP-UAB. Entre parèntesis s'indica el nombre de paràmetres que conté cada conjunt de dades. ENE fa referència als paràmetres d'energia. El camp *Complet* fa referència al fet que es consideren tots els segments de la frase, sense excepció.

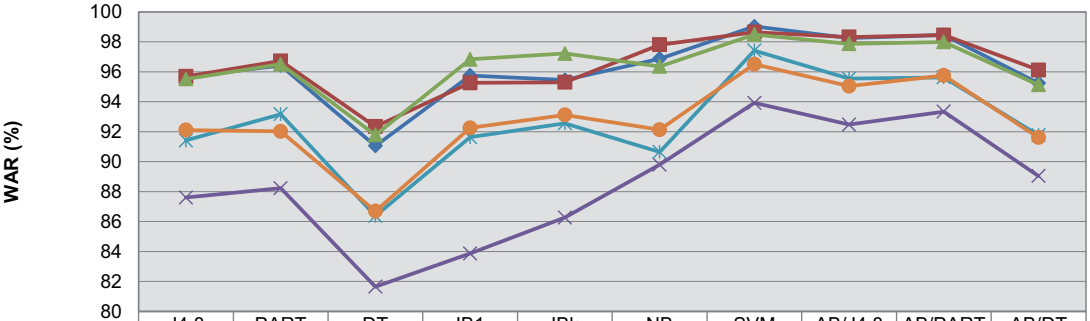
L'experiment avalua els mateixos 8 classificadors bàsics de (Planet *et al.*, 2006) i inclou el mateix metaclassificador de *boosting* en combinació amb els algorismes J4.8, PART i DT. En tots els casos es mesura la WAR avaluada mitjançant una validació creuada de 10 iteracions.

5.1.2.2 Resultats

La figura 5.6 mostra els resultats de l'experiment. L'algorisme de referència ZeroR, que el seu resultat no es mostra en la gràfica per simplicitat, aconsegueix una taxa de reconeixement del 22,60% en assignar a totes les locucions l'etiqueta majoritària "agressiu". Des d'un punt de vista dels algorismes de classificació, SVM obté la millor taxa de classificació tant en valor mitjà (97,34% de mitjana) com de forma puntual en combinació amb el conjunt de dades *data1G* (99,03%) (Iriondo *et al.*, 2007c). Aquesta combinació proporciona una millora absoluta del 76,43% respecte a l'algorisme de referència. A més, i tal com ja es notava en l'experiment de la secció 5.1.1, els metaclassificadors proporcionen millors resultats que els algorismes que empren quan aquests s'estudien forma individual. No obstant això, el cost computacional és més elevat.

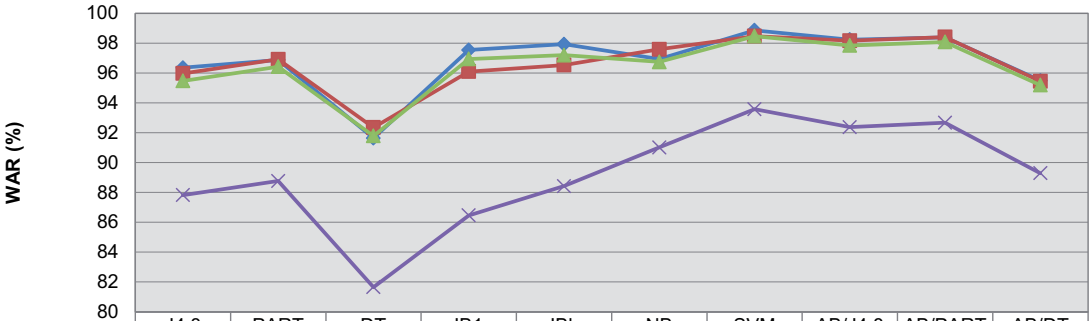
Des d'un punt de vista dels conjunts de dades s'aprecien diferents resultats en funció de l'algorisme. Així, SVM obté el millor resultat amb el conjunt de dades *data1G*. No obstant això, altres algorismes (J4.8, IB1 i IBk) obtenen millors taxes de classificació amb els subconjunts que han estat generats a partir de l'eliminació de la segona derivada i, posteriorment, s'ha realitzat una reducció mitjançant l'algorisme genètic. Finalment, un tercer grup d'algorismes (DT i NB) proporciona millors resultats quan s'elimina la redundància lineal/logarítmica de la F0 i l'energia. D'altra banda, els resultats obtinguts pels conjunts de dades *data1G* i *data1L* són bastant similars, si bé *data1LC* compta amb menys

de la meitat de paràmetres, efecte que també es percep en els conjunts que ometen la segona derivada. Referent als paràmetres temporals, els resultats empitjoren quan aquests s'ometen (conjunts de dades *data1N* i *data1NG*). I, finalment, els resultats empitjoren significativament quan els paràmetres es calculen únicament en les vocals tòniques (conjunts *data1LS* i *data2LS*). En mitjana, el conjunt de dades *data2G* obté el millor resultat, amb un 97,02% de mitjana (Iriondo *et al.*, 2007c).



	J4.8	PART	DT	IB1	IBk	NB	SVM	AB/J4.8	AB/PART	AB/DT
data1G	95,67	96,42	91,05	95,75	95,45	96,87	99,03	98,25	98,45	95,24
data1L	95,69	96,72	92,35	95,26	95,30	97,80	98,64	98,32	98,45	96,12
data1LC	95,52	96,51	91,79	96,83	97,22	96,36	98,49	97,87	97,99	95,13
data1LS	87,60	88,23	81,65	83,87	86,27	89,80	93,92	92,48	93,34	89,05
data1N	91,42	93,17	86,37	91,63	92,56	90,64	97,41	95,54	95,62	91,76
data1NG	92,11	92,02	86,70	92,26	93,12	92,13	96,51	95,04	95,75	91,61

(a)



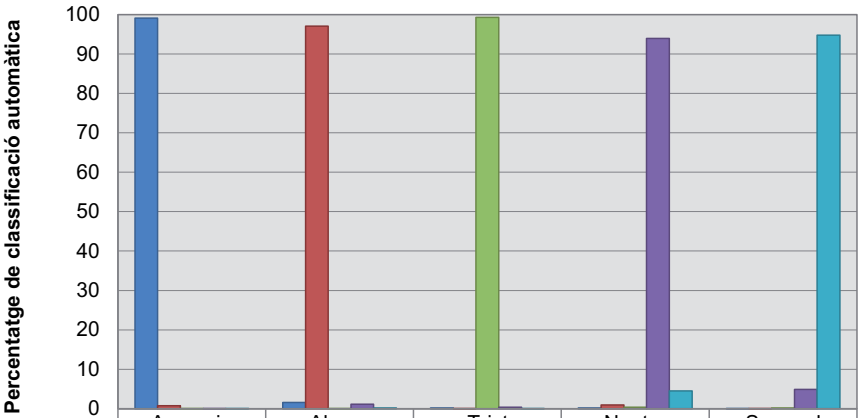
	J4.8	PART	DT	IB1	IBk	NB	SVM	AB/J4.8	AB/PART	AB/DT
data2G	96,36	96,87	91,63	97,54	97,93	96,94	98,86	98,23	98,38	95,49
data2L	95,97	96,92	92,35	96,10	96,53	97,59	98,47	98,17	98,40	95,45
data2LC	95,47	96,42	91,79	96,94	97,20	96,74	98,47	97,84	98,08	95,19
data2LS	87,82	88,77	81,65	86,46	88,42	91,01	93,57	92,37	92,67	89,28

(b)

Figura 5.6: Resultats WAR de l'experiment de reconeixement afectiu bàsic en el corpus BDP-UAB. La figura (a) mostra els resultats obtinguts pels subconjunts de dades derivats del conjunt de dades original complet. La figura (b) mostra els resultats obtinguts pels subconjunts de dades derivats del conjunt de dades original ometent la segona derivada. El resultat de l'algorisme de referència ZeroR és 22,60%.

La figura 5.7 mostra la matriu de confusió de l'algorisme SVM (que obté el millor resultat en mitjana) i el conjunt de dades *data2G* (que obté, també en mitjana, el millor rendiment). La conclusió que es desprèn és que els algorismes tendeixen a confondre, principalment, els estats neutre i sensual a diferència dels avaluadors subjectius, que ten-

deixen a confondre els estats agressiu i alegre (vegeu la figura 3.5). Aquesta diferència és deguda a l'absència de paràmetres relatius a la VoQ en aquest experiment ja que tots dos estils emotius tenen una prosòdia similar però la veu sensual és més xiuxiuejant que la neutra, diferència que sí és percebuda pels avaluadors subjectius (Iriondo *et al.*, 2007c).



	Agressiu	Alegre	Trist	Neutre	Sensual
Agressiu	99,07	1,57	0,17	0,22	0,04
Alegre	0,79	97,08	0,08	0,94	0,08
Trist	0,08	0,02	99,25	0,41	0,22
Neutre	0,04	1,17	0,40	93,93	4,89
Sensual	0,02	0,16	0,09	4,50	94,78

Figura 5.7: Matriu de confusió dels resultats de classificació de l'algorisme SVM i el conjunt de dades *data2G* en l'experiment de reconeixement afectiu bàsic amb el corpus BDP-UAB.

En (Monzo *et al.*, 2007) es recull un estudi posterior dut a terme considerant únicament els paràmetres de VoQ del corpus BDP-UAB. Aquest estudi constata el que s'ha explicat abans quant a la distinció entre els estils agressiu i alegre a través de paràmetres prosòdics, i neutre i sensual a través de paràmetres relatius a la VoQ. L'estudi mostra que els paràmetres de VoQ són suficients per destriar entre els diferents estils emocionals excepte, precisament, els dos primers. A més es demostra que l'estil alegre només es caracteritza per un únic paràmetre de VoQ (SFM) mentre que el sensual es caracteritza per fins a cinc d'ells (*jitter*, *shimmer*, HNR, GNE i do1000).

5.1.3 Conclusió

Els dos experiments anteriors mostren els resultats de classificar un corpus de veu actuada i un corpus de veu estimulada, és a dir, corpus gravats en condicions molt controlades tant a nivell de contingut com de mitjans tècnics per garantir uns requisits mínims de qualitat, i tots dos amb un únic locutor. Així, malgrat emprar parametritzacions diferents (segmental i fonètica), les taxes de classificació obtenen valors elevats, per damunt fins i tot de les obtingudes en els tests subjectius, aspecte que també s'observa als resultats obtinguts per diferents investigacions en condicions similars tal com es recull en l'estat de

la qüestió. No obstant això, seria incorrecte afirmar que això és l'objectiu desitjat. Podem afirmar que el sistema està reconeixent estils concrets de parla d'una veu determinada però no que està reconeixent una emoció en general. A més, si el corpus serà emprat en síntesi de veu, una validació automàtica dels estils expressius del corpus que ofereixi millor resultat (diferent, en general) que una subjectiva no és interessant atès que el sistema estaria donant per bones locucions en les quals els avaluadors subjectius no aconsegueixen reconèixer l'emoció desitjada.

Per aquest motiu, a continuació s'exposa un treball destinat a realitzar un reconeixement afectiu automàtic que aconsegueixi un resultat igual o molt similar al subjectiu. D'altra banda, en el següent capítol s'exposaran els experiments realitzats en un entorn de diversos locutors i, en general, més realista que el considerat en els treballs del present capítol.

5.2 Reconeixement afectiu de mapatge subjectiu

A diferència dels experiments de reconeixement afectiu bàsic, el reconeixement afectiu de mapatge subjectiu no pretén aconseguir la màxima taxa de classificació. Tal com s'ha mostrat en la secció anterior, els algorismes de classificació i els conjunts de dades presentats són capaços d'aconseguir taxes molt elevades, per damunt fins i tot de les aconseguides en les proves subjectives. En aquest cas, el reconeixement de mapatge subjectiu pretén aconseguir els mateixos resultats que les proves subjectives, classificant de forma automàtica cadascun dels estils emocionals tal com els classifiquen els avaluadors subjectius.

5.2.1 Plantejament general del sistema

Tal com es recull en (Iriondo *et al.*, 2009), un sistema de reconeixement afectiu de mapatge subjectiu consta d'una primera fase en la qual el sistema és entrenat amb un subconjunt petit procedent del corpus original, considerant al mateix temps l'avaluació subjectiva realitzada sobre el mateix subconjunt. Una vegada el sistema és entrenat, és capaç d'analitzar el corpus complet i eliminar totes aquelles locucions que no es corresponen amb l'estil emocional pretès i que, en conseqüència, també serien descartades en una prova subjectiva. Les locucions que millor es corresponen amb la classificació subjectiva romanen en el corpus resultant. Aquest procés és d'utilitat en la validació de corpus actuats, eliminant de forma automàtica aquells enregistraments en els quals el locutor no ha estat capaç d'expressar de forma convincent l'estil emocional requerit, sense necessitat de realitzar una prova subjectiva que suposi escoltar la totalitat del corpus. Així doncs, aquest sistema se centra en l'oient (vegeu la figura 3.1) i, per tant, considera més importants els resultats de l'avaluació subjectiva que els de el sistema automàtic.

L'estructura genèrica dels experiments de reconeixement afectiu de mapatge sub-

jectiu es pot observar en la figura 5.1. L'esquema és molt similar a la dels experiments de reconeixement afectiu bàsic però, en aquest cas, els resultats de l'avaluació subjectiva ajusten el procés d'avaluació objectiva. En aquesta secció es proposa un procés d'ajust tal com es detalla a continuació.

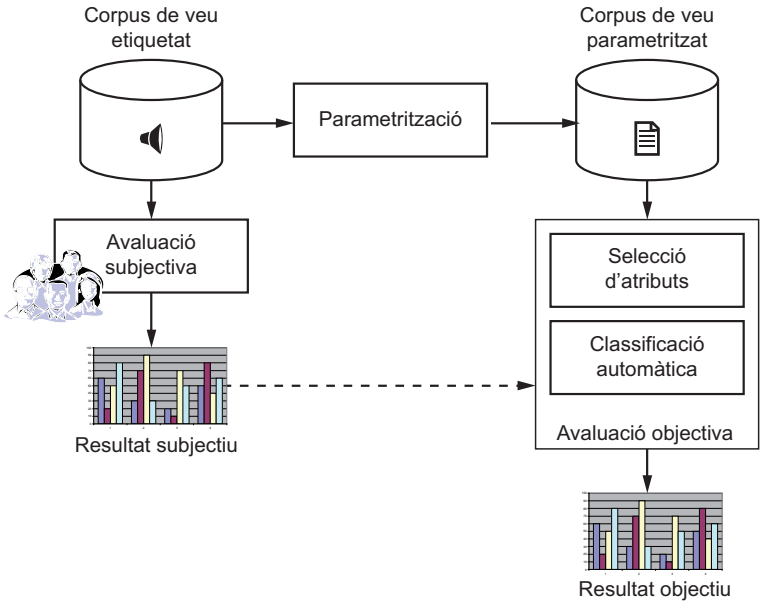


Figura 5.8: Esquema d'un experiment genèric de reconeixement afectiu de mapatge subjectiu.

5.2.2 Disseny del sistema

Per realitzar l'experiment de reconeixement afectiu de mapatge subjectiu s'empra el corpus BDP-UAB considerant la seva parametrització completa, tant a nivell de prosòdia com de VoQ. La prova subjectiva emprada per a l'entrenament del sistema és la mateixa que la indicada en la secció 3.4.2.1. De forma heurística, i sobre la base dels histogrames relacionats amb l'avaluació subjectiva (vegeu la figura 3.6), es determina que una locució és *incorrecta* (en el sentit de no correspondre's de forma adequada a l'estil emocional pretès) si el seu grau d'identificació és inferior al 50% o si el percentatge d'etiquetes "NS/NC" (és a dir, etiquetes assignades pels avaluadors en cas de no saber distingir l'estil expressiu) que rep és superior al 12%, ja que poden considerar-se atípics en els histogrames. En cas contrari, la locució es considera *correcta*. Segons aquestes dues regles heurístiques, 33 locucions són etiquetades com *incorrectes* sobre la base de l'avaluació subjectiva. Per la seva banda, el sistema de reconeixement afectiu de mapatge subjectiu assignarà una etiqueta categòrica a cada locució. Si l'etiqueta assignada no es correspon amb la que tenia en el seu etiquetatge original, la locució es considerarà *incorrecta*. Es considerarà *correcta* en cas contrari. L'objectiu de la fase d'entrenament és aconseguir que el sistema automàtic i l'avaluació subjectiva coincideixin en l'etiquetatge *correcta/incorrecta* que es faci de cada locució.

El funcionament global del sistema es mostra en la figura 5.9. Els algorismes de classificació són provats amb les 480 locucions involucrades en la prova subjectiva i entrenats amb les 4.158 locucions restants. A les locucions incorrectament classificades se'ls assigna l'etiqueta d'*incorrectes*. Aquesta assignació és comparada després amb la de la prova subjectiva. Per realitzar la comparació es calcula la mesura F1 que té en consideració tant la precisió com la cobertura de les locucions incorrectament classificades sobre la base de les 33 rebutjades per la prova subjectiva, i és aquesta mesura la que guia el procés de selecció de paràmetres. Aquest procés de selecció determina si el conjunt de dades ha de ser ajustat per optimitzar la mesura F1. L'ajust pot realitzar-se mitjançant una cerca voraç incremental, decremental o mixta, tal com es descriuen aquests processos en la secció 4.3.2.

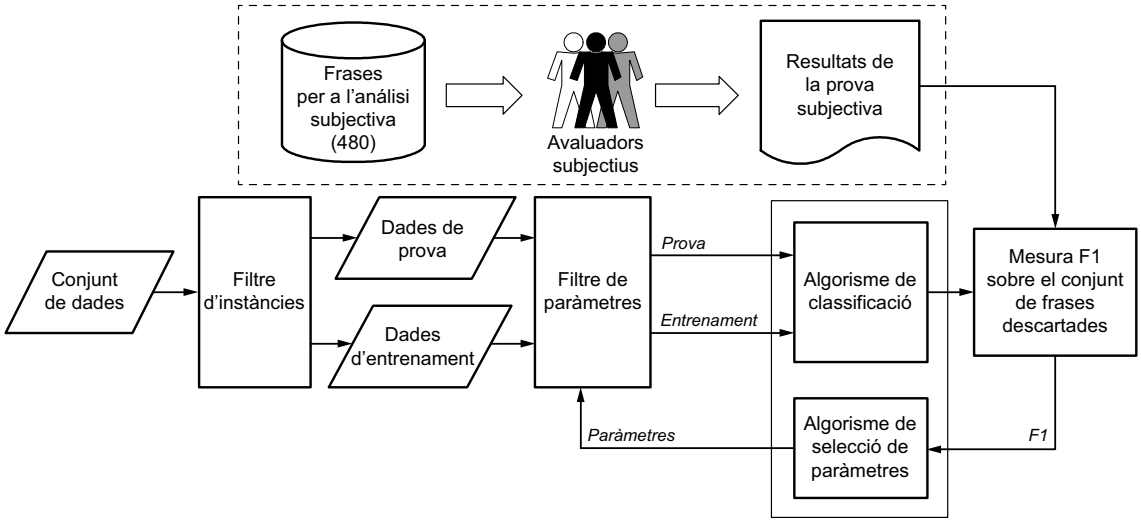


Figura 5.9: Ajust del sistema objectiu de reconeixement afectiu guiat pels resultats d'una prova subjectiva. Adaptat d'(Iriondo *et al.*, 2009).

5.2.3 Resultats preliminars

En (Iriondo *et al.*, 2007a) es recull un experiment preliminar basat en l'esquema proposat. El conjunt de dades de treball és el *data2LC* atès que aconsegueix un compromís òptim entre el nombre de paràmetres (68 paràmetres) i els resultats aconseguits (96,58% de taxa de reconeixement mitjana) (Iriondo *et al.*, 2009). Seguint l'esquema proposat en la figura 5.9, el procés de selecció d'atributs consta d'una cerca voraç en dues de les seves variants, incremental i decremental. Els algorismes de classificació analitzats són Naïve-Bayes, SVM i J4.8. El màxim valor de F1 (0,50) s'obté per l'algorisme SVM i la cerca voraç decremental. El nombre més gran de coincidències (18) entre les frases descartades de forma objectiva i de forma subjectiva s'obté amb l'algorisme J4.8 i la cerca voraç incremental. No obstant això, aquest esquema també obté un elevat nombre de classificacions incorrectes (51), la qual cosa es reflecteix en el valor de la mesura F1 (0,43).

No obstant això, aquest treball presenta algunes limitacions que s'intenten esmenar a continuació. En primer lloc, no es recull informació procedent de la VoQ sinó que es limita a analitzar únicament la prosòdia de les locucions. La incorporació d'aquesta informació pot ser útil per destriar entre determinats estils de la parla, tal com s'ha comentat anteriorment. D'altra banda, la cerca voraç exclusivament incremental o decremental no és capaç de desfer decisions prèvies, mentre que una estratègia combinada podria evitar aconseguir màxims locals associats a resultats pobres. I, finalment, atès que alguns classificadors són més precisos que uns altres, la combinació de diversos algorismes podria millorar el resultat final.

5.2.4 Inclusió de paràmetres de VoQ

La incorporació dels 55 paràmetres de VoQ extrets del corpus BDP-UAB (vegeu la secció 3.4.2.2) al conjunt de dades *data2LC* resulta en un conjunt de dades de 123 paràmetres la definició dels quals es mostra, de forma gràfica, en la figura 5.10, tal com es recull en (Iriondo *et al.*, 2007b). El nou conjunt de dades rep el nom de *data2LCVQ5*.

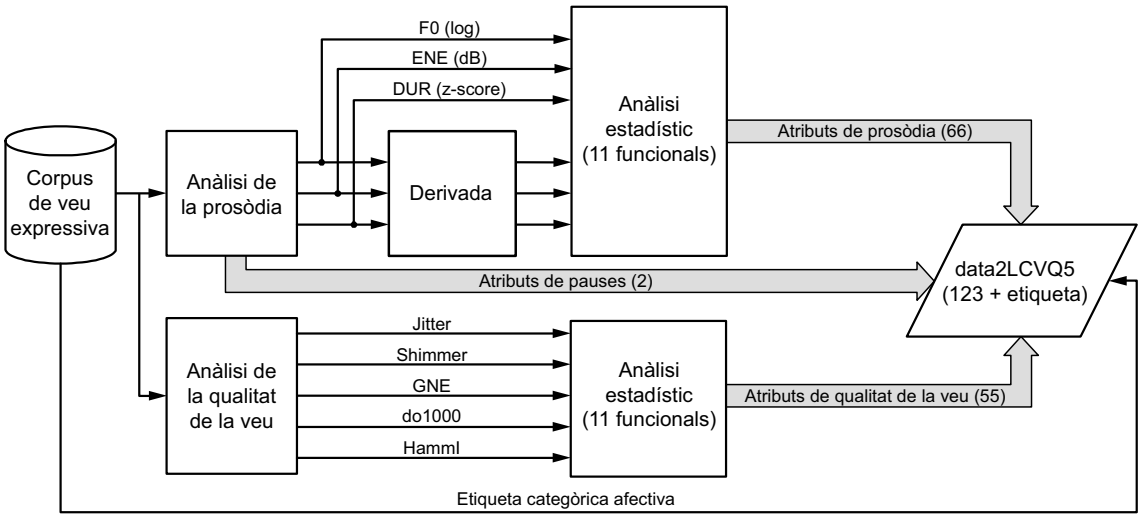


Figura 5.10: Descripció gràfica del corpus *data2LCVQ5*. ENE i DUR fan referència als paràmetres d'energia i durada, respectivament. Adaptat d'(Iriondo *et al.*, 2007b).

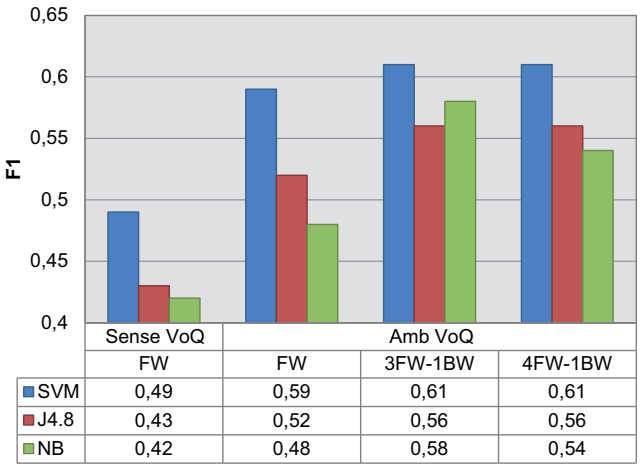
L'efecte de la incorporació dels paràmetres de VoQ als paràmetres prosòdics pot observar-se en les dues primeres columnes de la figura 5.11.

5.2.5 Selecció de paràmetres ampliada

En (Iriondo *et al.*, 2007b) es proposa una estratègia de selecció d'atributs basada en una cerca voraç bidireccional com la descrita en la secció 4.3.2. Si bé el primer ajust es realitza per a un esquema 3FW-1BW, en (Iriondo *et al.*, 2009) també es considera un

esquema 4FW-1BW. En cada iteració s'escull el subconjunt de paràmetres que maximitza la mesura F1, afegint 3 o 4 atributs en cada pas cap endavant, respectivament, i eliminant 1 en cada pas enrere.

La figura 5.11 mostra els resultats aconseguits pels classificadors SVM, J4.8 i Naïve-Bayes en combinació amb els diferents mètodes de selecció d'atributs 3FW-1BW i 4FW-1BW. Per a tots els casos, la presència de paràmetres de VoQ suposa una millora dels resultats. D'altra banda, l'estratègia 3FW-1BW també millora els resultats de la mesura F1. Els resultats més significatius s'obtenen per a J4.8 i Naïve-Bayes. Ambdues modificacions suposen una millora relativa superior al 20% per als tres classificadors. Respecte a l'estratègia de 3FW-1BW, l'estratègia 4FW-1BW iguala els resultats per SVM i J4.8, però els empitjora per Naïve-Bayes.



	Sense VoQ	Amb VoQ		
	FW	FW	3FW-1BW	4FW-1BW
SVM	0,49	0,59	0,61	0,61
J4.8	0,43	0,52	0,56	0,56
NB	0,42	0,48	0,58	0,54

Figura 5.11: Valors màxims de F1 per als algorismes SVM, J4.8 i Naïve-Bayes (NB) considerant l'absència i presència de paràmetres de VoQ (conjunts de dades *data2LC* i *data2LCVQ5*, respectivament). S'inclouen les estratègies de cerca cap endavant (FW), 3FW-1B i 4FW-1BW.

La figura 5.12 mostra l'evolució del valor màxim de F1 dels tres algorismes anteriors amb les estratègies de selecció d'atributs considerades.

5.2.6 Fusió de classificadors

La fusió de classificadors es realitza mitjançant dues estratègies de *stacking* (vegeu la secció 4.3.1) diferents: mitjançant un sistema de vot majoritari ponderat i a través d'un algorisme de regles PART.

El sistema de vot majoritari combina 7 dels 9 classificadors que consideren els paràmetres prosòdics i de VoQ de la figura 5.11. Es descarten, per tant, els algorismes J4.8/FW i NB/FW per obtenir els pitjors resultats. El sistema de votació té en compte l'assignació *correcta/incorrecta* (en referència a si es correspon amb l'estil emocional pretès o no, respectivament) que cada classificador realitza d'una mateixa locució i s'estableix un nombre

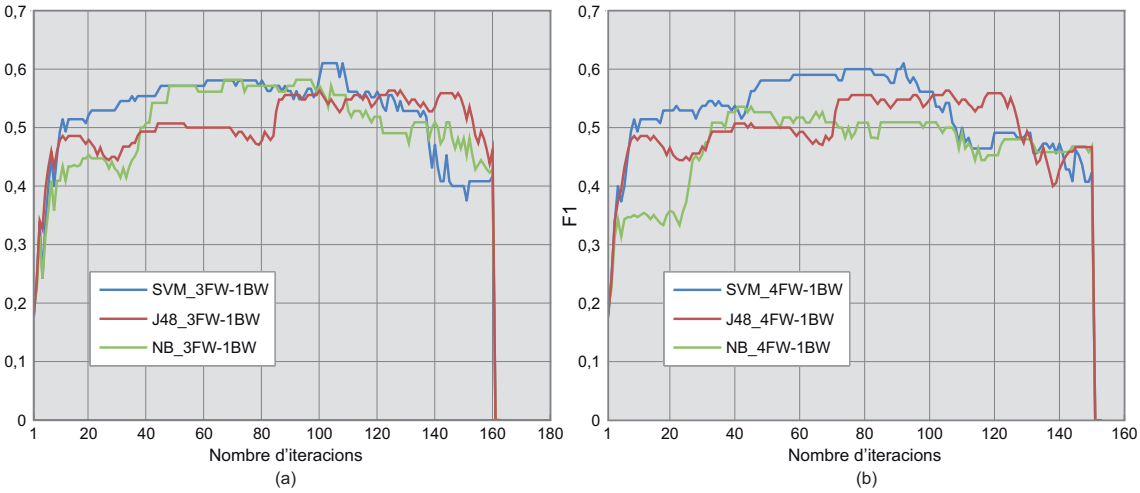


Figura 5.12: Valors màxims de F1 per iteració considerant el conjunt de dades *data2LCVQ5* amb les estratègies de selecció de paràmetres: (a) 3FW-1BW i (b) 4FW-1BW.

mínim de vots per considerar com *incorrectes* cadascuna d'elles. Atès que una anàlisi de cada estil mostra que les locucions corresponents a l'estil *agressiu* són detectades com *incorrectes* per pocs classificadors, es crea un sistema ponderat que multiplica per 2 el nombre de vots que reben les locucions d'aquest estil. La figura 5.13 mostra l'evolució de precisió, cobertura i mesura F1 del sistema en funció del nombre mínim de vots que es considerin necessaris per determinar que una locució sigui *incorrecta*. El valor màxim de F1 és 0,71, obtingut amb un mínim de 4 vots, la qual cosa millora el millor resultat individual (0,61) de forma significativa.

Una alternativa al sistema de votació és entrenar un classificador de nivell 1 que prengui una decisió final a partir dels resultats dels classificadors de la figura 5.11 (classificadors de nivell 0). En aquest cas s'escull l'algorisme PART de creació de regles de decisió. El resultat de la mesura F1 és 0,73, el qual és lleugerament superior al de votació. L'algorisme 1 mostra les regles creades per PART, en el qual els classificadors escollits són C1=SVM(3FW-1BW), C2=J4.8(3FW-1BW), C3=J4.8(4FW-1B) i C4=Naïve-Bayes(3FW-1BW). Els resultats de cada classificador s'expressen com 0 o 1, on 0 significa que la locució es classifica de forma correcta i 1 significa que la locució no es classifica correctament. El nombre de casos classificats de forma *correcta/incorrecta* es mostra entre parèntesis. Les regles s'apliquen en l'ordre establert i, en el cas que no es compleixin les condicions de cadascuna d'elles, s'assigna, per defecte, la classe majoritària.

5.3 Resum

Aquest capítol ha recollit diferents experiments de reconeixement afectiu automàtic relacionats amb corpus de veu actuada i estimulada gravats en condicions òptimes. S'han considerat dues perspectives diferents.

Algorisme 1 Algorisme PART per a la fusió de classificadors en el procés de reconeixement afectiu de mapatge subjectiu.

C1 = 0 i C2 = 0 i C3 = 0 i C4 = 0: Correcta (408/10)
C1 = 0 i Estil = AGR i C2 = 1: Incorrecta (7/2)
C1 = 0: Correcta (25)
C3 = 1: Incorrecta (18/3)
C2 = 0 i C4 = 0: Correcta (4/1)
C2 = 0: Incorrecta (2)
: Correcta

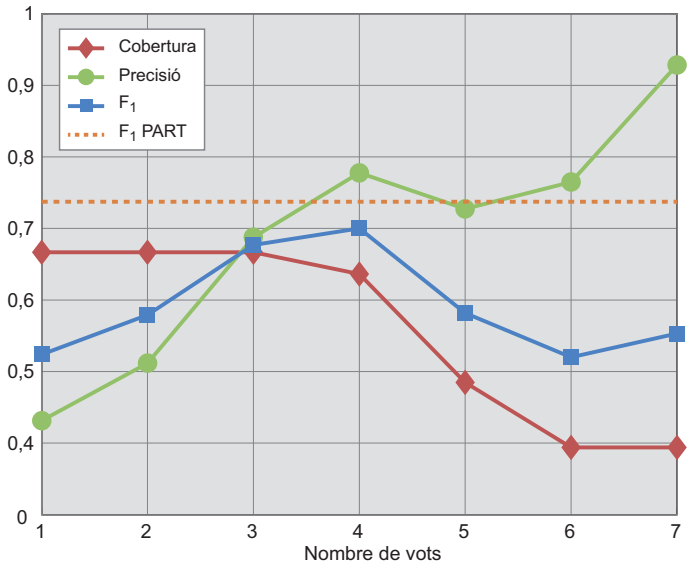


Figura 5.13: F1, cobertura i precisió de la fusió de classificadors per votació en funció del mínim nombre de vots necessari per considerar les locucions com *incorrectes*. S'inclou el resultat de la F1 aconseguida per la fusió mitjançant l'algorisme PART.

En primer lloc, s'ha plantejat un reconeixement afectiu automàtic a nivell bàsic consistent a maximitzar la WAR de dos corpus diferents mitjançant paràmetres acústics a nivell prosòdic i de VoQ. Comparant els resultats obtinguts pels classificadors i les proves subjectives realitzades per avaluadors humans es comprova que la taxa de reconeixement aconseguida pels algorismes objectius és, en ocasions, fins i tot superior a l'obtinguda de forma subjectiva. No obstant això s'aprecien diferències a nivell de categories concretes en funció de l'esquema d'aprenentatge analitzat en cada cas i de la parametrització escollida.

En segon lloc s'ha proposat un sistema automàtic orientat a obtenir uns resultats de classificació propers als obtinguts de forma subjectiva. És a dir, pretenent que la classificació automàtica segueixi el criteri subjectiu amb la finalitat de poder realitzar una validació automàtica d'un corpus extens sense necessitat d'invertir grans recursos humans en aquesta tasca. A diferència de la proposta realitzada en primer lloc, en aquesta ocasió els resultats subjectius obtinguts a partir d'un fragment del corpus serveixen per ajustar el sistema de reconeixement automàtic permetent aconseguir una aproximació del mapat-

ge desitjat. Aquest ajust es realitza mitjançant la selecció adequada dels paràmetres per maximitzar la mesura F1 calculada sobre la base de la classificació subjectiva i l'objectiva. Per a això s'empren algorismes de cerca voraç cap endavant, enrere i combinada. Així mateix, s'aplica la fusió de diversos classificadors mitjançant dues tècniques de *stacking* que permet millorar la classificació d'aquelles classes que són classificades amb diferent precisió per diferents algorismes.

Els resultats dels experiments anteriors posen de manifest que l'anàlisi de corpus de veu gravats per actors en condicions òptimes és una tasca que proporciona, en general, bons resultats. No obstant això, aquests experiments treballen dins d'un escenari difícilment reproduïble en una aplicació real. Aquests esquemes d'aprenentatge reconeixen emocions plenes concretes que procedeixen d'un únic locutor pel que disten molt d'emular el comportament humà, el qual és capaç de reconèixer una àmplia varietat d'estils afectius (amb els seus diferents graus d'intensitat) de locutors diferents. És per això que en el següent capítol s'aborden experiments destinats a treballar en un entorn més realista, en els quals es posaran de manifest diferències notables amb els resultats aquí obtinguts.

Capítol 6

Reconeixement afectiu en corpus de veu espontània

Aquest capítol se centra en el reconeixement afectiu automàtic realitzat en el corpus FAU Aibo, un corpus d'expressió vocal espontània de diversos locutors que presenta, per les seves pròpies condicions d'enregistrament, unes condicions molt diferents de les estudiades en els experiments del capítol 5 i que suposen, en conseqüència, un plantejament diferent dels experiments duts a terme. Aquest capítol detalla aquest nou plantejament i les diferències respecte als experiments presentats en l'anterior, tant en la seva definició com en els resultats obtinguts.

En primer lloc, aquest capítol se centra en la metodologia que segueixen els diferents experiments plantejats. A continuació es detallen els experiments de reconeixement afectiu agrupats en tres blocs segons el seu objecte d'anàlisi: paràmetres acústics i paràmetres lingüístics de forma independent, i la fusió d'ambdues modalitats per a un estudi conjunt.

En aquests experiments es tracta l'exercici de classificació des de diferents perspectives, considerant tant esquemes de classificació individuals com a estructures més complexes, agrupant diversos classificadors mitjançant estructures jeràrquiques, estructures en cascada o mitjançant fusió a nivell de decisió. Les modalitats acústica i lingüística es fusionen a nivell de decisió i a nivell de paràmetres. D'altra banda, s'aborden experiments de selecció de paràmetres destinats a reduir el volum de dades amb la finalitat de reduir la complexitat dels esquemes de reconeixement proposats i optimitzar les taxes de classificació aconseguides.

6.1 Metodologia

La metodologia dels experiments presentats en aquest capítol difereix de la presentada en el capítol anterior en diversos aspectes que es detallen a continuació. Bàsicament, aquestes diferències fan referència a l'esquema general dels experiments, a la mètrica que s'empra per mesurar i comparar els seus resultats i els mètodes d'avaluació dels esquemes d'aprenentatge.

6.1.1 Esquema general

A diferència dels experiments proposats per al reconeixement afectiu automàtic en corpus gravats per actors, com els exposats en el capítol 5, els experiments de reconeixement afectiu en corpus espontanis no poden dividir-se en els dos grups que s'havien proposat: experiments de reconeixement afectiu bàsic i de mapatge subjectiu. En primer lloc perquè, donada la pròpia naturalesa del corpus, no té sentit plantejar una validació del mateix. L'enregistrament recull locucions amb un estil afectiu natural que no se cenyeix a cap guió previ pel que, atès que no hi ha un estil pretès, no poden descartar-se locucions per no expressar correctament un estil afectiu concret. En segon lloc perquè un corpus espontani sempre és etiquetat en una fase posterior al seu enregistrament. Això implica que l'etiquetatge del corpus ja està realitzat sobre la base d'un criteri subjectiu pel que qualsevol experiment de reconeixement afectiu automàtic està realitzant, de forma implícita, un mapatge subjectiu. Així doncs, un experiment de reconeixement afectiu automàtic en un corpus espontani és, quant a estructura, molt similar a un de reconeixement afectiu bàsic, si bé existeix implícit un mapatge del criteri subjectiu.

La figura 6.1 mostra l'esquema d'un experiment de reconeixement afectiu automàtic basat en un corpus de veu espontània. A diferència de l'esquema proposat en la figura 5.1, l'etiquetatge del corpus és posterior a l'enregistrament del mateix i es realitza de forma subjectiva[1], procés que substitueix a la fase d'avaluació subjectiva. La resta de mòduls sí són iguals als de l'esquema anterior. Els experiments que es recullen en aquest capítol segueixen aquesta estructura genèrica.

6.1.2 Mètrica

A més de la diferència en l'esquema genèric proposat, els experiments de reconeixement afectiu automàtic en el corpus de veu espontània FAU Aibo recollits en aquest capítol també es diferencien dels del capítol anterior en la mètrica emprada per mesurar el rendiment dels esquemes d'aprenentatge. En aquest cas es mesura la UAR en lloc de

[1] En el cas del corpus FAU Aibo, aquest etiquetatge subjectiu consisteix en l'assignació d'una etiqueta categòrica a cada paraula de cadascuna de les locucions del corpus. Aquest procés involucra a cinc avaluadors. Posteriorment, un algorisme heurístic assigna una etiqueta global a la locució a partir de l'avaluació individual de cada paraula (Steidl, 2009).

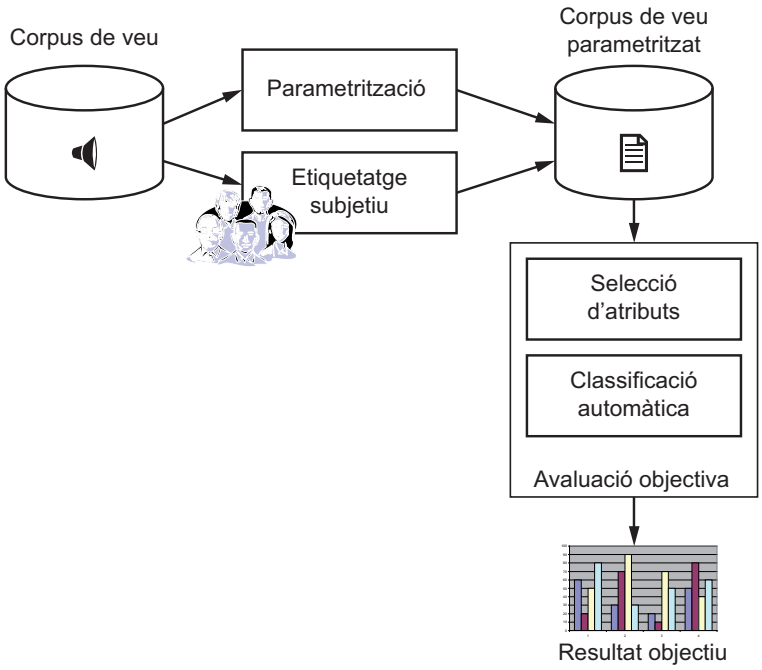


Figura 6.1: Esquema d'un experiment genèric de reconeixement afectiu en corpus de veu espontània.

la WAR atès que el corpus FAU Aibo és un corpus no equilibrat, és a dir, existeix una notable diferència entre el nombre de casos d'uns estats afectius enfront d'uns altres (vegeu la secció 3.4.3.1). Per exemple, la classe majoritària és la corresponent a l'estat neutre, la qual cosa és conforme amb un escenari real, en el qual les interaccions neutres són les més abundants ja que rarament es produeixen expressions emocionals (Batliner *et al.*, 2003). Per aquest motiu, en un escenari d'aquestes característiques és tan important detectar les locucions amb contingut emocional com les neutres i és per això que la UAR és d'especial utilitat, ja que no mesura únicament el nombre de casos correctament classificats sinó que té en consideració el nombre de casos de cada classe que es classifiquen de forma correcta (vegeu la secció 4.3.4).

No obstant això, en els primers experiments d'aquest capítol també s'inclourà el resultat de la WAR amb la finalitat de comparar resultats i analitzar el comportament general dels classificadors estudiats en un corpus d'aquestes característiques. No obstant això, la mesura a maximitzar serà la UAR per les raons abans exposades.

6.1.3 Mètodes d'avaluació

El corpus FAU Aibo ofereix dos subconjunts diferents que Schuller *et al.* (2009) escullen com a conjunt de dades d'entrenament i conjunt de dades de prova (vegeu la secció 3.4.3.1). Aquesta divisió permet que els esquemes d'aprenentatge puguin ser entrenats i provats de forma independent. Això permet destacar diversos aspectes:

- Es treballarà en un entorn de diversos locutors en els quals el sistema serà provat amb veus diferents a les quals es van emprar per al seu entrenament. Això permetrà observar si el sistema és capaç de reconèixer expressions emocionals genèriques i no veus concretes ni expressions emocionals pròpies d'usuaris específics.
- Atès que tots dos conjunts estan formats per locucions gravades per persones diferents, les expressions emocionals seran lleugerament diferents en la forma en la qual són expressades. L'anàlisi per separat permetrà comprovar si el sistema és capaç de reconèixer l'emoció subjacent dins de les particularitats de cada expressió concreta.
- Entrenar i provar amb conjunts diferents permetrà emular el funcionament del sistema de reconeixement en un entorn real en el qual el sistema hauria de ser capaç de reconèixer els estats afectius sense necessitat d'haver estat entrenat amb els mateixos usuaris que després el faran servir.

Aquest mètode d'avaluació permet comparar els resultats de diferents experiments que segueixen la mateixa metodologia tal com es detalla en la secció 4.4.2. Això no impedeix, però, que no es puguin aplicar altres metodologies utilitzant, únicament, el subconjunt de dades d'entrenament com a pas previ a l'avaluació mitjançant els dos subconjunts, tals com una validació creuada de diverses iteracions o independent dels subjectes. Això facilitaria la tasca d'elecció de paràmetres i algorismes en treballar amb un conjunt de dades més reduït i, a més, permetria una posterior validació mitjançant el subconjunt de prova que, en tot moment excepte al final, romandria fora de l'experiment.

Els experiments que es recullen en aquest capítol empren diferents mètodes d'avaluació si bé tots ells realitzen, com a pas final, una avaluació mitjançant els subconjunts d'entrenament i prova per permetre la comparació dels seus resultats, tant entre ells com amb la resta d'experiments que segueixen la metodologia aquí explicada, que és l'emprada majoritàriament pels treballs realitzats sobre el corpus FAU Aibo.

Schuller *et al.* (2009) proposen un resultat de referència per a un entorn d'estudi com l'aquí descrit, dins del marc de l'*Emotion Challenge 2009* (vegeu la secció 6.2.1). Aquesta referència es correspon amb un valor de UAR del 38,2%, obtingut mitjançant un algorisme SVM de kernel lineal en analitzar el conjunt de dades acústiques estandarditzat i remostrejat mitjançant la tècnica SMOTE (Chawla *et al.*, 2002), tal com s'introduïa en la secció 4.4.2. Aquest és el resultat que també es considerarà com a referència al llarg d'aquest capítol.

6.2 Reconeixement afectiu basat en l'anàlisi de paràmetres acústics

Aquesta secció recull els experiments realitzats a nivell acústic sobre el corpus FAU Aibo, seguint la línia exposada en el capítol anterior. Es detallen dos experiments que

analitzen el corpus únicament a nivell acústic emprant la parametrització descrita per Schuller *et al.* (2009). Tots dos experiments estudien, en primer lloc, el subconjunt d'entrenament mitjançant dues tècniques diferents: una validació creuada de 10 iteracions (secció 6.2.1) i una estratègia LOSO (secció 6.2.2), respectivament, garantint la independència entre el subconjunt d'entrenament i el de prova durant la fase d'elecció del millor esquema de classificació. En segon lloc es validen els resultats obtinguts mitjançant un entrenament amb el primer subconjunt complet i la classificació dels elements del segon subconjunt. D'altra banda, la secció 6.2.1 presenta uns resultats preliminars tal com es recullen en (Planet *et al.*, 2009), mostrant diferents esquemes de classificació i una primera selecció de paràmetres. La secció 6.2.2 empra un conjunt de dades millorat i presenta altres esquemes de classificació, incloent una aproximació bidimensional, i una selecció de paràmetres més precisa.

6.2.1 Proposta per l'*Emotion Challenge 2009*

El primer experiment de reconeixement afectiu automàtic amb veu espontània es realitza a partir de les regles establertes per Schuller *et al.* (2009) en el marc de l'*Emotion Challenge* de la X Conferència *INTERSPEECH* 2009, analitzant el corpus FAU Aibo. Aquestes regles determinen el corpus FAU Aibo com a element d'estudi per a tots els participants, obrint-se per primera vegada per a la comunitat científica en general i definint els dos subconjunts abans citats. Així mateix ofereix una parametrització acústica realitzada mitjançant una versió preliminar del programari openSmile inclòs en el paquet openEAR (Eyben *et al.*, 2009), tal com es descriu en la secció 3.4.3.2. Els paràmetres es proporcionen en el format adequat per ser processats mitjançant l'eina WEKA.

Aquest experiment s'emmarca en dues de les tres parts que formen l'*Emotion Challenge 2009* (tal com s'assenyalava en la secció 4.4.2). L'objectiu general és maximitzar la UAR. Els detalls de les parts es descriuen a continuació:

- Desafiament de classificació (*Classifier Sub-Challenge*). Consisteix en l'estudi dels paràmetres acústics proporcionats per l'organització de l'*Emotion Challenge 2009* (vegeu la taula 3.4), de forma exclusiva i sense afegir cap altre paràmetre si bé es permet la modificació dels paràmetres originals. L'anàlisi pot realitzar-se mitjançant un únic classificador o combinant dos o més. Els experiments es detallen en la secció 6.2.1.1.
- Desafiament de paràmetres (*Feature Sub-Challenge*). Aquesta part consisteix en la cerca dels paràmetres més rellevants, fins a un màxim de 100, per a la maximització dels resultats de la UAR aconseguida pels esquemes de classificació. Es permet la inclusió de nous paràmetres als ja proporcionats pels organitzadors. Aquesta anàlisi es detalla en la secció 6.2.1.2.

Les propostes i els resultats obtinguts en ambdues parts de l'estudi es recullen en (Planet *et al.*, 2009). A continuació s'expliquen els detalls dels mateixos.

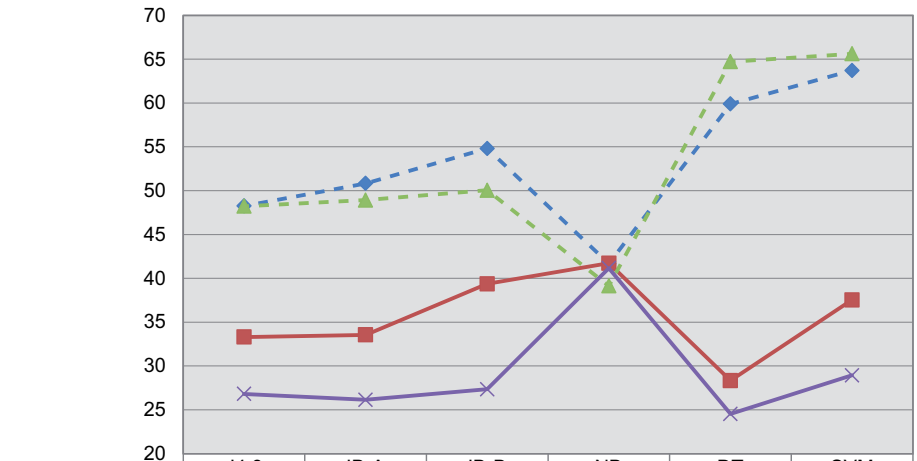
6.2.1.1 Desafiament de classificació

La proposta presentada en el desafiament de classificació de l'*Emotion Challenge 2009* consta de quatre estratègies de classificació. En primer lloc es planteja una bateria de proves amb diferents classificadors, similar al plantejament que es va presentar en la secció 5.1.1. Després, amb la finalitat de millorar els resultats obtinguts, es dissenyen dos metaclassificadors jeràrquics i un més amb estructura en cascada. En tots els casos l'avaluació consisteix a avaluar diversos classificadors mitjançant una validació creuada de 10 iteracions considerant, únicament, el subconjunt d'entrenament. Després es repeteix l'avaluació mitjançant un entrenament amb el primer subconjunt i la consegüent avaluació amb el segon. Els resultats exposats en aquesta secció recullen els resultats d'ambdues avaluacions.

Avaluació d'algorismes simples de classificació. En el primer cas, en el qual s'avaluen diferents algorismes de classificació, els algorismes escollits són els determinats com més rellevants sobre la base dels resultats obtinguts en l'experiment preliminar de la secció 5.1.2: l'arbre de decisió J4.8, dos algorismes KNN amb selecció d'un veí més proper i basats en la distància euclidiana (IB-A) i en la distància Manhattan (IB-B), Naïve-Bayes amb discretització supervisada prèvia del conjunt de dades, l'algorisme de taules de decisió DT i un algorisme SVM de kernel lineal, entrenament mitjançant l'algorisme d'optimització seqüencial mínima i emprant discriminació per parelles.

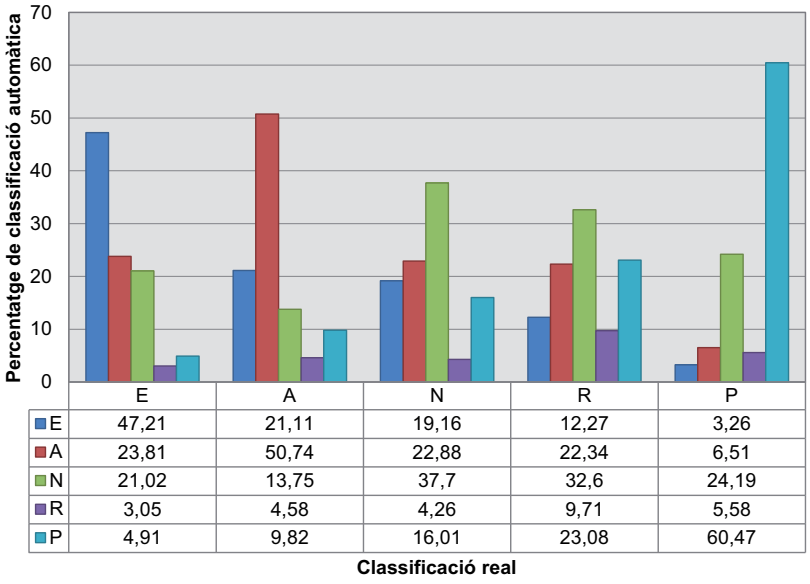
La figura 6.2 mostra els resultats dels classificadors anteriors segons els dos mètodes d'avaluació abans descrits. Es poden observar diferències entre els resultats obtinguts per la validació creuada i l'entrenament i prova amb els dos conjunts independents. En general, els resultats procedents de l'avaluació de conjunts independents és pitjor, en termes de UAR, que els obtinguts per la validació creuada. No obstant això, l'algorisme Naïve-Bayes no és particularment susceptible a l'avaluació amb un mètode o un altre, encara que obté un resultat baix en termes de WAR. Naïve-Bayes, a més, millora el resultat de referència en obtenir un valor de UAR en l'avaluació amb dos conjunts independents del 41,16%, en un 2,96% absolut (7,75% relatiu). La figura 6.3 mostra la matriu de confusió d'aquest classificador en la seva avaluació mitjançant els dos conjunts independents. Noti's la dificultat per classificar correctament els casos de la classe R, la qual cosa és conforme amb la pròpia indefinició d'aquesta categoria afectiva. El millor resultat en termes de WAR s'obté mitjançant l'algorisme SVM de kernel lineal, essent la seva UAR la segona millor de l'experiment.

Estructura jeràrquica de dos nivells basada en un classificador binari i un de general. La idea d'aquest algorisme jeràrquic és classificar les locucions de la classe N en un primer nivell de decisió mitjançant un classificador especialitzat i, posteriorment, classificar la resta de locucions mitjançant un classificador de segon nivell. La figura 6.4 representa aquest classificador jeràrquic en el qual es planteja un primer nivell de decisió en el que les



	J4.8	IB-A	IB-B	NB	DT	SVM
10-FCV WAR	48,26	50,83	54,82	41,67	59,89	63,71
10-FCV UAR	33,31	33,55	39,37	41,72	28,33	37,54
Test set WAR	48,23	48,92	50,05	39,14	64,71	65,62
Test set UAR	26,82	26,14	27,35	41,16	24,53	28,94

Figura 6.2: Taxes de classificació ponderada (WAR) i no ponderada (UAR), en percentatge, dels classificadors segons la seva avaluació mitjançant una validació creuada de 10 iteracions (10-FCV) i mitjançant els dos conjunts independents (Test set). Adaptat de (Planet *et al.*, 2009).



	E	A	N	R	P
E	47,21	21,11	19,16	12,27	3,26
A	23,81	50,74	22,88	22,34	6,51
N	21,02	13,75	37,7	32,6	24,19
R	3,05	4,58	4,26	9,71	5,58
P	4,91	9,82	16,01	23,08	60,47

Figura 6.3: Matriu de confusió de l'algorisme Naïve-Bayes avaluat mitjançant els dos conjunts independents d'entrenament i de prova. Adaptat de (Planet *et al.*, 2009).

locucions són classificades de forma binària, segons pertanyin a la classe N o a una altra diferent (¬N). En el segon nivell de decisió un classificador és entrenat amb el corpus complet després d'haver estat remostrejat per aconseguir una distribució uniforme de

les classes[2]. En aquesta figura, com en totes les d'aquesta secció, els elements acolorits en blau indiquen relació amb la fase d'entrenament i els acolorits en vermell indiquen relació amb la fase de prova. Les línies de punts indiquen fluxos de dades que serveixen per entrenar un classificador mentre que les línies sòlides indiquen fluxos de dades que són avaluats en un classificador, independentment de si s'aquests fluxos intervenen en la fase d'entrenament o la de prova[3].

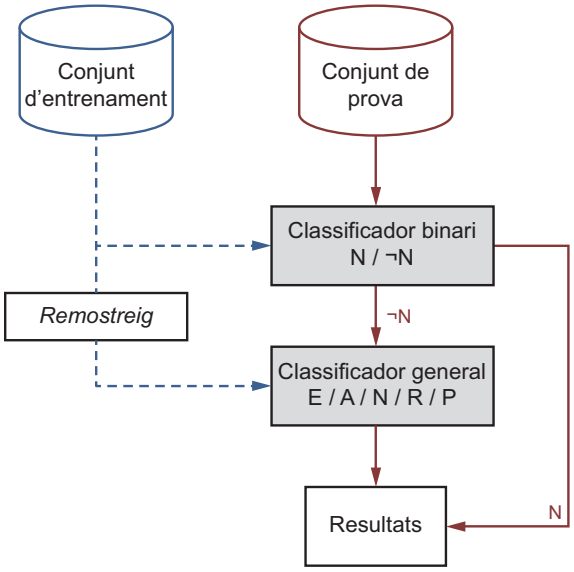


Figura 6.4: Esquema del classificador jeràrquic de dos nivells basat en un classificador binari de primer nivell i un classificador general de segon nivell. Els elements acolorits en blau indiquen relació amb la fase d'entrenament i els acolorits en vermell indiquen relació amb la fase de prova. Les línies de punts indiquen fluxos de dades que serveixen per entrenar un classificador mentre que les línies sòlides indiquen fluxos de dades que són avaluades en un classificador.

S'escullen dos classificadors de tipus SVM lineal com el descrit en l'experiment anterior[4]. En aquest cas es treballa amb dues versions diferents del conjunt de dades: l'original i una versió estandarditzada. La taula 6.1 mostra els resultats obtinguts en tots dos casos, els quals superen els de l'experiment anterior excepte els de l'algorisme Naïve-Bayes, que manté el millor resultat en termes de UAR. En comparació del classificador SVM de la secció anterior (el resultat del qual es mostra en la primera fila de la taula 6.1), el valor de WAR disminueix i augmenta el valor de UAR, la qual cosa és conforme amb

[2]La tècnica de remostreig emprada és senzilla i consisteix a eliminar alguns casos de les classes majoritàries i repetir alguns casos de les minoritàries. És, per tant, un procés de filtrat automàtic supervisat i permet reduir, mantenir o incrementar, segons sigui la configuració de l'experiment, el nombre de casos del conjunt de dades final.

[3]No és el cas d'aquest esquema però, en ocasions (per exemple, en l'esquema de la figura 6.5), és possible que durant la fase d'entrenament s'avaluïn algunes dades del conjunt d'entrenament en un classificador que ja ha estat entrenat amb la finalitat d'emprar el resultat de classificació per entrenar un segon mòdul. Aquestes dades apareixerien representats en color blau (per estar relacionats amb la fase d'entrenament) i una línia sòlida, ja que són avaluades per un classificador.

[4]Per simplificar, no s'inclouen els resultats d'altres algorismes atès que obtenen resultats inferiors en aquesta estructura.

l'objectiu plantejat. Atès que el conjunt de dades estandarditzat millora la UAR, aquesta versió serà l'escollida per a la resta d'esquemes d'aprenentatge d'aquesta secció.

Taula 6.1: Resultats per a l'avaluació mitjançant dos conjunts independents de l'estructura jeràrquica basada en dos nivells: un primer nivell binari i un segon nivell amb un únic classificador general. S'incorpora en la primera fila el resultat obtingut per un classificador SVM com a referència, tal com s'especifica en l'experiment anterior (figura 6.2). Adaptat de (Planet *et al.*, 2009).

Conjunt de dades	Classificador	WAR	UAR
Original	SVM (1 nivell)	65,62	28,94
Original	SVM (2 nivells)	60,78	30,82
Estandarditzat	SVM (2 nivells)	58,71	32,30

Estructura jeràrquica de dos nivells basada en un classificador binari i dos generals. L'estratègia d'aquesta segona estructura jeràrquica de dos nivells és aconseguir conjunts equilibrats de dades en un primer nivell amb els quals entrenar els classificadors de segon nivell mitjançant un mètode natural, sense emprar una tècnica de remostreig com en el primer esquema proposat. Aquesta estructura consisteix, igual que en el cas anterior i tal com es mostra en la figura 6.5, en un classificador de primer nivell (1C) en el qual les locucions són classificades de forma binària segons la seva pertinença a la classe N o a una altra diferent (¬N). A diferència de l'estructura anterior, dos classificadors generals són entrenats en el segon nivell emprant el conjunt de dades d'entrenament després d'haver estat classificat pel classificador binari de primer nivell. El primer classificador (2C-1) és entrenat amb aquells casos de la classe N que han estat correctament classificats pel classificador de primer nivell (veritables positius) i amb els casos que, sense pertànyer a la classe N, han estat classificats per aquest mateix classificador com a tals (falsos positius). El segon classificador (2C-2) és entrenat amb aquells casos de la classe N classificats incorrectament com no pertanyents a aquesta classe (¬N) (falsos negatius) i amb els casos no pertanyents a la classe N que han estat correctament classificats com a tals (veritables negatius).

La taula 6.2 mostra els resultats d'aquesta segona estructura jeràrquica de dos nivells, considerant diferents classificadors en les etapes de les quals consta l'estructura. Els classificadors SVM són els mateixos que els descrits en els experiments anteriors si bé s'introdueix un classificador SVM de kernel gaussià, el qual millora els resultats en termes de WAR però els empitjora en termes de UAR. El millor resultat s'obté amb la configuració de dos classificadors SVM i un classificador Naïve-Bayes, per als casos classificats com a ¬N, que millora el resultat de l'experiment anterior però es manté per darrere de l'obtingut per un únic classificador Naïve-Bayes en el primer experiment.

Estructura de classificació en cascada L'estructura de classificació en cascada proposada apareix representada en la figura 6.6. En cadascun dels 5 nivells de l'estructura es classifiquen les locucions segons pertanyin o no a una classe concreta. L'ordre de les classes

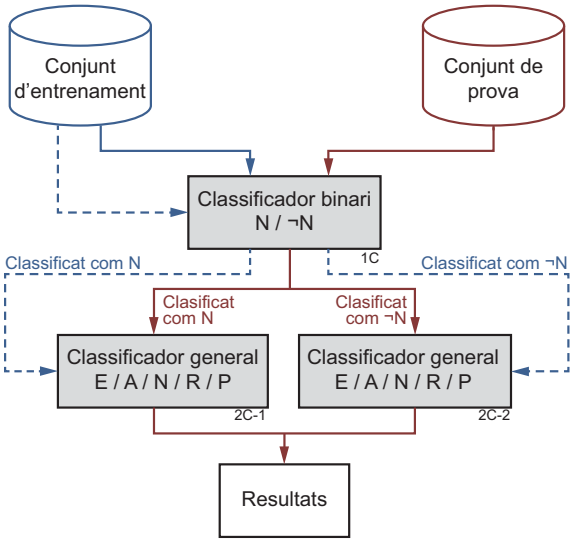


Figura 6.5: Esquema del classificador jeràrquic de dos nivells basat en un classificador binari de primer nivell (1C) i dos classificadors generals de segon nivell (2C-1 i 2C-2). Els elements acolorits en blau indiquen relació amb la fase d'entrenament i els acolorits en vermell indiquen relació amb la fase de prova. Les línies de punts indiquen fluxos de dades que serveixen per entrenar un classificador mentre que les línies sòlides indiquen fluxos de dades que són avaluades en un classificador, independentment de si s'aquests fluxos intervenen en la fase d'entrenament o prova.

Taula 6.2: Resultats per a l'avaluació mitjançant dos conjunts independents de l'estructura jeràrquica basada en dos nivells: un primer nivell binari i un segon nivell amb dos classificadors generals. 1C indica l'algorisme de classificació emprat en el primer nivell. 2C-1 i 2C-2 indiquen els algorismes emprats en el primer i segon classificador, respectivament, del segon nivell. Adaptat de (Planet *et al.*, 2009).

1C	2C-1	2C-2	WAR	UAR
SVM lineal	SVM lineal	SVM lineal	63,04	30,20
SVM gaussià	SVM gaussià	SVM gaussià	64,28	27,97
SVM lineal	SVM lineal	Naïve-Bayes	60,20	32,63

s'estableix en funció de la població de les mateixes (les majoritàries primer i les minoritàries al final) amb la finalitat de descartar casos en cada nivell i crear conjunts de dades més equilibrats per entrenar el següent classificador. Així doncs, el primer classificador és entrenat amb el conjunt de dades d'entrenament complet i la seva finalitat és distingir els casos de la classe N de la resta. És a dir, el primer classificador és binari i separa les locucions segons pertanyin a la classe N o a la resta (¬N). El segon classificador, en un nivell inferior, és entrenat per dur a terme la mateixa funció però analitzant la segona classe més poblada del corpus FAU Aibo, és a dir, distingint els casos de la classe E de la resta. Els casos classificats com a ¬N pel primer classificador són usats per entrenar el segon classificador. L'estructura és repetida fins al cinquè classificador, que analitza la classe menys poblada del corpus, la classe P. En aquest cas, i per ser l'últim, els casos identificats com a ¬P són assignats, per defecte, a la classe N atès que és la classe majoritària.

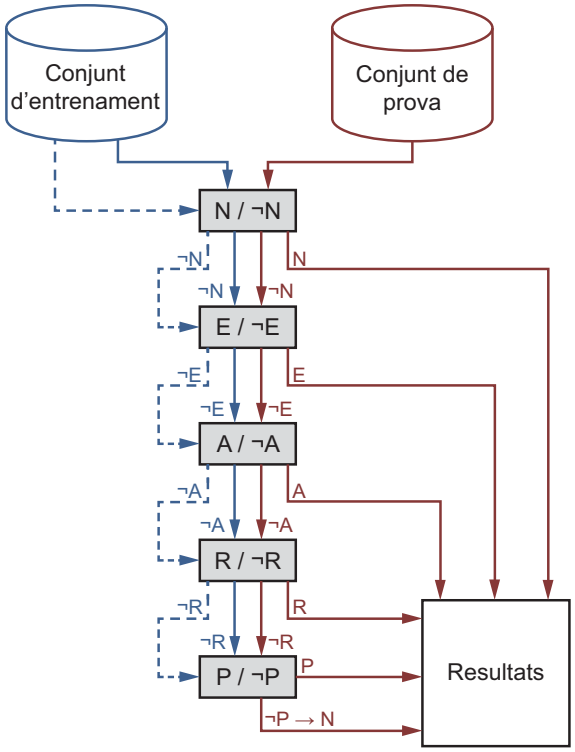


Figura 6.6: Esquema del classificador en cascada basat en 5 nivells de classificació ordenats segons la població de cadascuna de les classes del corpus FAU Aibo. Els elements que es mostren acolorits en blau indiquen relació amb la fase d'entrenament i els acolorits en vermell indiquen relació amb la fase de prova. Les línies de punts indiquen fluxos de dades que serveixen per entrenar un classificador mentre que les línies sòlides indiquen fluxos de dades que són avaluats en un classificador, independentment de si s'aquests fluxos intervenen en la fase d'entrenament o en la de prova.

Emprant classificadors SVM de kernel lineal en cada nivell, la WAR és de 65,53% i la UAR de 28,03%. Atès que la classe R és la que presenta una definició més difusa i és, per tant, més difícil de caracteritzar, es planteja el canvi d'ordre dels nivells R i P. Amb aquesta modificació la WAR és de 65,33% i la UAR de 28,12%. En comparació dels experiments anteriors, aquest plantejament millora la WAR però empitjora la UAR. No obstant això la fase d'entrenament és més ràpida perquè la grandària dels conjunts de dades en cada nivell és inferior al dels altres experiments.

6.2.1.2 Desafiament de paràmetres

Pel que fa a la selecció de paràmetres plantejada a l'inici d'aquesta secció, en primer lloc es completa la parametrització original del corpus. Per completar aquesta parametrització s'afegeix una sèrie de paràmetres relatius a la VoQ que han demostrat ser útils en estudis anteriors (Iriondo *et al.*, 2007b, 2009). Aquests paràmetres, la descripció dels quals es pot consultar en la secció 3.3.1.1, són els següents descriptors de baix nivell: *jitter*, *shim-*

mer, HammI, SFM, do1000 i pe1000. Aquesta parametrització es duu a terme únicament en les parts sonores de les locucions i els funcionals extrets són els mateixos que els de la resta de descriptors de baix nivell originals del corpus FAU Aibo (vegeu la secció 3.4.3.2), amb l'excepció del segon quartil. Addicionalment s'afegeixen dos paràmetres relacionats amb el ritme: durada dels grups accentuals i durada de les síl·labes.

La durada dels grups accentuals s'estima mitjançant la distància entre màxims locals del contorn freqüencial, tal com es defineix en (Planet *et al.*, 2009). Aquest contorn s'obté restant un contorn de base al contorn original. El contorn de base consisteix en el contorn original filtrat passa-baixes amb una freqüència de tall de 5 Hz[5]. La durada de les síl·labes s'estima mitjançant la distància entre pics d'intensitat del contorn d'energia del senyal (Planet *et al.*, 2009). Tots dos contorns s'obtenen mitjançant l'eina d'anàlisi Praat. Per a cadascun d'aquests paràmetres es calculen dos funcionals: mitjana i desviació estàndard. Considerant els paràmetres de VoQ anteriors, en total s'afegeixen 70 paràmetres al conjunt de dades original. El conjunt resultant final consta de 454 paràmetres per locució, excloent l'etiqueta emocional.

Per a aquest experiment preliminar s'escull el mètode de filtre de selecció de paràmetres mRMR (vegeu la secció 4.3.2) emprant la implementació en MATLAB de Peng *et al.* (2005). El conjunt de dades d'entrenament és prèviament processat en dues etapes. En primer lloc, donat el gran volum de dades, el conjunt de dades es redueix a la meitat mantenint la mateixa distribució d'etiquetes afectives. A continuació els paràmetres són discretitzats mitjançant un procés descrit per (Peng *et al.*, 2005) pel qual primer són estandarditzats, després multiplicats per un factor *k* i, finalment, arrodonits a l'enter més proper, tal com es mostra en l'equació 6.1. En aquest cas s'escull el valor de $k = 10$ de forma heurística.

$$param_i = round(k \times z\text{-}score(param_i)) \tag{6.1}$$

L'algorisme es configura per obtenir 200 paràmetres rellevants però tan sols s'escullen els 100 primers d'acord amb les especificacions establertes per Schuller *et al.* (2009) per provar el seu rendiment sobre el conjunt de prova. La figura 6.7 mostra una simulació del rendiment dels conjunts d'entre 1 i 200 paràmetres seleccionats per l'algorisme mRMR a través de l'avaluació amb un classificador SVM de kernel lineal mitjançant una validació creuada de 10 iteracions realitzada sobre el conjunt de dades d'entrenament. El conjunt de dades de prova no intervé en el procés de selecció ni s'empra en aquesta simulació. Amb 100 paràmetres el conjunt de dades no aconsegueix el millor rendiment en la validació creuada. No obstant això, fora de l'escenari de la simulació representada en la figura anterior, un classificador SVM de kernel lineal entrenat amb aquest subconjunt de 100 paràmetres i avaluat amb el conjunt de dades de prova aconsegueix una WAR de

[5]Aquest mètode pretén ser independent de la transcripció fonètica en treballar directament amb la corba de F0. L'eliminació de la variació lenta de F0 fa que els màxims de la corba resultant coincideixin amb els accents i que es pugui estimar, per tant, el temps que transcorre entre els mateixos. Aquest mètode es va validar verificant un conjunt reduït del corpus.

64,95% i una UAR de 21,32%, la qual cosa suposa una degradació absoluta del -0,67% i del -7,62%, respectivament, respecte al resultat obtingut pel mateix classificador amb el conjunt de dades original (vegeu la figura 6.2), essent aquest un 284% més gran.

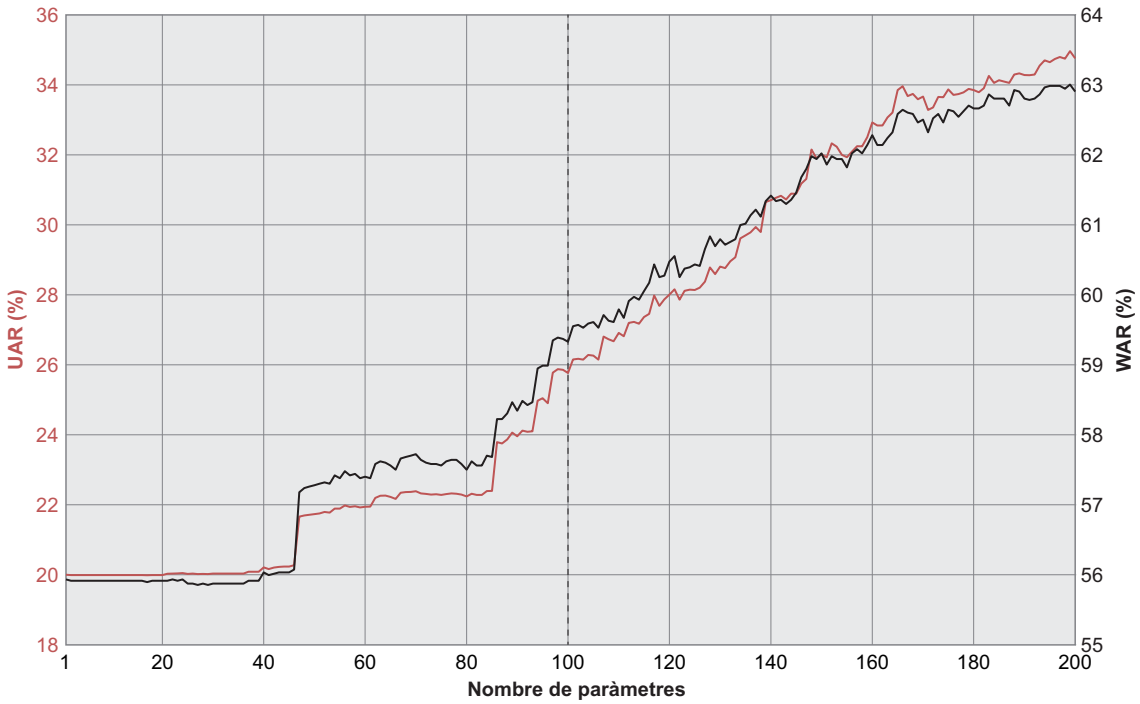


Figura 6.7: Taxes de classificació ponderada (WAR, en negre) i no ponderada (UAR, en vermell), de l'algorisme SVM avaluat mitjançant una validació creuada de 10 iteracions, considerant un conjunt de dades d'entre 1 i 200 paràmetres prèviament escollits pel criteri mRMR. Adaptat de (Planet *et al.*, 2009).

6.2.1.3 Discussió

Els resultats preliminars obtinguts després de la primera anàlisi del corpus FAU Aibo mostren la major complexitat present en l'estudi d'un corpus espontani enfront de la d'un corpus actuat. Es comprova la dificultat per obtenir unes taxes de classificació elevades tant ponderades com no ponderades i, generalment, la millora d'una mesura suposa la degradació de l'altra. A més, es pot comprovar com el fet de treballar en un entorn de diversos locutors suposa una dificultat afegida, la qual cosa s'observa en els resultats obtinguts en entrenar i provar els sistemes de classificació emprant conjunts de dades procedents de locutors diferents. Aquesta prova proporciona resultats per sota dels obtinguts en la validació creuada realitzada en el conjunt de dades d'entrenament, atès que en aquest últim mètode s'empren dades d'un mateix usuari (a més dels d'uns altres) tant per entrenar com per avaluar. No obstant això, aquest efecte de degradació no és tan notable en el cas de l'algorisme Naïve-Bayes, el qual obté els millors resultats, en termes de UAR, que en altres esquemes de classificació més complexos.

D'altra banda, l'algorisme mRMR emprat per realitzar una selecció automàtica de paràmetres rellevants proporciona un conjunt heterogeni i, en conseqüència, de difícil interpretació. Per aquest motiu, al final d'aquest capítol (vegeu la secció 6.4.2.2) es plantejarà un procés de selecció manual que permeti interpretar adequadament els paràmetres escollits.

6.2.2 Sistema de classificació acústica millorat

Després de comprovar que els algorismes Naïve-Bayes i SVM són els que presenten un millor rendiment en la tasca de classificació del corpus FAU Aibo, en aquesta secció són els classificadors emprats per dur a terme un reconeixement basat en paràmetres acústics analitzant un conjunt de dades actualitzat. En aquest sentit, no s'empra la parametrització utilitzada en la secció anterior sinó que es realitza una rèplica de la mateixa mitjançant una versió més recent del programari openSmile. Si bé la definició del corpus segueix sent la mateixa quan al nombre i la definició dels paràmetres, el càlcul dels mateixos és més precís en aquesta versió[6].

Es realitzen dos tipus de classificacions: una categòrica i una altra dimensional, amb l'objectiu de millorar els resultats previs. La classificació categòrica no es basa en cap model emocional i planteja l'exercici de reconeixement afectiu com un cas de classificació de les locucions en classes concretes que, en aquest cas, són etiquetes afectives. La classificació dimensional sí es basa en els models emocionals d'activació-avaluació per realitzar una classificació específica per a estats afectius.

D'altra banda, i per evitar que durant la fase de prova els esquemes siguin entrenats i provats amb dades d'un mateix locutor, es prescindeix del mètode d'avaluació de validació creuada per nombre d'iteracions i se substitueix pel mètode LOSO.

Finalment, i de forma anàloga a com es va realitzar en la secció anterior, es realitza una selecció de paràmetres acústics però, en aquesta ocasió, sense les limitacions establertes en el desafiament de paràmetres abans exposat.

6.2.2.1 Classificació categòrica

En aquest experiment de classificació acústica s'analitzen 4 esquemes de classificació basats en els dos classificadors Naïve-Bayes i SVM, abans esmentats. Aquests esquemes són els següents: un classificador Naïve-Bayes discretitzant de forma supervisada el conjunt de dades (esquema anàleg al de la secció anterior), un classificador SVM de kernel lineal després de la normalització del conjunt de dades (SVM-A), un classificador SVM de kernel lineal després de la normalització i remostreig del conjunt de dades creant

[6]La nova versió del programari openSmile corregeix errors de la versió prèvia i modifica la representació numèrica al format exponencial, la qual cosa suposa una major precisió.

una distribució uniforme de les etiquetes emocionals[7] (SVM-B) i, finalment, un esquema que fusiona a nivell de decisió, mitjançant una estratègia de *stacking* emprant un arbre de decisió J4.8 com a classificador de nivell 1 (vegeu la secció 4.3.1), els resultats dels dos esquemes que presenten un millor rendiment. Tal com s'observa en la figura 6.8 aquests són el primer i tercer esquema aquí descrits (NB+SVM-B). Aquests dos esquemes actuen com a classificadors de nivell 0. Aquest esquema de fusió es descriu detalladament a continuació.

A diferència de la secció anterior, en aquest cas el mètode d'avaluació emprat en la primera fase de l'experiment segueix una estratègia LOSO realitzada sobre el subconjunt d'entrenament. En la segona fase s'analitzen els subconjunts d'entrenament i de prova de forma independent. En aquest cas, l'estratègia LOSO consta de 26 iteracions ja que el subconjunt de dades d'entrenament conté informació de 26 locutors.

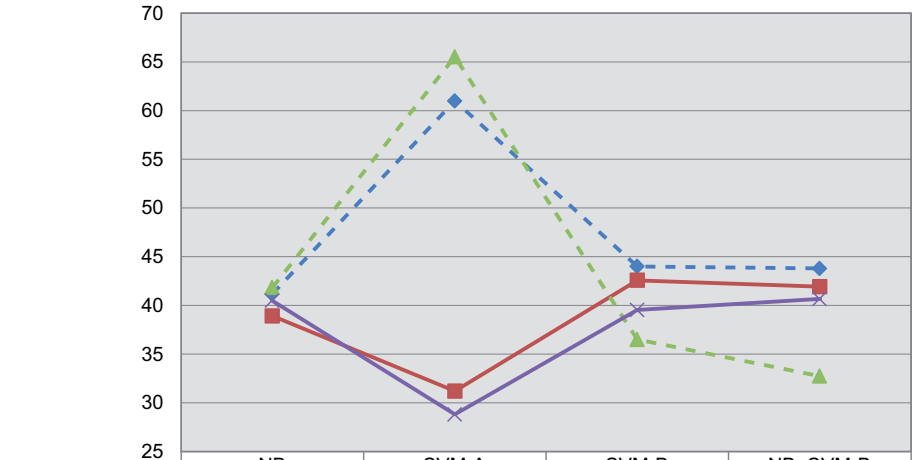
Per a la fusió dels classificadors mitjançant l'esquema *stacking* abans introduït s'empren les decisions categòriques[8] de cada classificador atès que el rendiment d'aquest tipus de decisions acostuma a ser elevat i, al mateix temps, són decisions que aporten un volum petit de dades a l'esquema de fusió (Gabrys i Ruta, 2006). Aquest tipus de fusió mostra el seu bon rendiment en treballs previs de reconeixement afectiu (Morrison *et al.*, 2007; Iriondo *et al.*, 2009).

Un esquema de fusió *stacking* com l'aquí plantejat requereix d'un disseny específic per a la seva avaluació. En la fase d'entrenament —tant en la validació LOSO com en l'avaluació mitjançant els dos conjunts independents—, els esquemes de classificació de nivell 0, Naïve-Bayes i SVM-B, són entrenats amb les dades d'entrenament corresponents (convenientment preprocessades d'acord a la definició de cada esquema). A continuació, aquestes mateixes dades d'entrenament són avaluades per aquests dos algorismes. A partir les decisions categòriques obtingudes es crea un conjunt de dades format per les sortides de tots dos classificadors i l'etiqueta afectiva real de cada cas. Aquest conjunt de dades servirà per entrenar l'algorisme de classificació de nivell 1, l'arbre J4.8, però abans és remostrejat per aconseguir una distribució uniforme de classes. L'objectiu és intentar maximitzar la UAR atès que l'algorisme J4.8 no està especialment dissenyat per obtenir un rendiment similar per classe sinó per maximitzar la WAR. Noti's que el remostreig només afecta a les dades d'entrenament i mai a les de prova. En la fase de prova, les dades de prova són avaluades pels dos classificadors de nivell 0. Les decisions categòriques obtingudes són després avaluades pel classificador de nivell 1 que assigna una etiqueta afectiva final a cada locució. En el cas de la validació creuada aquest procés es realitza per a cada iteració. En el cas de l'anàlisi dels dos conjunts independents aquest procés només es realitza una vegada.

[7] Només de les dades emprades per entrenar el classificador.

[8] Les decisions categòriques (o "dures", de l'anglès *hard decisions*) són la resposta d'un classificador en forma d'atribut discret, en aquest cas en forma de l'etiqueta afectiva que el classificador assigna a cada cas concret. Les decisions "toves" (de l'anglès, *soft decisions*) fan referència a la resposta d'un classificador en forma de grau de pertinença de cada cas concret a cada classe afectiva.

La figura 6.8 mostra els resultats dels esquemes de classificació proposats en termes de taxes de classificació ponderades i no ponderades.



	NB	SVM-A	SVM-B	NB+SVM-B
LOSO WAR	41,18	60,99	44,00	43,80
LOSO UAR	38,92	31,20	42,58	41,92
Test set WAR	41,83	65,53	36,50	32,74
Test set UAR	40,48	28,81	39,53	40,67

Figura 6.8: Taxes de classificació ponderada (WAR) i no ponderada (UAR), en percentatge, dels classificadors segons la seva avaluació mitjançant una estratègia LOSO i mitjançant els dos conjunts independents (Test set).

Pel que fa als resultats obtinguts pel sistema LOSO, el classificador SVM-B obté el millor resultat en termes de UAR. Comparat amb l'algorisme Naïve-Bayes, que obté el segon millor resultat i era el millor esquema de classificació de l'experiment anterior, obté una millora absoluta del 3,66% (9,40% relativa).

En comparació de l'esquema SVM-A, el qual no considera el remostreig de les dades d'entrenament, l'algorisme SVM-B obté una millora absoluta de l'11,38% (36,47% relativa). No obstant això, en termes de WAR, el resultat de l'esquema SVM-A és un 16,99% millor que el resultat de l'esquema SVM-B. Això il·lustra que l'algorisme SVM està orientat a maximitzar el resultat global de classificació, encara que en aquest cas l'interès se centri a obtenir la millor classificació per totes i cadascuna de les etiquetes emocionals, per la qual cosa la creació de conjunts equilibrats de dades afavoreix la classificació de les classes individuals.

Els esquemes Naïve-Bayes i SVM-B es fusionen en l'esquema NB+SVM-B, atès que són els que millors resultats obtenen. En termes de UAR, l'esquema de fusió NB+SVM-B obté una millora absoluta del 3% (7,71% relativa) enfront de l'algorisme Naïve-Bayes per si només. No obstant això, el resultat està lleugerament per sota de l'obtingut per l'algorisme SVM-B.

Quant als resultats obtinguts avaluant els conjunts independents d'entrenament i de prova, l'esquema de fusió NB+SVM-B és el que obté el millor resultat en termes de

UAR, millorant els resultats obtinguts de forma independent pels dos esquemes que ho formen. Pel que fa a l'esquema Naïve-Bayes la millora és 0,19% absoluta (0,47% relativa) i pel que fa a l'esquema SVM-B la millora és de l'1,14% absoluta (2,88% relativa). Comparant el resultat d'aquest algorisme amb el proporcionat per Schuller *et al.* (2009) com a referència (38,20%), la millora és 2,47% absoluta (6,47% relativa).

L'esquema Naïve-Bayes és l'únic que millora el seu valor de UAR en el cas de l'avaluació amb els subconjunts de dades independents enfront de la validació creuada. La resta d'esquemes obtenen un resultat per sota de l'obtingut en la validació creuada. Aquesta degradació és més notable en els esquemes SVM-A i SVM-B (-2,39% i -3,05%, respectivament) que en l'esquema de fusió NB+SVM-B (-1,25%). Això posa de manifest que l'esquema Naïve-Bayes és més generalizable, doncs suporta el canvi de dades en la fase de prova després d'haver estat entrenat amb altres diferents.

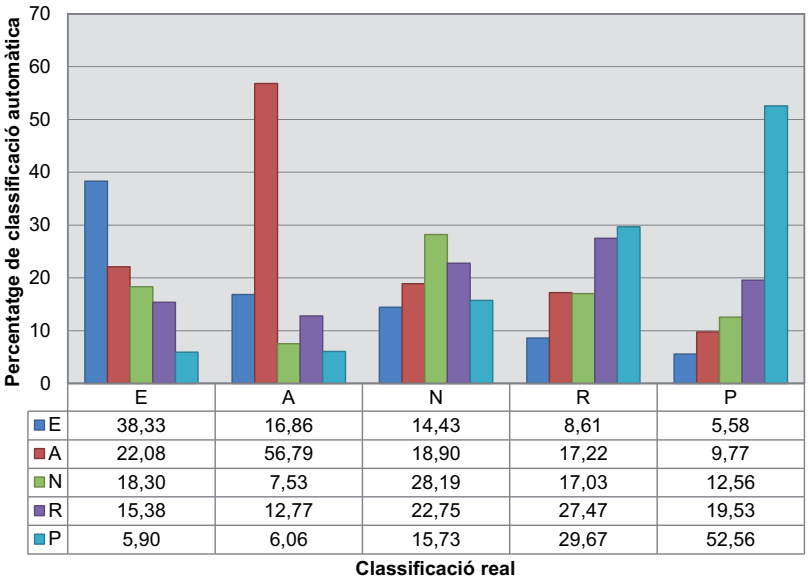
Novament, i tal com succeïa en l'experiment anterior, l'esquema Naïve-Bayes és el que presenta una menor variació entre els resultats de WAR i UAR. No obstant això, el fet de remostrejar les dades d'entrenament a una distribució uniforme de classes permet que aquesta diferència no sigui molt elevada en altres esquemes. Així, és notable la diferència entre resultats WAR i UAR de l'esquema SVM-A mentre que aquesta es veu reduïda en l'esquema SVM-B.

La figura 6.9 mostra la matriu de confusió de l'esquema de fusió NB+SVM-B avaluat mitjançant els dos conjunts de dades independents. Es pot observar que els casos de la classe R són els pitjor classificats, si bé el resultat és millor que la classificació de l'experiment anterior (vegeu la figura 6.2). Novament es comprova la dificultat de classificar aquesta classe, associada a un valor baix d'afinitat prototípica, donada la seva definició poc exacta.

6.2.2.2 Classificació dimensional

Els esquemes proposats en els experiments anteriors no es basen en cap model emocional. Per completar l'anàlisi acústica del corpus FAU Aibo en aquesta secció s'exposa un plantejament basat en el model d'activació-avaluació explicat en la secció 2.1.3.3. La figura 6.10 mostra el mapatge de les etiquetes afectives del corpus FAU Aibo (vegeu la secció 3.4.3.1) en l'espai d'activació-avaluació descrit per Scherer (1984), segons les definicions de Steidl (2009) i Schuller *et al.* (2009) de cadascuna d'elles. El mapatge en un espai bidimensional permet la proposta de dos classificadors jeràrquics, un per cada dimensió.

A continuació s'explica, en primer lloc, el disseny d'aquests dos classificadors jeràrquics per separat i, en segon lloc, s'explica el procés de fusió i es detallen els resultats obtinguts.



	E	A	N	R	P
E	38,33	16,86	14,43	8,61	5,58
A	22,08	56,79	18,90	17,22	9,77
N	18,30	7,53	28,19	17,03	12,56
R	15,38	12,77	22,75	27,47	19,53
P	5,90	6,06	15,73	29,67	52,56

Figura 6.9: Matriu de confusió de l'esquema de fusió NB+SVM-B avaluat mitjançant els dos conjunts independents d'entrenament i de prova. Adaptat de (Planet *et al.*, 2009).

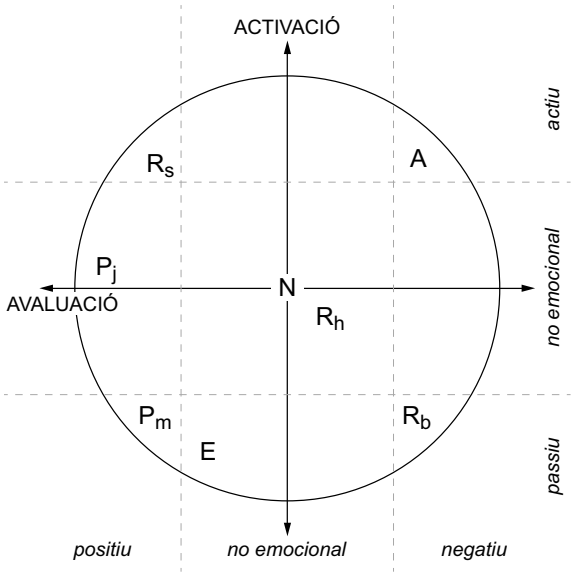


Figura 6.10: Mapatge en el model d'activació-avaluació de Scherer (1984) de les etiquetes afectives del corpus FAU Aibo, segons les definicions de Steidl (2009) i Schuller *et al.* (2009) de cadascuna d'elles. Les etiquetes afectives són neutre (N), enuig (A), emfàtic (E), positiu (P) (format per maternal (P_m) i alegre (P_j)) i resta (R) (format per sorpresa (R_s), impotència (R_h) i avorriment (R_b)).

Classificació sobre la base de la dimensió d'avaluació. La figura 6.11 mostra la divisió teòrica de les classes afectives del corpus FAU Aibo sobre la base de, en primer lloc, la

seva prominència emocional (emocional o no emocional)[9] i, en segon lloc, la seva avaluació (positiva o negativa), segons el mapatge il·lustrat en la figura 6.10. D'acord amb la definició de les etiquetes afectives de Schuller *et al.* (2009), els estats maternal (P_m) i alegre (P_j) estan coberts per la classe P, la qual es considera d'avaluació positiva, la classe A es considera d'avaluació negativa i les classes E i N es consideren sense prominència emocional. D'altra banda, atès que la classe R està formada per 3 estats afectius diferents (sorpresa (R_s), avorriment (R_b) i impotència (R_h)), es consideren 3 subclasses per realitzar el mapatge de forma adequada assignant 3 valors d'avaluació diferents (avaluació positiva, negativa i sense prominència emocional, respectivament). No obstant això, aquestes subclasses han de ser agrupades en una etiqueta única (R) durant la fase de prova dels classificadors atès que aquesta distinció no existeix en l'etiquetatge original del corpus.

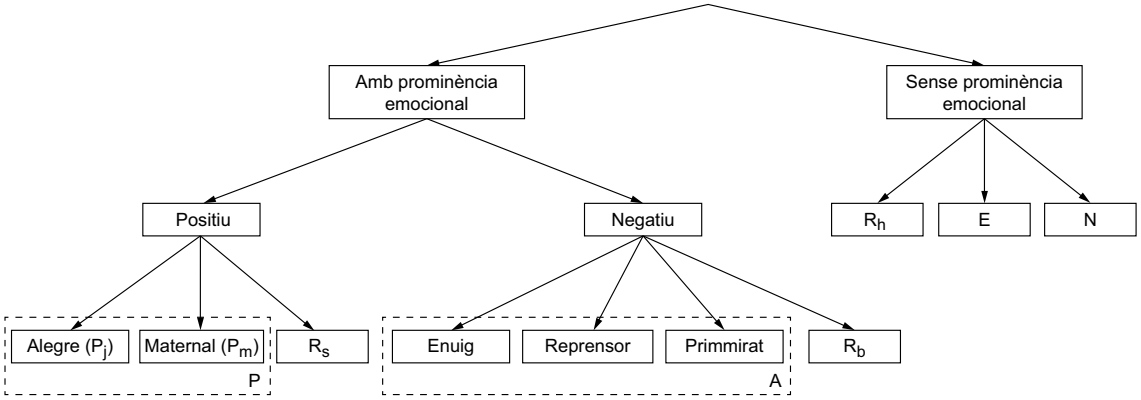


Figura 6.11: Classificació de les etiquetes afectives del corpus FAU Aibo basada en la dimensió d'avaluació.

La figura 6.12 mostra el classificador jeràrquic proposat sobre la base de la divisió de les etiquetes afectives realitzada a dalt. L'estructura consta de 5 classificadors especialitzats, denominats microclassificadors i representats pel símbol μC_i on $1 \leq i \leq 5$. El microclassificador μC_1 s'entrena per distingir les classes E, N i R (sense prominència emocional) de les classes P, A i R (amb prominència emocional), d'acord amb la primera divisió de l'arbre de classificació de la figura 6.11. El microclassificador μC_2 s'entrena per distingir les 3 etiquetes sense prominència emocional E, N i R, mentre que el microclassificador μC_3 es dissenya per classificar els casos d'acord al seu valor d'avaluació: positiva (classes P i R) o negativa (classes A i R). Els microclassificadors μC_4 i μC_5 tenen l'objectiu de realitzar la darrera classificació per separar les classes P, A i R.

Per implementar aquesta estructura s'analitzen els dos algorismes que han obtingut els millors resultats en els experiments anteriors: Naïve-Bayes i SVM. A partir dels resultats de UAR aconseguits per aquests algorismes en cadascuna de les parts de l'estructura jeràrquica s'escullen els millors per implementar els diferents microclassificadors. Així, els algorismes són: Naïve-Bayes pels microclassificadors μC_1 i μC_2, i SVM

[9]No ha de confondre's el concepte de prominència emocional aquí emprat amb l'explicat en la secció 3.3.2.1, que es referia a la parametrització lingüística dels elements de les frases d'un corpus. En aquest cas es fa referència a si la categoria afectiva és emocionalment representativa de la dimensió afectiva analitzada.

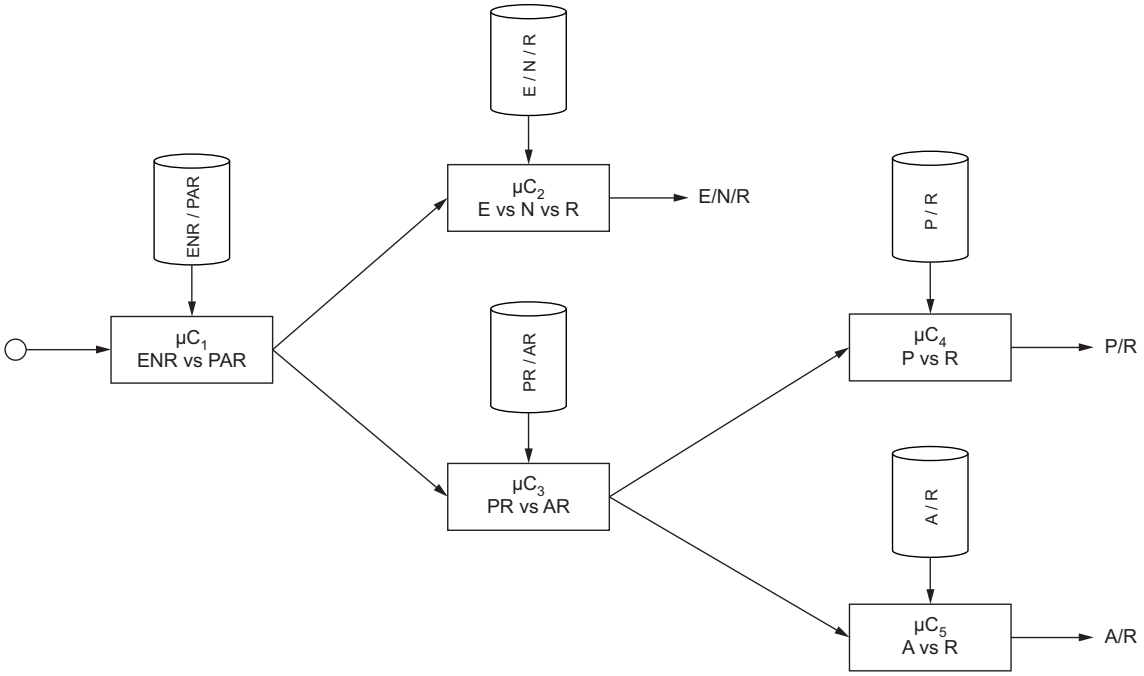


Figura 6.12: Disseny del classificador jeràrquic basat en la dimensió d'avaluació.

pels microclassificadors μC_3, μC_4 i μC_5.

Classificació sobre la base de la dimensió d'activació. La figura 6.13 mostra la divisió teòrica de les classes afectives del corpus FAU Aibo sobre la base de la seva prominència emocional en primer lloc i, en segon lloc, sobre la base de la seva activació (actives o passives). En aquest cas, les classes P_m i P_j estan en branques diferents de l'arbre de classificació atès que la seva dimensió d'activació és diferent. A més, R_b, R_s i R_h apareixen una altra vegada en 3 grups diferents, tal com ja succeïa en l'anàlisi a través de la dimensió d'avaluació. No obstant això, i tal com s'ha explicat abans, atès que no existeix diferenciació en l'etiquetatge original del corpus totes elles es consideren agrupades en una etiqueta única.

La figura 6.14 mostra el classificador jeràrquic proposat sobre la base de la divisió de les etiquetes afectives realitzada a dalt. L'estructura consta de 5 microclassificadors (μC_i, $1 \leq i \leq 5$). El microclassificador μC_1 s'entrena per distingir les classes N, P i R (sense prominència emocional) de les classes A, E, P i R (amb prominència emocional), d'acord amb la primera branca de l'arbre de la figura 6.13. El microclassificador μC_2 té l'objectiu de classificar les 3 classes sense prominència emocional (N, P i R), mentre que el microclassificador μC_3 classifica la resta de casos sobre la base de la seva activació: activa (classes A i R) o passiva (E, P i R). Els microclassificadors μC_4 i μC_5 s'encarreguen de l'últim pas de la classificació en les classes A, E, P i R.

En aquest cas els algorismes per implementar cadascun dels microclassificadors,

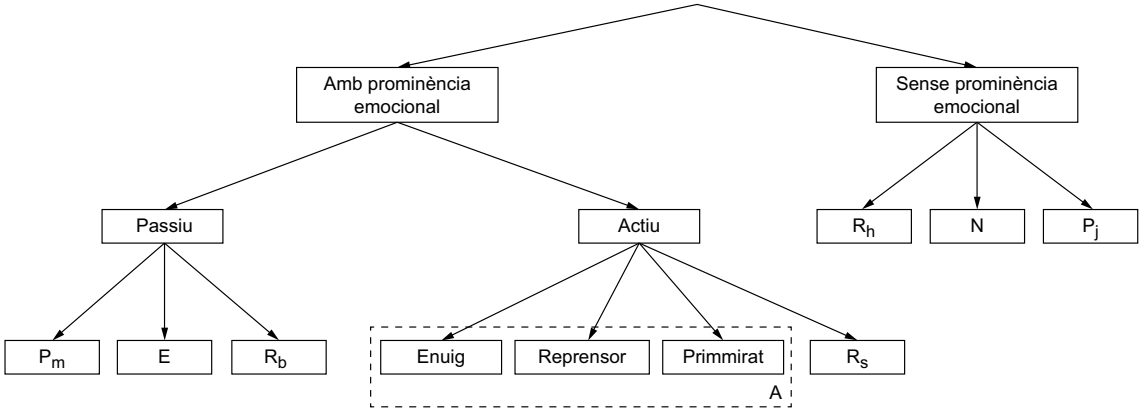


Figura 6.13: Classificació de les etiquetes afectives del corpus FAU Aibo basada en la dimensió d'activació.

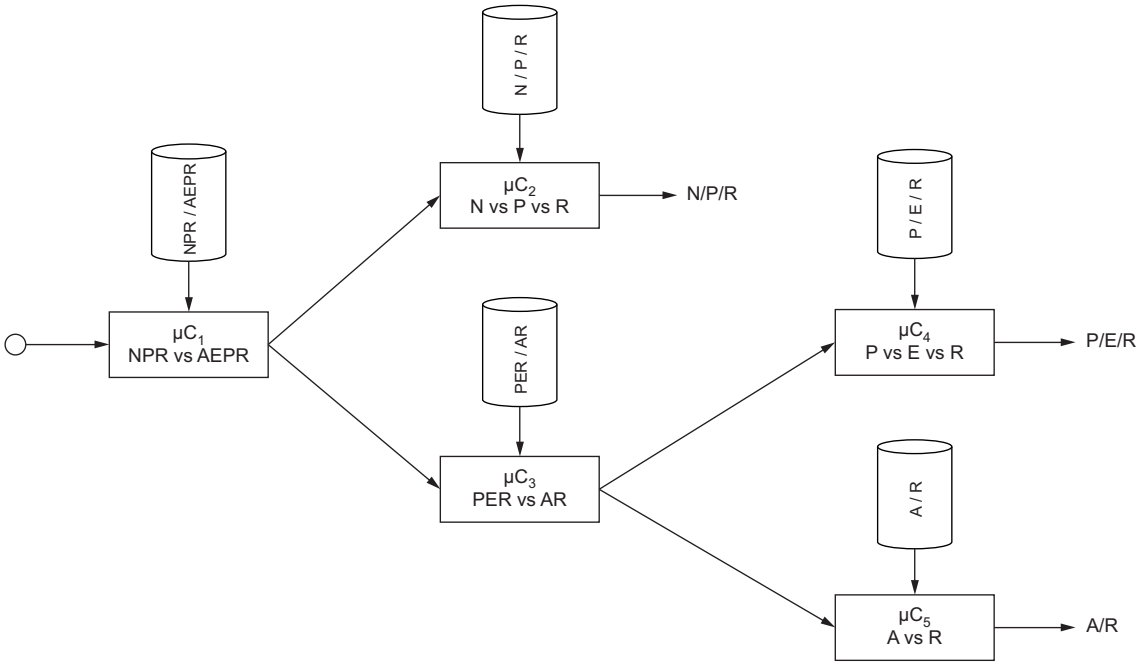
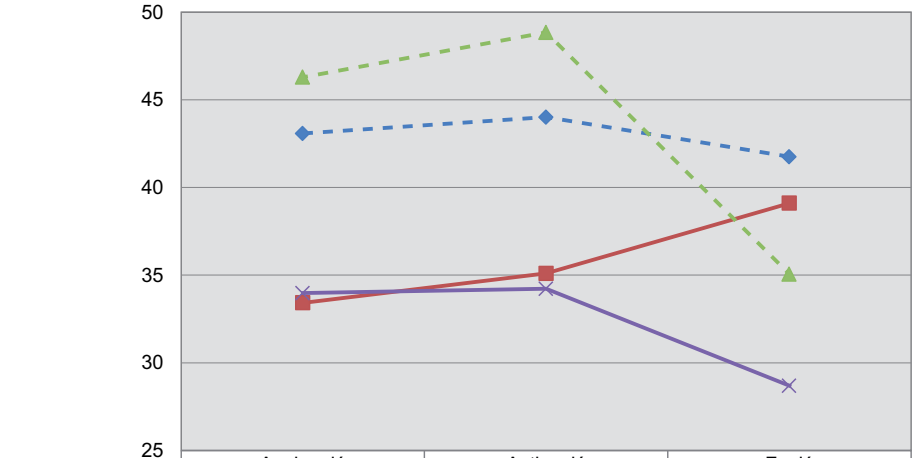


Figura 6.14: Disseny del classificador jeràrquic basat en la dimensió d'activació.

escollits segons el mateix procediment que en el cas anterior, són: Naïve-Bayes pels microclassificadors μC_2 i μC_3, i SVM pels microclassificadors μC_1, μC_4 i μC_5.

Fusió. L'esquema de fusió emprat per combinar els resultats de tots dos classificadors dimensionals és el mateix esquema de *stacking* proposat en la secció 6.2.2.1, mitjançant un arbre de decisió J4.8 com a classificador de nivell 1 i un remostreig previ de les dades d'entrenament.

Els resultats dels classificadors basats en cada dimensió i la fusió de tots dos es



	Avaluació	Activació	Fusió
LOSO WAR	43,08	44,01	41,75
LOSO UAR	33,42	35,10	39,10
Test set WAR	46,29	48,84	35,06
Test set UAR	33,98	34,22	28,69

Figura 6.15: Taxes de classificació ponderada (WAR) i no ponderada (UAR), en percentatge, dels esquemes jeràrquics basats en les dimensions d'avaluació i activació, així com la fusió de tots dos mitjançant una estratègia de *stacking*, segons la seva avaluació mitjançant una validació LOSO i mitjançant els dos conjunts independents (Test set).

recullen en la figura 6.15. Com es pot observar, una vegada més els resultats referents a la mesura WAR se situen per sobre dels de la mesura UAR, si bé la fusió dels classificadors dimensionals apropa tots dos valors. A més, a diferència dels casos anteriors, no existeixen grans diferències entre els resultats obtinguts per la validació LOSO i l'avaluació mitjançant els dos conjunts independents de dades, amb l'excepció de l'esquema de fusió on aquesta diferència és més gran. Quant a l'anàlisi dels resultats obtinguts segons l'avaluació mitjançant els dos conjunts independents de dades, l'esquema de fusió obté uns resultats inferiors als obtinguts de forma individual per cada dimensió. Això es deu, principalment, al fet que l'esquema de fusió es basa únicament en dos paràmetres (els resultats de classificació basats en cadascuna de les dues dimensions emocionals), al fet que aquests paràmetres provenen de classificadors amb baixes taxes de classificació i al fet que, a més, existeix un baix consens entre tots dos classificadors (la qual cosa es relaciona amb el baix rendiment de cadascun d'ells).

6.2.2.3 Reducció del conjunt de paràmetres

Per realitzar una selecció dels paràmetres més representatius del conjunt de dades acústic es realitza una cerca voraç de forma incremental i decremental (vegeu la secció 4.3.2). L'algorisme d'avaluació emprat és el Naïve-Bayes amb discretització supervisada prèvia. Per agilitar el procés, de forma prèvia a l'inici de la selecció de paràmetres el conjunt de dades d'entrenament és remostrejat reduint el volum de dades a la meitat i

creant una distribució uniforme de classes. Així mateix, per garantir la independència de locutors, s'empra només el conjunt de dades d'entrenament per seleccionar i avaluar els subconjunts de paràmetres en les fases de cerca i avaluació.

En el cas de la cerca voraç incremental, el conjunt de dades acústic es redueix de 384 a 28 paràmetres: 21 relacionats amb els paràmetres MFCC, 3 relacionats amb l'energia, 2 relacionats amb la F0, 1 relacionat amb la HNR i 1 relacionat amb la ZCR. La cerca voraç decremental suposa una aproximació més conservadora i crea un conjunt de 305 paràmetres.

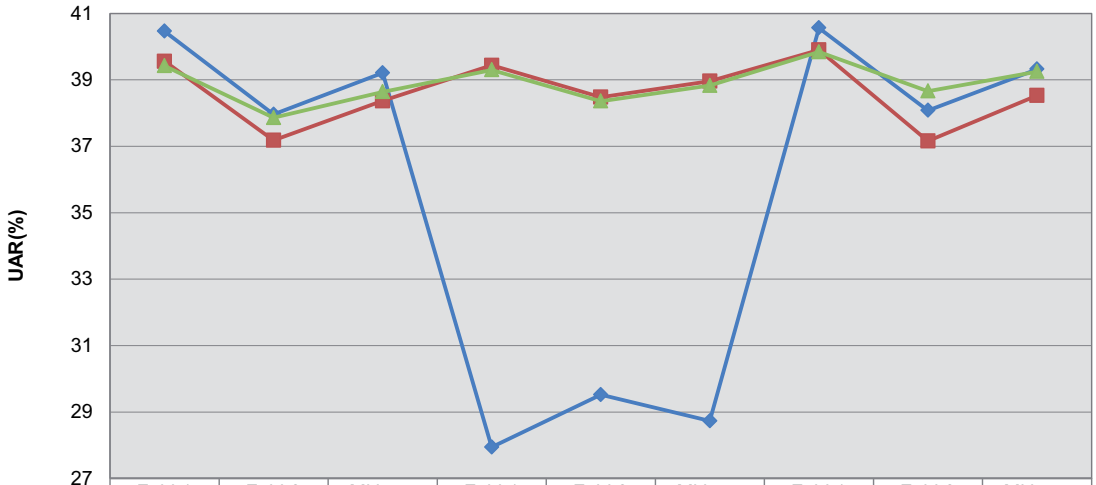
Per observar el rendiment obtingut amb aquests conjunts de dades s'avaluen tres classificadors diferents. Els dos primers es corresponen amb els algorismes Naïve-Bayes i SVM de kernel lineal descrits en les seccions anteriors, als quals s'afegeix un arbre de decisió que empra un model de regressió logística en l'última etapa de classificació, tal com descriuen Landwehr *et al.* (2005). La implementació de WEKA d'aquest algorisme, emprada en aquesta tesi, rep el nom de *Simple Logistic* (SL).

L'avaluació dels algorismes anteriors es realitza mitjançant una validació creuada de dues iteracions en la qual els dos conjunts de dades emprats són els de entrenament i prova ja descrits, emprant-los per a les seves respectives finalitats en la primera iteració (Fold 1)[10] i intercanviant-los en la segona (Fold 2). El valor mitjà d'ambdues avaluacions també és calculat. Per simplificar el text, l'anàlisi de resultats ja no inclou els resultats de validacions creuades realitzades sobre el conjunt de dades d'entrenament així com tampoc els valors de WAR. Com s'ha pogut observar en els resultats anteriors, existeix una estreta relació entre el comportament dels resultats obtinguts mitjançant una avaluació LOSO duta a terme sobre el conjunt d'entrenament i els obtinguts mitjançant l'avaluació dels conjunts independents d'entrenament i prova. Així doncs, s'opta per la validació creuada de dues iteracions abans descrita pels motius anteriors, aprofitant totes les dades del corpus. D'altra banda, això permet superar les limitacions imposades per l'*Emtion Challenge 2009* sense perdre la capacitat de comparació dels resultats obtinguts.

La figura 6.16 mostra els resultats de l'avaluació dels tres classificadors anteriors usant el conjunt de dades acústiques complet i les seves dues versions reduïdes. Observant els valors mitjans de les dues iteracions, els valors de UAR són lleugerament millors en el cas dels conjunts de dades reduïts que en el complet, exceptuant el cas de l'algorisme Naïve-Bayes. En el cas de l'algorisme Naïve-Bayes, el conjunt de dades creat mitjançant la cerca voraç incremental degrada notablement el rendiment del classificador, reduint la taxa de classificació en un -10,48%. Això és a causa de què una de les condicions d'aquest algorisme és la independència dels paràmetres del conjunt de dades a analitzar (Rish, 2001). No obstant això, aquest mètode de selecció no té per objectiu vetllar per aquesta condició pel que el conjunt creat no és idoni per a aquest algorisme. Aquesta degradació no és apreciable en el cas de la cerca decremental atès que el conjunt de dades seleccionat

[10] Aquesta primera iteració es correspon, de fet, amb els experiments *Test set* duts a terme amb anterioritat, en els quals l'entrenament es realitza amb el conjunt de dades Ohm i la prova amb el conjunt de dades Mont (vegeu la secció 3.4.3.1).

és molt més gran. Cal assenyalar també el bon rendiment de l'algorisme SL en l'aplicació de reconeixement afectiu.



	Fold 1	Fold 2	Mitjana	Fold 1	Fold 2	Mitjana	Fold 1	Fold 2	Mitjana
	Conjunt acústic complet (384)			Selecció acústica FW (28)			Selecció acústica BW (305)		
NB	40,46	37,96	39,21	27,94	29,52	28,73	40,56	38,08	39,32
SVM	39,56	37,18	38,37	39,44	38,48	38,96	39,90	37,16	38,53
SL	39,42	37,86	38,64	39,30	38,36	38,83	39,84	38,66	39,25

Figura 6.16: Taxa de classificació no ponderada (UAR), en percentatge, dels classificadors analitzant el conjunt complet de 384 paràmetres acústics i els conjunts reduïts de 28 i 305 paràmetres seleccionats mitjançant una cerca voraç incremental i decremental, respectivament. Adaptat de (Planet i Iriondo, 2012b).

6.2.2.4 Discussió

Quant a l'experiment de classificació categòrica, en aquest segon experiment de reconeixement afectiu basat únicament en paràmetres acústics es comprova, en contrast amb l'anterior, que la diferència entre els mètodes LOSO i l'avaluació amb dos conjunts independents és menor que l'existent entre la validació creuada de diverses iteracions i la dels dos conjunts independents. No obstant això, el comportament de les validacions creuades segueix la tendència observada en l'anàlisi dels conjunts independents.

En aquest cas, l'esquema SVM-B obté el millor valor de UAR, a través d'un conjunt de dades remostrejat cap a una distribució uniforme de classes, quan s'analitza mitjançant una validació creuada. No obstant això el rendiment decreix quan s'empren dos conjunts independents. Aquest efecte no és perceptible en l'esquema Naïve-Bayes, el qual funciona fins i tot millor en analitzar dos subconjunts diferents. La combinació de tots dos esquemes mitjançant l'estratègia de *stacking* proporciona un valor de UAR lleugerament superior als obtinguts per cada esquema de forma individual, sent molt similar a la de l'esquema Naïve-Bayes. No obstant això, observant la classificació per classe a través de la matriu de confusió es comprova que la classificació dels casos de la classe R és millor

en el cas de l'esquema de fusió que en l'esquema Naïve-Bayes.

Quant a l'experiment de classificació dimensional, basada en les dimensions d'activació i avaluació a partir de la localització de les etiquetes afectives en un espai emocional bidimensional, els resultats obtinguts se situen per sota dels experiments anteriors, en termes de UAR. Les taxes de classificació d'ambdues dimensions és similar, si bé és lleugerament superior en el cas de la dimensió d'activació. Ambdues, però, són baixes i la serva fusió proporciona un resultat per sota dels experiments anteriors.

A pesar que els esquemes proposats no aconsegueixen superar el valor de UAR de la secció anterior, el procés de selecció de paràmetres mitjançant una cerca voraç incremental i decremental aconsegueix crear dos conjunts reduïts de dades que mantenen els resultats obtinguts pel conjunt de dades complet. No obstant això, l'esquema Naïve-Bayes veu degradat el seu rendiment amb el conjunt de dades reduït perquè l'algorisme de selecció no garanteix la independència dels paràmetres escollits.

Per aconseguir millorar els resultats de classificació es proposa l'anàlisi de paràmetres lingüístics, atès que les estratègies basades en informació acústica plantejades no suposen millores en els resultats ja obtinguts, tal com es mostra a continuació.

6.3 Reconeixement afectiu basat en l'anàlisi de paràmetres lingüístics

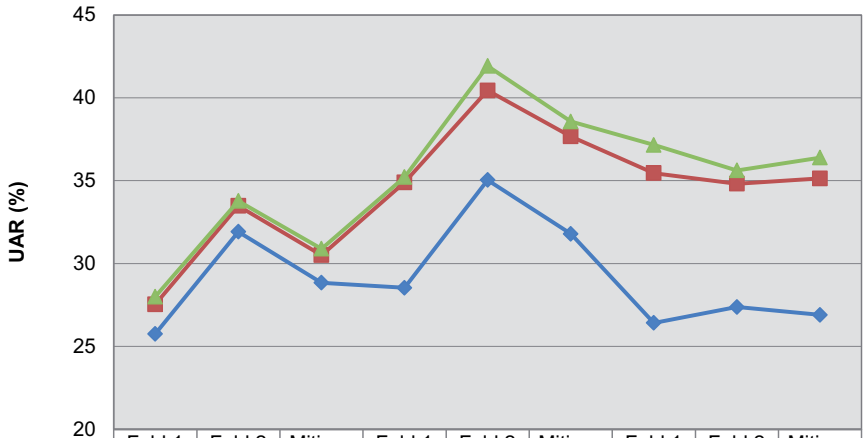
Els resultats de la classificació basada en paràmetres acústics posen de manifest que en un escenari realista, com el plantejat pel corpus FAU Aibo, l'anàlisi exclusiva de la informació acústica pot no ser suficient per dur a terme una tasca de reconeixement automàtic afectiu a partir del senyal de veu, tal com assenyalen Batliner *et al.* (2003). Per aquest motiu s'introdueix l'anàlisi de la modalitat lingüística a partir de la parametrització descrita en la secció 3.4.3.2, basada en l'anàlisi a nivell de paraula de les transcripcions del corpus sobre la base d'unigrames i bigrames i de l'anàlisi de categoria morfològica sobre la base de trigrames a nivell POS.

La finalitat d'aquest apartat és, de forma anàloga a la de l'apartat anterior, comprovar el rendiment dels classificadors quan treballen amb els paràmetres lingüístics de manera independent per posteriorment, en la secció següent, realitzar la seva fusió amb els paràmetres acústics.

6.3.1 Classificació dels subconjunts de paràmetres

La figura 6.17 mostra els resultats UAR dels classificadors Naïve-Bayes, SVM i SL descrits en les seccions anteriors, avaluats mitjançant la validació creuada de dues iteracions abans exposada. Els conjunts de dades es corresponen amb els tres tipus de pa-

ràmetres extrets a nivell de paraula i de categoria morfològica en forma d'unigrames i bigrames a nivell de paraula i trigrames a nivell POS, analitzats de forma independent. Es pot comprovar que la tendència dels tres classificadors és pràcticament la mateixa, si bé l'algorisme Naïve-Bayes és el que presenta un rendiment inferior en tots els conjunts de dades lingüístiques analitzats. Per la seva banda, SVM i SL presenten uns resultats bastant similars.



	Unigrames			Bigrames			Trigrames (POS)		
	Fold 1	Fold 2	Mitjana	Fold 1	Fold 2	Mitjana	Fold 1	Fold 2	Mitjana
NB	25,76	31,92	28,84	28,54	35,04	31,79	26,42	27,38	26,90
SVM	27,54	33,48	30,51	34,90	40,44	37,67	35,46	34,82	35,14
SL	28,00	33,78	30,89	35,22	41,92	38,57	37,16	35,62	36,39

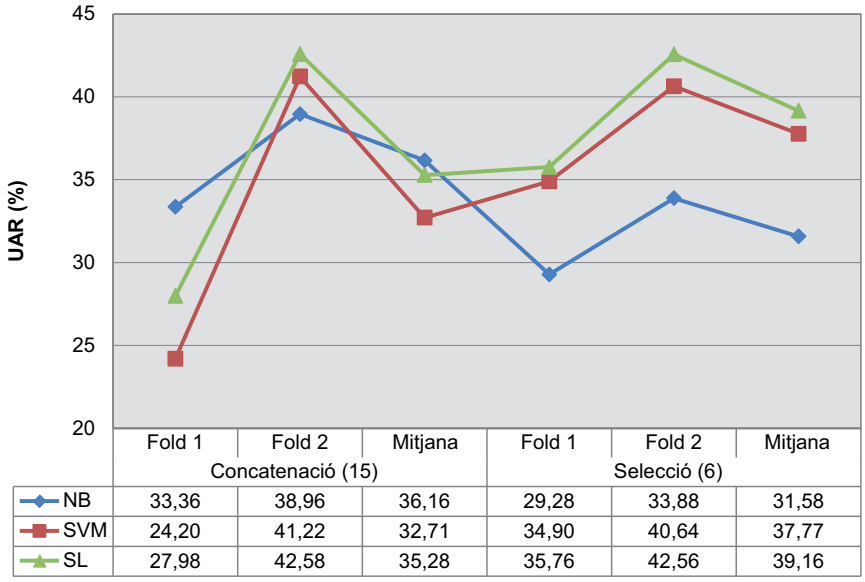
Figura 6.17: Taxa de classificació no ponderada (UAR), en percentatge, dels classificadors analitzant 3 conjunts de dades lingüístiques: els formats pels unigrames i bigrames a nivell de paraula i el format per trigrames a nivell POS.

Atenent als resultats de l'esquema SL i considerant el valor mitjà de les dues iteracions, el rendiment UAR del conjunt de dades format pels bigrames a nivell de paraula (38,57%) és superior a la resta de conjunts. Així, la millora obtinguda amb aquest conjunt de dades i l'algorisme SL és de 2,18% absolut (5,99% relatiu) pel que fa al conjunt de trigrames a nivell POS. En comparació del conjunt d'unigrames a nivell de paraula aquesta millora es d'un 7,68% absolut (24,86% relatiu).

A efectes comparatius amb la resta d'experiments, i seguint el marc de treball plantejat per Schuller *et al.* (2009) i adoptat en aquesta tesi, s'han de considerar els resultats de la columna Fold 1 els quals recullen l'entrenament i la classificació dels conjunts de dades definides a aquest efecte. El millor resultat l'obté l'anàlisi del conjunt de dades format pels trigrames a nivell POS mitjançant l'algorisme SL, amb un valor de UAR del 37,16%. No obstant això, aquest resultat se situa per sota de la major part dels resultats obtinguts mitjançant l'anàlisi acústica realitzada en la secció anterior, així com també és inferior en un 1,04% absolut (2,72% relatiu) al resultat de referència presentat per Schuller *et al.* (2009), obtingut en les mateixes condicions a partir d'una anàlisi acústica.

6.3.2 Fusió i selecció de paràmetres lingüístics

En la secció anterior es realitzava l'anàlisi de cada conjunt de dades lingüístiques de forma independent donada la diferent naturalesa de cadascun d'aquests conjunts: malgrat referir-se a l'àmbit lingüístic cada conjunt provenia de l'anàlisi de diferents unitats (unigrames i bigrames a nivell de paraula i trigrames a nivell POS). En aquesta secció es fusionen els tres conjunts anteriors a nivell de paràmetres mitjançant concatenació creant un conjunt de dades de 15 paràmetres lingüístics. El rendiment d'aquest conjunt de dades pot observar-se en els resultats recollits en la figura 6.18. Naïve-Bayes obté millors resultats que en l'anàlisi individual si bé els algorismes SVM i SL, exceptuant l'anàlisi d'unigrames, tenen millor rendiment amb els conjunts individuals.



	Concatenació (15)			Selecció (6)		
	Fold 1	Fold 2	Mitjana	Fold 1	Fold 2	Mitjana
NB	33,36	38,96	36,16	29,28	33,88	31,58
SVM	24,20	41,22	32,71	34,90	40,64	37,77
SL	27,98	42,58	35,28	35,76	42,56	39,16

Figura 6.18: Taxa de classificació no ponderada (UAR), en percentatge, dels classificadors analitzant dos conjunts de dades lingüístiques: el que fusiona per concatenació els tres conjunts anteriors (15 paràmetres) i la selecció automàtica realitzada per una cerca voraç incremental (6 paràmetres).

D'altra banda, malgrat la reduïda grandària d'aquest conjunt de dades, es realitza una cerca voraç incremental per seleccionar els paràmetres més rellevants. El resultat és un conjunt de 6 paràmetres format per 4 elements relatius als bigrames a nivell de paraula i 2 relatius als trigrames a nivell POS. Com era d'esperar, el conjunt de dades seleccionades no inclou paràmetres relatius als unigrames a nivell de paraula donat el seu baix rendiment (vegeu la figura 6.17). La figura 6.18 recull el rendiment d'aquest conjunt reduït de dades. Naïve-Bayes obté un rendiment baix pels mateixos motius que ja s'exposaven en la secció anterior, relacionats amb la independència dels paràmetres seleccionats. Per contra, SVM i SL milloren el valor mitjà de UAR del conjunt complet, així com també la de la iteració Fold 1.

6.3.3 Discussió

L'anàlisi lingüística duta a terme en aquesta secció ha avaluat tres conjunts de dades des de diferents punts de vista. En primer lloc s'han analitzat de forma independent donada la diferent naturalesa dels mateixos ja que aquests provenen de l'anàlisi d'unigrames i bigrames a nivell de paraula i de trigrames a nivell POS. També s'han fusionat els tres conjunts creant un únic conjunt mitjançant la concatenació dels mateixos. I finalment, de forma anàloga al procés seguit en l'estudi acústic, s'ha creat un conjunt reduït mitjançant una cerca voraç incremental.

Bigrames a nivell de paraula i trigrames a nivell POS s'han revelat com els paràmetres més rellevants per a l'experiment de reconeixement afectiu automàtic aquí plantejat, tant per l'anàlisi individual com per l'algorisme de selecció automàtica.

No obstant això, el rendiment dels paràmetres lingüístics no és prou elevat com per superar el resultat de referència de Schuller *et al.* (2009). Així doncs, el millor resultat individual procedent de l'anàlisi dels subconjunts d'entrenament i prova de forma independent és un 1,04% absolut (2,72% relatiu) inferior a aquesta referència. Igual succeeix amb la concatenació dels paràmetres i el conjunt reduït dels mateixos, els millors rendiments dels quals són un 4,84% absolut (33,36% relatiu) i un 2,44% absolut (6,82% relatiu), respectivament, inferiors al de referència.

La següent secció mostra, no obstant això, l'efecte constructiu d'aquests paràmetres lingüístics en combinació amb els acústics analitzats en la secció anterior [11].

6.4 Reconeixement basat en l'anàlisi de paràmetres acústics i lingüístics

L'anàlisi independent de les modalitats acústica i lingüística mostra diferències en els seus rendiments. Mentre que la modalitat acústica es mostra com a vàlida per dur a terme el reconeixement afectiu automàtic del corpus FAU Aibo, amb resultats de classificació superiors al de referència, la modalitat lingüística proposada en aquest treball no aconsegueix els mateixos resultats i obté taxes de classificació fins i tot inferiors a aquest últim.

En aquesta secció s'aborda l'experiment de reconeixement afectiu combinant ambdues modalitats mitjançant tècniques de fusió a nivell de decisió i a nivell de paràmetres. No obstant això, la parametrització acústica completa suposa un volum molt més gran

[11] Cal assenyalar que, en el desenvolupament de la línia de recerca d'aquesta tesi, el procés de fusió va evolucionar en paral·lel amb el procés de selecció i optimització dels paràmetres de forma individual. Per aquest motiu, la secció 6.4.1 empra els unigrames a nivell de paraula per determinar quin esquema de fusió és el més convenient malgrat que aquests no són els paràmetres lingüístics òptims per millorar la taxa de classificació. Les subsegüents seccions sí consideren la informació lingüística al complet.

de dades (384 paràmetres) que la parametrització lingüística (15 paràmetres, en la seva versió íntegra). Malgrat que en un esquema de fusió a nivell de decisió aquest fet no seria tan important atès que la classificació per a cada modalitat es realitza de forma independent, la concatenació de tots dos conjunts de dades resultaria en vectors amb dos tipus d'informació en proporció poc equilibrada.

Per aquest motiu en aquesta secció es planteja una anàlisi preliminar destinada a estudiar la millor tècnica de reducció de paràmetres i el millor mètode de fusió a partir dels dos conjunts reduïts de dades acústiques i el conjunt de dades d'unigrames a nivell de paraula. A continuació s'estudia l'efecte de combinar la parametrització acústica i la lingüística considerant la resta de paràmetres amb l'objectiu d'optimitzar els resultats de reconeixement.

6.4.1 Anàlisi preliminar de selecció de paràmetres i tècniques de fusió

L'anàlisi preliminar plantejada en aquesta secció consisteix en la fusió tant a nivell de decisió com de paràmetres de les modalitats acústica i lingüística, tal com es descriu a (Planet i Iriondo, 2012b). La parametrització acústica considera els dos conjunts de dades reduïdes mitjançant la cerca voraç incremental i decremental exposada en la secció 6.2.2.3. La parametrització lingüística considera el conjunt de dades d'activació referent als unigrames a nivell de paraula. L'objectiu és determinar, mitjançant els conjunts de dades reduïdes, quin dels dos conjunts acústics és l'òptim així com la millor tècnica de fusió per a l'experiment de reconeixement afectiu automàtic.

Es consideren tres esquemes de classificació: Naïve-Bayes, SVM i SL, seguint les mateixes configuracions que les descrites en les seccions anteriors. Quant als esquemes de fusió, per a la fusió a nivell de paràmetres s'empra la tècnica que uneix els paràmetres acústics i lingüístics de cada element del corpus en una única instància mitjançant concatenació, i per a la fusió a nivell de decisió s'empra la tècnica de *stacking* descrita en la secció 6.2.2.1 considerant un esquema J4.8 com a classificador de nivell 1 i els tres esquemes abans citats com a classificadors de nivell 0. Dos classificadors de nivell 0 del mateix tipus serveixen per classificar ambdues modalitats i les seves decisions categòriques són després fusionades pel classificador de nivell 1. La figura 6.19 mostra de forma esquemàtica la metodologia d'aquest experiment.

La figura 6.20 recull els resultats d'aquesta anàlisi preliminar. Els rendiments obtinguts pels conjunts acústics i lingüístics reduïts analitzats de forma independent (com en les figures 6.16 i 6.17) s'inclouen per facilitar la comparació dels resultats.

Focalitzant en el valor mitjà dels experiments Fold 1 i Fold 2, es pot observar que el rendiment dels classificadors que analitzen el conjunt de 28 paràmetres acústics seleccionats per la cerca voraç incremental és millor, en general, que el rendiment dels mateixos classificadors que empren tan sols els 5 paràmetres lingüístics. En el cas de l'esquema SVM, l'ús dels paràmetres acústics millora el valor de UAR en un 8,45% absolut (27,70%

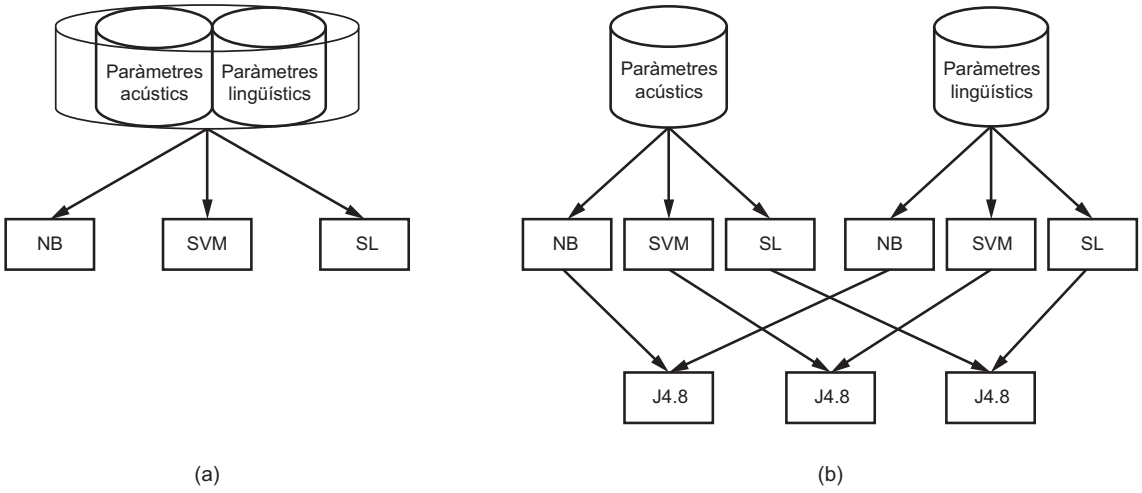


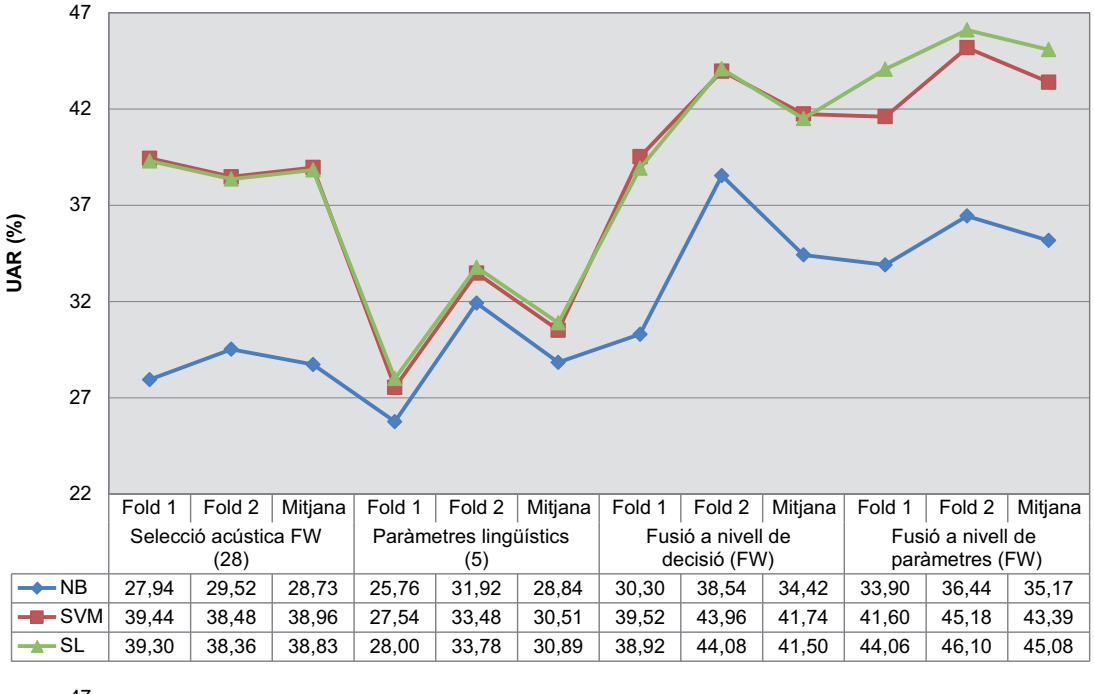
Figura 6.19: Representació esquemàtica dels esquemes de fusió de les modalitats acústica i lingüística (a) a nivell de paràmetres mitjançant concatenació i (b) a nivell de decisió mitjançant un esquema de *stacking*.

relatiu) respecte a l'ús dels paràmetres lingüístics. En el cas de l'esquema SL la millora és 7,94% absoluta (25,70% relativa). Només l'esquema Naïve-Bayes obté una lleugera millora en el seu rendiment en utilitzar els paràmetres lingüístics en lloc dels acústics (0,11% absoluta, 0,38% relativa).

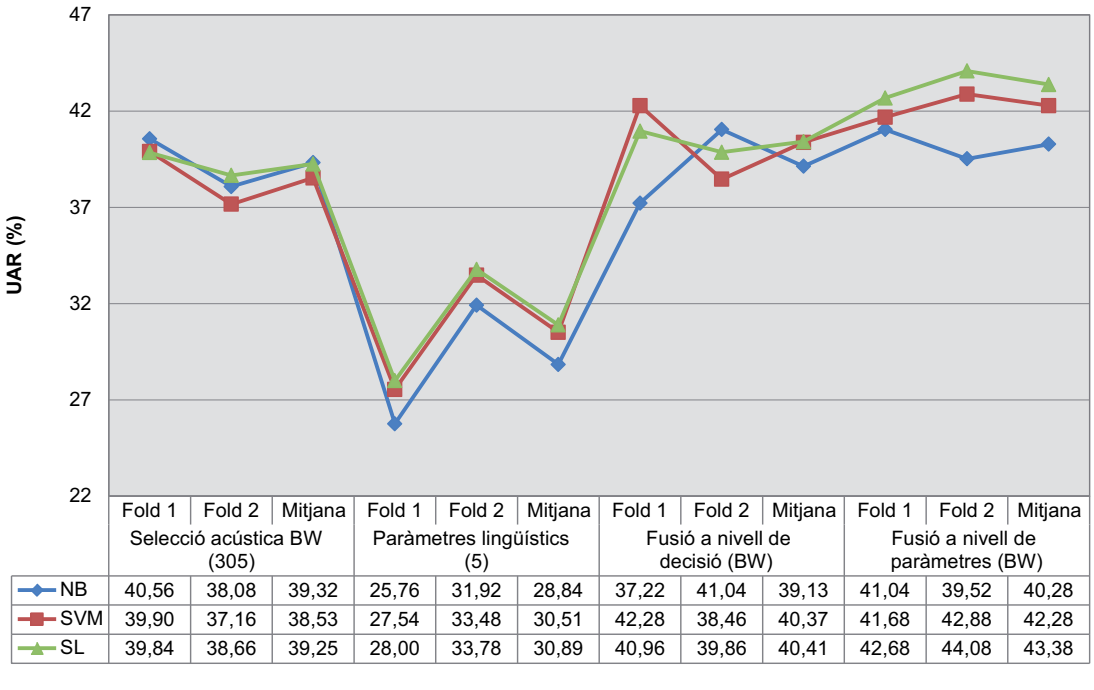
En el cas de la selecció realitzada per la cerca voraç decremental, el rendiment dels classificadors emprant els 305 paràmetres acústics és millor, en tots els casos, que emprant els 5 paràmetres lingüístics. La millora en el cas de l'esquema Naïve-Bayes, SVM i SL és 10,48% absoluta (36,34% relativa), 8,02% absoluta (26,29% relativa) i 8,36% absoluta (27,06% relativa), respectivament.

No obstant això, la combinació dels paràmetres acústics i lingüístics tant a nivell de decisió com a nivell de paràmetres millora les taxes de classificació dels classificadors que consideren ambdues modalitats de forma independent. En el cas de la selecció acústica realitzada per la cerca voraç incremental, la fusió a nivell de decisió obté una millora en la mitjana de les taxes de classificació aconseguides pels conjunts acústics i lingüístics en els casos dels algorismes Naïve-Bayes, SVM i SL del 5,63% absoluta (19,56% relativa), 7,00% absoluta (20,15% relativa) i 6,64% absoluta (19,05% relativa), respectivament. En mitjana aquesta millora és del 3,71%. En el cas de la fusió a nivell de paràmetres la millora és 6,38% absoluta (22,16% relativa), 8,65% absoluta (24,90% relativa) i 10,22% absoluta (29,32% relativa), respectivament. En mitjana, aquesta millora és del 8,42%.

En el cas de la selecció acústica realitzada per la cerca voraç decremental, la fusió a nivell de decisió obté una millora en la mitjana de les taxes de classificació aconseguides pels conjunts acústics i lingüístics en els casos dels algorismes Naïve-Bayes, SVM i SL del 5,85% absoluta (16,95% relativa) i 5,34% absoluta (15,23% relativa), respectivament. En el cas del classificador Naïve-Bayes es produeix una degradació del rendiment (-5,05% abso-



	Fold 1	Fold 2	Mitjana	Fold 1	Fold 2	Mitjana	Fold 1	Fold 2	Mitjana	Fold 1	Fold 2	Mitjana
	Selecció acústica FW (28)			Paràmetres lingüístics (5)			Fusió a nivell de decisió (FW)			Fusió a nivell de paràmetres (FW)		
NB	27,94	29,52	28,73	25,76	31,92	28,84	30,30	38,54	34,42	33,90	36,44	35,17
SVM	39,44	38,48	38,96	27,54	33,48	30,51	39,52	43,96	41,74	41,60	45,18	43,39
SL	39,30	38,36	38,83	28,00	33,78	30,89	38,92	44,08	41,50	44,06	46,10	45,08



	Fold 1	Fold 2	Mitjana	Fold 1	Fold 2	Mitjana	Fold 1	Fold 2	Mitjana	Fold 1	Fold 2	Mitjana
	Selecció acústica BW (305)			Paràmetres lingüístics (5)			Fusió a nivell de decisió (BW)			Fusió a nivell de paràmetres (BW)		
NB	40,56	38,08	39,32	25,76	31,92	28,84	37,22	41,04	39,13	41,04	39,52	40,28
SVM	39,90	37,16	38,53	27,54	33,48	30,51	42,28	38,46	40,37	41,68	42,88	42,28
SL	39,84	38,66	39,25	28,00	33,78	30,89	40,96	39,86	40,41	42,68	44,08	43,38

Figura 6.20: Taxa de classificació no ponderada (UAR), en percentatge, dels classificadors analitzant el conjunt de dades acústiques (28 paràmetres seleccionats per la cerca voraç incremental (FW, a dalt) i 305 paràmetres seleccionats per la cerca voraç decremental (BW, a baix)), el conjunt de dades lingüístiques basades en els unigrames a nivell de paraula i els esquemes de fusió a nivell de decisió i de paràmetres. Adaptat de (Planet i Iriondo, 2012b).

luta). En mitjana aquesta millora és del 2,05%. En el cas de la fusió a nivell de paràmetres la millora és 6,20% absoluta (18,19% relativa) per al classificador Naïve-Bayes, 7,76% absoluta (22,48% relativa) per al classificador SVM i 8,31% absoluta (23,70% relativa) per al classificador SL. En mitjana, aquesta millora és del 7,42%.

Com es pot observar, la millora aconseguida amb la fusió de les modalitats acústica i lingüística és més important quan es considera el conjunt acústic seleccionat per la cerca voraç incremental fusionat a nivell de paràmetres amb el conjunt de dades lingüístiques (8,42%).

Com en les seccions anteriors, tan sols ha de considerar-se la columna Fold 1 de la figura 6.20 per poder comparar aquests resultats amb els d'altres autors en el mateix escenari. Tal com es recull en la secció 4.4.2, Schuller *et al.* (2011) fusionen 7 esquemes de classificació i diferents conjunts de dades per obtenir un valor de UAR del 44,01%. El millor resultat obtingut en aquesta secció mitjançant un classificador SL i la fusió a nivell de paràmetres dels 28 paràmetres acústics seleccionats mitjançant la cerca incremental i els 5 paràmetres lingüístics (és a dir, un total de 33 paràmetres únicament) millora aquest resultat en un 0,06% absolut (0,14% relatiu). Encara que tots dos rendiments són bastant similars cal destacar que el nombre de paràmetres és molt més petit en l'esquema aquí plantejat, així com també és inferior la complexitat de l'esquema de classificació proposat.

6.4.2 Fusió de paràmetres seleccionats acústics i lingüístics

En aquesta secció s'aborda el reconeixement afectiu automàtic a través de l'anàlisi conjunta de les modalitats acústica i lingüística a partir dels resultats anteriorment obtinguts. L'esquema de fusió considerat és el de fusió a nivell de paràmetres mitjançant concatenació, tal com es determina en la secció prèvia. La reducció acústica es realitza mitjançant la cerca voraç incremental, si bé també es considera el conjunt de dades acústiques complet per realitzar, al final d'aquesta secció, una reducció de forma conjunta amb els paràmetres lingüístics. A nivell lingüístic s'estudien els diversos paràmetres d'activació que inclouen també els no analitzats en la secció anterior.

L'estructura bàsica d'aquesta secció consta de tres parts. La primera defineix els diferents conjunts de dades per simplificar la notació en el text. La segona combina la parametrització acústica amb els diferents conjunts lingüístics per determinar quins optimitzen el rendiment del reconeixement afectiu. Aquesta combinació tracta cadascun d'aquests conjunts íntegrament, considerant-los indivisibles i irreductibles, i es considera, per tant, manual. La tercera considera un procés automàtic per reduir, en primer lloc, la parametrització lingüística i, en segon lloc, el conjunt complet de paràmetres acústics i lingüístics.

Taula 6.3: Descripció dels conjunts de dades creats a partir dels conjunts complets de 384 paràmetres acústics i 15 lingüístics extrets del corpus FAU Aibo.

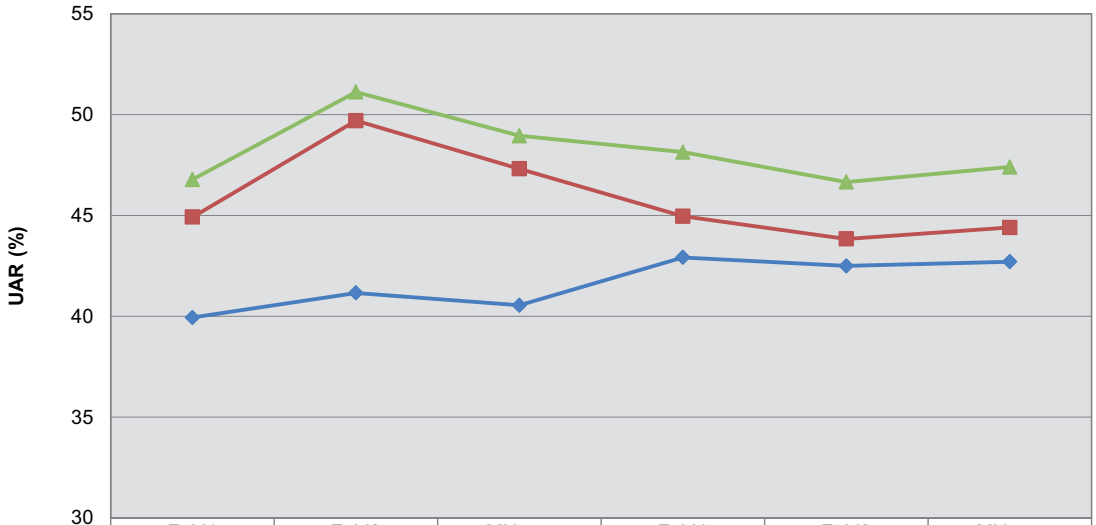
Tipus	Nom	Descripció
Acústic	Data1	28 paràmetres acústics seleccionats per una cerca voraç incremental a partir de *Data2*
	Data2	Conjunt original de 384 paràmetres acústics
Lingüístic	Data3	Conjunt original de 15 paràmetres lingüístics (equivalent a la concatenació de *Data4*, *Data5* i *Data6*)
	Data4	5 paràmetres lingüístics d'activació de l'anàlisi dels unigrames a nivell de paraula
	Data5	5 paràmetres lingüístics d'activació de l'anàlisi dels bigrames a nivell de paraula
	Data6	5 paràmetres lingüístics d'activació de l'anàlisi dels trigrames a nivell POS
	Data7	6 paràmetres lingüístics seleccionats per una cerca voraç incremental a partir de *Data3*
Acústic i lingüístic	Data8	159 paràmetres seleccionats per un algorisme genètic a partir de la concatenació de *Data2* i *Data3*

6.4.2.1 Definició dels conjunts de dades

Per simplificar la notació tant en el text com en les figures, s'ha assignat un nom específic a cadascun dels diferents conjunts de dades la definició de les quals es recull en la taula 6.3. *Data1* i *Data2* són conjunts de dades acústiques tal com es defineixen segons es recull en la secció 6.2.2.3: *Data1* es refereix als 28 paràmetres acústics seleccionats per una cerca voraç incremental mentre que *Data2* és el conjunt complet de paràmetres acústics. Els conjunts de dades *Data3*, *Data4*, *Data5*, *Data6* i *Data7* són lingüístics i es refereixen als paràmetres detallats en la secció 3.4.3.2. *Data3* és el conjunt de dades lingüístiques complet. Els altres conjunts de dades, excepte *Data7*, són diferents subconjunts creats a partir d'aquest conjunt complet. Així doncs, *Data4* i *Data5* recullen, respectivament, els 5 paràmetres d'activació calculats a partir dels unigrames i bigrames a nivell de paraula, mentre que *Data6* recull els 5 paràmetres d'activació calculats a partir dels trigrames a nivell POS. *Data7* és el subconjunt de paràmetres lingüístics seleccionat automàticament mitjançant una cerca voraç incremental (tal com es detalla en la secció 6.3.2) a partir del conjunt lingüístic complet. Finalment, *Data8* és un conjunt de 159 paràmetres acústics i lingüístics creat mitjançant un procés de selecció de paràmetres basat en l'algorisme genètic referenciat en la secció 4.3.2 i a partir de la concatenació dels conjunts *Data2* i *Data3*. *Data8* conté 149 paràmetres relacionats amb l'anàlisi acústica i 10 amb l'anàlisi lingüística. Només 2 d'aquests 10 paràmetres lingüístics es refereixen a l'activació dels unigrames a nivell de paraula mentre que 4 es refereixen a l'activació dels bigrames a nivell de paraula i els altres 4 als trigrames a nivell POS.

6.4.2.2 Fusió de paràmetres acústics i paràmetres lingüístics seleccionats manualment

En primer lloc s'estudia el rendiment dels classificadors en analitzar el conjunt de dades lingüístiques complet en combinació amb el conjunt de dades acústiques complet i reduït. Aquest plantejament és similar al de la secció anterior si bé en aquest cas s'empra el conjunt de dades lingüístiques complet. La figura 6.21 mostra els resultats UAR dels classificadors Naïve-Bayes, SVM i SL en ser avaluats, tal com s'ha establert, amb el conjunt de dades *Data3* en combinació, mitjançant concatenació, amb els conjunts *Data1* i *Data2*.



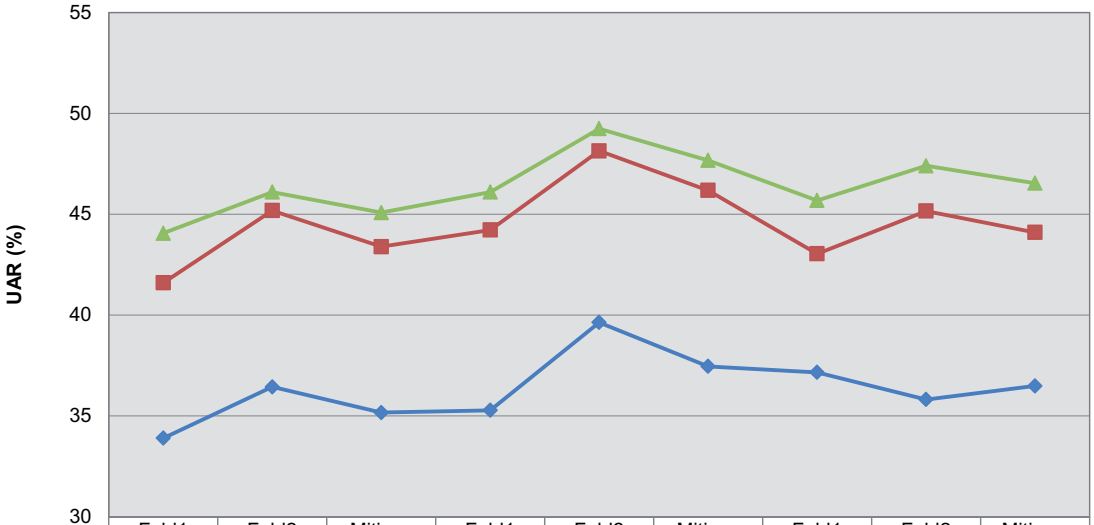
	Fold1	Fold2	Mitjana	Fold1	Fold2	Mitjana
		Data1 + Data3			Data2 + Data3	
NB	39,94	41,16	40,55	42,92	42,50	42,71
SVM	44,92	49,70	47,31	44,96	43,84	44,40
SL	46,78	51,12	48,95	48,14	46,66	47,40

Figura 6.21: Taxa de classificació no ponderada (UAR), en percentatge, dels classificadors analitzant el conjunt complet de paràmetres lingüístics (*Data3*) en combinació amb el conjunt reduït de paràmetres acústics (*Data1*) i el conjunt acústic complet (*Data2*).

Tal com succeïa en els casos anteriors, els rendiments dels classificadors milloren en utilitzar el conjunt reduït de dades acústiques, excepte en el cas de l'algorisme Naïve-Bayes. D'acord al valor mitjà dels experiments recollits en les columnes Fold1 i Fold2, l'ús del conjunt de dades *Data1* suposa una millora en comparació de l'ús del conjunt *Data2* d'un 2,91% absolut (6,55% relatiu) en el cas de l'esquema SVM i d'un 1,55% absolut (3,27% relatiu) en el cas de l'esquema SL. Per aquest motiu, *Data1* és el conjunt de dades acústiques considerat a partir d'aquest moment per a la resta d'experiments.

En la secció 6.3.2 es presentava la selecció lingüística realitzada per una cerca voraç incremental. No obstant això, la naturalesa del conjunt de dades lingüístiques permet una anàlisi detallada dels mateixos que observa, a diferència de l'estudi realitzat de forma independent en aquesta secció i en la secció 6.3.1, com afecta al rendiment dels classificadors la seva fusió amb els paràmetres acústics. La figura 6.22 mostra els rendiments obtin-

guts pels tres classificadors anteriors a través de l'anàlisi del conjunt reduït de paràmetres acústics (*Data1*) en combinació amb cadascun dels tres grups de paràmetres lingüístics: els paràmetres d'activació dels unigrames a nivell de paraula (*Data4*), bigrames a nivell de paraula (*Data5*) i trigrames a nivell POS (*Data6*).



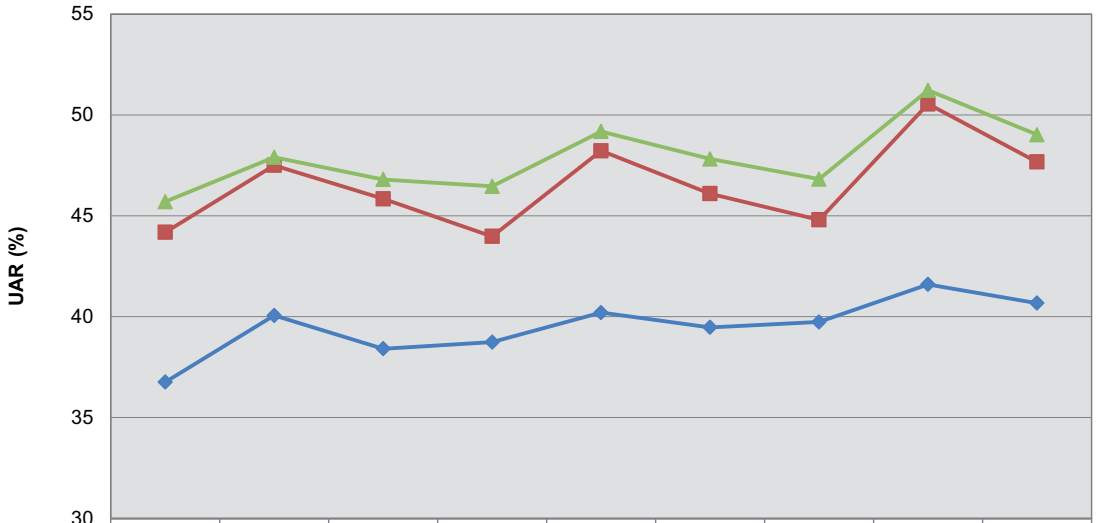
	Data1 + Data4			Data1 + Data5			Data1 + Data6		
	Fold1	Fold2	Mitjana	Fold1	Fold2	Mitjana	Fold1	Fold2	Mitjana
NB	33,90	36,44	35,17	35,28	39,64	37,46	37,16	35,82	36,49
SVM	41,60	45,18	43,39	44,22	48,14	46,18	43,04	45,16	44,10
SL	44,06	46,10	45,08	46,10	49,24	47,67	45,68	47,40	46,54

Figura 6.22: Taxa de classificació no ponderada (UAR), en percentatge, dels classificadors en fusionar el conjunt reduït de paràmetres acústics (*Data1*) amb els 3 subconjunts de paràmetres lingüístics (*Data4*, *Data5* i *Data6*) de forma independent.

D'acord a la mitjana dels dos experiments, la combinació de *Data1* i *Data4* obté el pitjor resultat d'aquest experiment. Per contra, la concatenació dels conjunts *Data1* i *Data5* obté el millor rendiment millorant en un 2,29% absolut (6,51% relatiu) el de l'esquema Naïve-Bayes usant la combinació anterior de conjunts de dades i millorant, també, els algorismes SVM i SL en un 2,79% absolut (6,43% relatiu) i en un 2,59% absolut (5,75% relatiu), respectivament, amb aquesta mateixa combinació. La combinació de *Data1* i *Data6* obté els segons millors resultats, millorant els esquemes Naïve-Bayes, SVM i SL en un 1,30% absolut (3,69% relatiu), 0,71% absolut (1,64% relatiu) i 1,46% absolut (3,24% relatiu), respectivament, els resultats obtinguts amb la primera combinació de conjunts de dades.

Per determinar si la combinació de dos subconjunts lingüístics és capaç de millorar el rendiment obtingut en l'experiment anterior mitjançant només un d'aquests subconjunts es creen tres combinacions dels mateixos dos a dos (*Data4* i *Data5*, *Data4* i *Data6*, *Data5* i *Data6*) fusionant-les amb la parametrització acústica de *Data1*. Els resultats apareixen representats en la figura 6.23. Com calia esperar, la combinació dels conjunts *Data5* i *Data6* (els dos millors conjunts de dades trobats en els resultats exposats en la figura 6.22) millora les altres dues combinacions. En comparació de la segona millor combina-

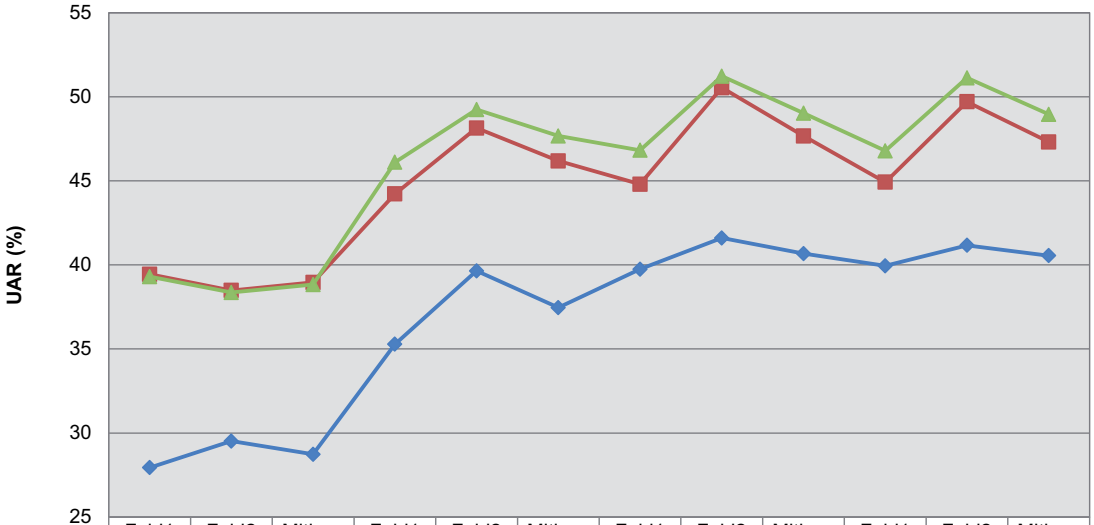
ció (*Data4* i *Data6*), el rendiment dels algorismes Naïve-Bayes, SVM i SL és millor en un 1,20% absolut (3,04% relatiu), 1,57% absolut (3,41% relatiu) i 1,20% absolut (2,51% relatiu), respectivament.



	Fold1	Fold2	Mitjana	Fold1	Fold2	Mitjana	Fold1	Fold2	Mitjana
	Data1 + Data4 + Data5			Data1 + Data4 + Data6			Data1 + Data5 + Data6		
NB	36,76	40,06	38,41	38,74	40,20	39,47	39,74	41,60	40,67
SVM	44,18	47,50	45,84	43,98	48,22	46,10	44,80	50,54	47,67
SL	45,70	47,90	46,80	46,46	49,18	47,82	46,82	51,22	49,02

Figura 6.23: Taxa de classificació no ponderada (UAR), en percentatge, dels classificadors fusionant el conjunt reduït de paràmetres acústics (*Data1*) amb tres combinacions dos a dos dels subconjunts lingüístics.

Finalment, per determinar els paràmetres lingüístics més rellevants, els millors conjunts de dades abans trobats (*Data1*, *Data5* i *Data6*) es comparen en la figura 6.24, en la qual els rendiments de *Data1* i la combinació de *Data1* i *Data3* s'inclouen com a referència. Com es pot observar, la inclusió dels paràmetres lingüístics millora el rendiment dels classificadors que només analitzen paràmetres acústics. Els millors resultats s'obtenen mitjançant la combinació de *Data5* i *Data6* concatenats amb *Data1*. D'acord amb el valor mitjà de les dues iteracions, aquest conjunt de dades procedent de la concatenació dels tres conjunts citats (el valor de UAR per a l'esquema SL és del 49,02%) millora els resultats obtinguts per la combinació de *Data1* i *Data5* (vegeu la figura 6.22), per als algorismes Naïve-Bayes, SVM i SL, en un 3,21% absolut (8,57% relatiu), 1,49% absolut (3,23% relatiu) i 1,35% absolut (2,83% relatiu), respectivament. Els resultats obtinguts per la combinació dels 3 datasets també milloren els resultats obtinguts per la combinació de *Data1* i *Data3*, és a dir, emprant la totalitat dels paràmetres lingüístics.



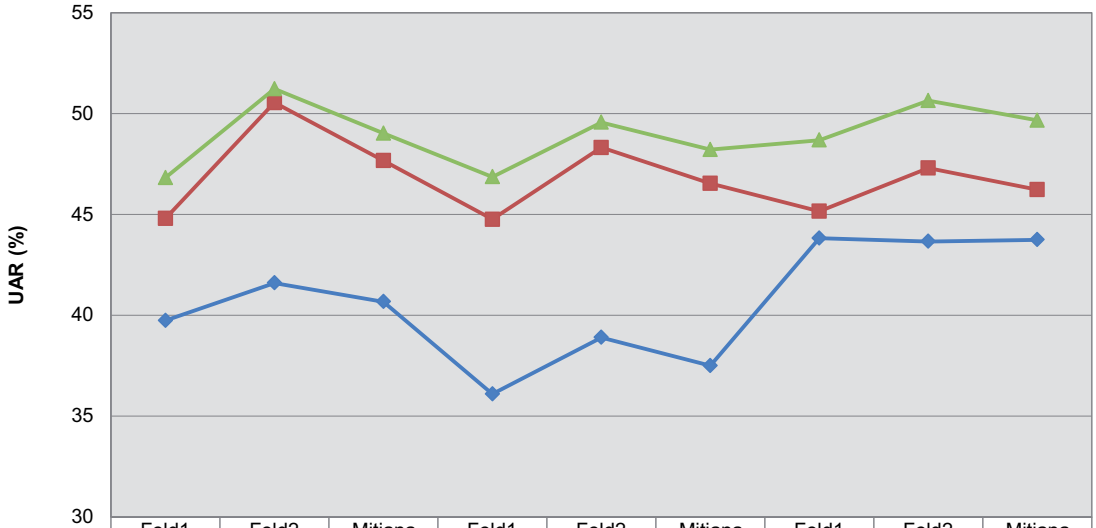
	Fold1	Fold2	Mitjana	Fold1	Fold2	Mitjana	Fold1	Fold2	Mitjana	Fold1	Fold2	Mitjana
	Data1			Data1 + Data5			Data1 + Data5 + Data6			Data1 + Data3		
NB	27,94	29,52	28,73	35,28	39,64	37,46	39,74	41,60	40,67	39,94	41,16	40,55
SVM	39,44	38,48	38,96	44,22	48,14	46,18	44,80	50,54	47,67	44,92	49,70	47,31
SL	39,30	38,36	38,83	46,10	49,24	47,67	46,82	51,22	49,02	46,78	51,12	48,95

Figura 6.24: Comparació de les taxes de classificació no ponderades (UAR) dels classificadors en analitzar el conjunt reduït de paràmetres acústics (*Data1*) de forma independent i en afegir, de forma incremental, els paràmetres d'activació dels bigrames a nivell de paraula (*Data5*) i els paràmetres d'activació dels trigrames a nivell POS, així com la seva fusió amb la totalitat de paràmetres lingüístics (*Data3*).

6.4.2.3 Fusió de paràmetres acústics i paràmetres lingüístics seleccionats automàticament

L'anàlisi conjunta dels paràmetres acústics i cadascun dels grups que formen part del conjunt de dades lingüístiques mostra que la millor combinació és la resultant de la concatenació dels conjunts de dades *Data1*, *Data5* i *Data6* (acústics i lingüístics de bigrames a nivell de paraula i trigrames a nivell POS, respectivament). Aquest conjunt parteix de la selecció acústica realitzada de forma automàtica i la parametrització lingüística seleccionada mitjançant una anàlisi manual i és coherent amb la selecció realitzada de forma automàtica en la secció 6.3.2 (la qual es recull en el conjunt *Data7*), ja que els paràmetres procedeixen dels dos mateixos grups si bé la selecció automàtica no els selecciona tots.

La figura 6.25 mostra els resultats UAR dels classificadors en analitzar tres conjunts de dades diferents: la combinació de *Data1*, *Data5* i *Data6*, que és la combinació dels paràmetres acústics seleccionats per una cerca voraç incremental i els paràmetres lingüístics més rellevants trobats manualment; la combinació de *Data1* i *Data7*, similar a l'anterior però usant els paràmetres lingüístics seleccionats per una cerca voraç incremental (vegeu la secció 6.3.2); i *Data8*, el conjunt de dades reduït creat a partir de la parametrització acústica i lingüística completa mitjançant un algorisme genètic (vegeu la secció 6.4.2.1).



	Data1 + Data5 + Data6			Data1 + Data7			Data8		
	Fold1	Fold2	Mitjana	Fold1	Fold2	Mitjana	Fold1	Fold2	Mitjana
NB	39,74	41,60	40,67	36,10	38,90	37,50	43,82	43,66	43,74
SVM	44,80	50,54	47,67	44,76	48,32	46,54	45,16	47,30	46,23
SL	46,82	51,22	49,02	46,86	49,56	48,21	48,68	50,64	49,66

Figura 6.25: Taxa de classificació no ponderada (UAR), en percentatge, dels classificadors considerant la fusió del conjunt reduït de paràmetres acústics (*Data1*) amb els conjunts de dades *Data5* i *Data6*, amb el conjunt reduït de paràmetres lingüístics (*Data7*) i el subconjunt seleccionat per un algorisme genètic a partir del conjunt complet acústic i lingüístic (*Data8*).

Els resultats mostren que la combinació de *Data1* i *Data7* obté el rendiment més baix dels conjunts de dades analitzats en aquesta part encara que és bastant similar a l'obtingut per la combinació de *Data1*, *Data5* i *Data6*. *Data8* obté, en general (excepte pel cas de l'algorisme SVM), el millor resultat de classificació (valor de UAR, com a mitjana de les iteracions, del 49,66% per a l'algorisme SL). Comparat amb la combinació de *Data1*, *Data5* i *Data6*, obté una millora en el rendiment per als algorismes Naïve-Bayes i SL del 3,07% absolut (7,55% relatiu) i del 0,64% absolut (1,31% relatiu), respectivament. Considerant els algorismes de classificació, la millora més significativa es produeix en el cas de l'algorisme Naïve-Bayes. Malgrat que els conjunts reduïts degradaven el rendiment d'aquest classificador en els experiments anteriors, *Data8* és un conjunt creat mitjançant un procés que afavoreix una correlació creuada mínima entre els paràmetres escollits, condició requerida per l'algorisme Naïve-Bayes per a un funcionament òptim (Rish, 2001).

6.4.2.4 Comparació global de resultats

Per comparar els resultats amb els obtinguts per altres autors en el mateix escenari cal considerar la columna Fold 1 ja que equival al resultat de l'avaluació amb els dos conjunts independents d'entrenament i prova. La figura 6.26 recull alguns dels resultats citats en aquesta tesi comparant-los amb els més representatius trobats en aquesta secció.

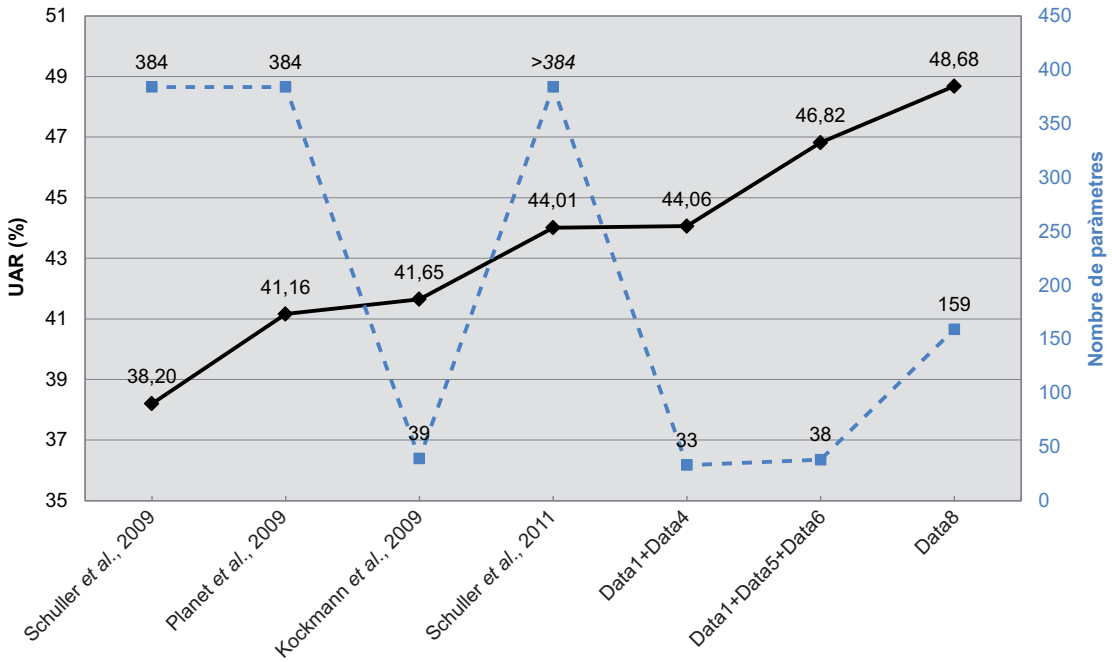


Figura 6.26: Comparació entre els resultats de diversos experiments de l'estat de la qüestió i els resultats més rellevants presentats en aquest capítol en emprar dos conjunts independents per a entrenament i prova. Es mostra la taxa de classificació no ponderada (UAR, en negre) i el nombre de paràmetres del conjunt de dades amb el qual s'aconsegueix aquest resultat (en blau). En el cas de (Schuller *et al.*, 2011) el nombre de paràmetres mostrat ha de ser considerat com a orientatiu ja que és superior a l'indicat perquè prové de la fusió de 7 esquemes de classificació diferents.

Kockmann *et al.* (2009) aconsegueixen un valor de UAR del 41,65% mitjançant 4 classificadors GMM i 13 paràmetres MFCC i les seves derivades. D'altra banda, la fusió de 7 esquemes de classificació realitzada per Schuller *et al.* (2011) aconsegueix una taxa de classificació del 44,01%. El nombre de paràmetres emprat per cadascun d'aquests 7 esquemes és diferent i entre ells figura la parametrització de 384 elements de Schuller *et al.* (2009) pel que el nombre total ascendeix a un valor superior[12]. D'altra banda, l'esquema SL (que aconsegueix el millor rendiment per a les tres combinacions de conjunts de dades presentades, tal com es comprova en la figura 6.25) obté el millor valor de UAR en analitzar el conjunt de dades *Data8* (48,68%). Aquest resultat millora el de Kockmann *et al.* (2009) en un 7,03% absolut (16,88% relatiu) i el de Schuller *et al.* (2011) en un 4,68% absolut (10,64% relatiu). Cal destacar que, a més de la millora respecte al resultat de Schuller *et al.* (2011), el volum de dades és de 159 paràmetres i emprant un únic esquema de classificació. Aquest resultat també millora l'obtingut per la combinació dels conjunts de dades *Data1* i *Data4*, que és d'un 44,06% de UAR (vegeu la secció 6.4.1). La millora és d'un 4,62% absolut (10,49% relatiu). En aquest cas, no obstant això, s'empra un volum de dades més

[12] Alguns dels estudis fusionats a nivell de decisió en (Schuller *et al.*, 2011) comparteixen part o la totalitat de la parametrització original de Schuller *et al.* (2009) i incorporen altres paràmetres addicionals. Per això pot considerar-se que, indirectament, s'analitza un corpus de grandària superior a l'especificat.

gran ja que la combinació de *Data1* i *Data4* ascendeix a un total de 33 paràmetres davant dels 159 de *Data8*. En una posició intermèdia es troba la selecció lingüística manual en combinació amb l'acústica (*Data1*, *Data5* i *Data6*). Amb 38 paràmetres, el seu rendiment se situa gairebé al mateix nivell que la combinació de *Data1* i *Data4* però per sota de *Data8*.

Tant *Data8* com la fusió de *Data1* i *Data4* superen el valor de UAR de referència establert per Schuller *et al.* (2009) i el resultat preliminar mostrat en la secció 6.2.1 tal com es recull en (Planet *et al.*, 2009). En el cas de *Data8* en comparació de Schuller *et al.* (2009), la millora absoluta és d'un 10,48% (27,43% relativa), emprant 159 paràmetres acústics i lingüístics enfront dels 384 paràmetres acústics del corpus analitzat per Schuller *et al.* (2009), és a dir, un 41,41% del total. La millora de la fusió entre *Data1* i *Data4* pel que fa al resultat de referència és inferior (5,86% absoluta, 15,34% relativa), encara que també és molt inferior el volum de dades emprat: 33 paràmetres enfront de 384, és a dir, un 8,59% del total. Referent a la combinació de *Data1*, *Data5* i *Data6*, la qual obté un resultat intermedi, aquesta millora és 8,62% absoluta (22,57% relativa) emprant 38 paràmetres, un 9,90% del total.

6.5 Resum

En aquesta secció s'ha tractat el reconeixement afectiu automàtic basat en un corpus de veu espontània, la qual cosa suposa treballar amb unes condicions específiques i molt diferents als corpus de veu actuada destacant, entre d'altres, la presència d'emocions no prototípiques i amb una baixa expressivitat afectiva. A més, el plantejament dels experiments és diferent atès que, en el cas de corpus actuats, es tracta amb locucions el contingut afectiu de les quals ha de ser validat posteriorment ja que el locutor podria no haver expressat correctament l'estat afectiu pretès. D'aquesta forma, el reconeixement pot plantejar-se com un exercici de mapatge d'un criteri subjectiu que accepti o rebutgi certes locucions sobre la base del seu estil afectiu. És a dir, que s'adapti a l'usuari que escolta i interpreta, en lloc del locutor que actua l'emoció. De no ser així, el sistema de reconeixement automàtic podria superar el propi reconeixement humà, la qual cosa suposaria que s'està produint el reconeixement de l'estil actuat en lloc del reconeixement de la pròpia emoció. En canvi, en corpus de veu espontània no té sentit aquesta validació ja que les locucions no pretenen un estil afectiu sinó que aquest estil s'assigna posteriorment en forma d'etiqueta, després d'escoltar les mateixes, ja que aquestes s'han obtingut d'una interacció espontània. L'exercici de reconeixement ja estaria adaptat a l'usuari que escolta i interpreta les emocions. A més, en aquest cas, s'ha treballat en un entorn de diversos locutors, més d'acord amb un escenari de propòsit general i, per tant, menys específic.

El corpus en el qual es basa aquesta anàlisi és el FAU Aibo el qual presenta una distribució de categories afectives no equilibrada. Per aquest motiu la mesura emprada és la UAR, a efectes de tenir en compte la correcta classificació de les categories afectives majoritàries i minoritàries per igual. D'altra banda, i per mantenir la total independència entre les dades emprades per entrenar i provar el sistema així com poder comparar els

resultats amb altres estudis de l'estat de la qüestió, s'empren dos subconjunts diferents i independents del corpus, consistent en locucions de locutors diferents.

L'estudi s'estructura en tres parts diferenciant l'anàlisi a nivell acústic i a nivell lingüístic, tots dos de forma independent, i a nivell de la fusió dels paràmetres acústics i lingüístics.

A nivell acústic s'han realitzat dos experiments. El primer d'ells, preliminar, s'emmarca dins de l'escenari proposat per Schuller *et al.* (2009), a nivell de classificació i de selecció de paràmetres. A nivell de classificació s'estudien diversos classificadors simples, dues estructures jeràrquiques i una estructura en cascada. De tots ells, l'esquema Naïve-Bayes és el que presenta un millor rendiment. A nivell de selecció de paràmetres es presenta una selecció de 100 elements realitzada mitjançant l'algorisme mRMR, la qual mostra una lleugera degradació en la taxa de reconeixement de referència malgrat suposar una reducció important en el volum de dades.

El segon experiment a nivell acústic compta amb una parametrització més precisa del corpus FAU Aibo i també realitza una tasca de classificació i una de selecció de paràmetres. A nivell de classificació es realitza, en primer lloc, un experiment categòric que analitza diversos algorismes de classificació i introdueix la fusió a nivell de classificació de dos d'ells i, en segon lloc, un experiment de classificació dimensional basat en el mapatge de les categories afectives en un espai bidimensional d'activació-avaluació, el resultat de la qual és inferior als obtinguts anteriorment. A nivell de selecció de paràmetres, en aquest cas s'ha emprat un algorisme de cerca voraç en les seves variants incremental i decremental, el qual presenta un millor rendiment que l'algorisme mRMR anteriorment estudiat.

A nivell lingüístic el corpus s'analitza a nivell de paraula sobre la base d'unigrames i bigrames i a nivell de la categoria morfològica de les paraules sobre la base de trigrames POS. En primer lloc s'estudien els tres conjunts de forma independent, sent els últims conjunts els que ofereixen un millor rendiment. D'altra banda, es presenta el rendiment de la fusió a nivell de paràmetres dels tres conjunts mitjançant concatenació. A més, malgrat que la modalitat lingüística compta amb un volum reduït de dades, també es presenta la reducció de les mateixes mitjançant una cerca voraç incremental. En tots els casos els resultats d'aquesta modalitat són inferiors als de referència, si bé s'observa posteriorment que l'anàlisi lingüística aporta un efecte constructiu a la classificació realitzada a nivell acústic.

La fusió acústica i lingüística es realitza, de forma preliminar, a nivell de decisió i a nivell de paràmetres, analitzant les seleccions de paràmetres acústics realitzades de forma incremental i decremental. D'aquest primer estudi es comprova que la selecció de paràmetres acústics incremental és la que presenta un millor rendiment, considerant el seu menor volum de dades i la seva taxa de classificació. D'altra banda, es determina que l'esquema de fusió a nivell de paràmetres és, donada la seva senzillesa i els bons resultats proporcionats, la més adequada per a aquest treball.

Determinat el millor esquema de fusió i la selecció incremental de paràmetres acústics com la més adequada, es realitzen diferents combinacions amb els paràmetres lingüístics. En primer lloc es duen a terme combinacions considerant els tres conjunts lingüístics detallats. En segon lloc es considera la fusió dels paràmetres acústics amb la reducció lingüística realitzada amb anterioritat així com una reducció mitjançant un algorisme genètic aplicat a la totalitat dels paràmetres acústics i lingüístics, sent aquest últim conjunt de dades el que aconsegueix un rendiment òptim. En combinació amb l'algorisme SL, el seu resultat millora notablement els resultats obtinguts per esquemes més complexos de l'estat de la qüestió.

Capítol 7

Conclusions i futures línies de recerca

Aquest últim capítol recull les principals conclusions i les línies futures de recerca. Aquestes conclusions apareixen agrupades en tres blocs. En primer lloc, en la secció 7.1, es recullen les conclusions generals de la tesi, les quals deriven directament dels objectius específics detallats en la secció 1.2. La secció 7.2 recull les conclusions de l'experimentació realitzada amb corpus de veu actuada. El tercer bloc, corresponent a la secció 7.3, detalla les conclusions de l'experimentació relacionada amb el corpus de veu espontània FAU Aibo. Finalment, la secció 7.4 proposa algunes línies de recerca que podrien abordar-se en el futur.

7.1 Conclusions generals

La motivació que dóna origen a aquesta tesi és cobrir la branca de l'anàlisi afectiva en la comunicació oral persona-màquina. El GTM comptava amb experiència en el camp de la síntesi audiovisual i, en el moment d'iniciar-se aquest treball de recerca, existien en desenvolupament diversos treballs relacionats amb la síntesi expressiva. No obstant això, existia també la necessitat de plantejar un estudi en el vessant analític per completar el procés comunicatiu, permetent avançar en el camp de la interacció persona-màquina de forma completa.

Així mateix, en general i tal com s'aprecia en els estudis analitzats en l'estat de la qüestió, fins a fa uns anys l'anàlisi dels senyals d'entrada de les interfícies d'usuari no havia estat tan tractada com sí ho havia estat la generació de contingut en la sortida de les mateixes. Així, mentre començaven a desenvolupar-se avatars virtuals amb síntesi audiovisual expressiva o motors de diàleg capaços de processar llenguatge natural, la interacció per part de l'usuari quedava reduïda, en la majoria de les ocasions, a la introducció de text

a través del teclat o, en el millor dels casos, al dictat de frases a través d'un micròfon. No obstant això, les millores en els mòduls de reconeixement de veu van permetre crear aplicacions més amigables i amb elles, i gràcies a l'augment de les capacitats computacionals dels dispositius, es va produir un auge en els estudis relacionats amb el reconeixement afectiu automàtic.

No obstant això, gran part d'aquests estudis se centraven en l'anàlisi d'enregistraments realitzats en condicions molt controlades, sovint actuades en estudis d'enregistrament professionals, amb una gran supervisió del contingut i de les condicions de l'entorn. Això desembocava en sistemes amb taxes de reconeixement molt elevades que, quan eren provats en condicions realistes d'interacció natural, produïen resultats molt per sota dels inicials. És per això que arran d'aquestes proves es van originar cada vegada més estudis focalitzats en l'anàlisi de locucions amb expressions emocionals espontànies el que suposava treballar, sovint, amb corpus poc equilibrats, que incloïen dades no prototípiques i locucions d'una baixa intensitat emocional i, freqüentment, poc definida, com és el cas del corpus FAU Aibo estudiat en aquesta tesi.

Per aquest motiu, aquesta tesi planteja com a objectiu general la creació d'un sistema de reconeixement afectiu automàtic basat en expressió oral espontània, la qual cosa comportava en la secció 1.2 a plantejar quatre objectius específics, les conclusions sobre els quals es desenvolupen a continuació. Les seccions següents detallaran aquests punts emmarcant-los en el seu context d'experimentació, a partir dels treballs realitzats en corpus de veu actuada i en corpus de veu espontània.

Realització de l'exercici de reconeixement afectiu en condicions de laboratori per a la seva comparació amb els realitzats en corpus de veu espontània. Seguint la línia dels treballs realitzats per diversos investigadors en l'àmbit del reconeixement afectiu automàtic i, en concret, el complet treball desenvolupat per Oudeyer (2003), en aquesta tesi s'ha plantejat l'anàlisi de dos corpus de veu actuada i estimulada, el corpus ESDLA i el corpus BDP-UAB, respectivament, mitjançant un banc d'algorismes de classificació individuals i parametritzacions basades en paràmetres prosòdics i de VoQ. L'estudi dels resultats posa de manifest que el sistema automàtic obté una taxa de classificació fins i tot superior al reconeixement que, de forma subjectiva, realitza un conjunt de persones. En l'estudi dut a terme amb el corpus de veu espontània FAU Aibo no cal considerar el reconeixement subjectiu atès que les locucions s'etiqueten després del seu enregistrament, és a dir, l'etiqueta afectiva associada s'assigna després que l'enregistrament s'hagi realitzat mentre que en l'enregistrament actuat el locutor ha d'expressar l'emoció que se li sol·licita. No obstant això la taxa de reconeixement és molt inferior en el cas de la veu espontània, més encara quan només es consideren paràmetres acústics per a l'anàlisi.

Parametrització a nivell acústic i lingüístic i posterior selecció dels paràmetres més rellevants. El corpus FAU Aibo no està gravat en condicions òptimes d'estudi sinó que, malgrat haver-se controlat diferents condicions ambientals, s'ha realitzat en un entorn

d'interacció realista. A més, les locucions no pertanyen a actors i es corresponen amb veus de diferents persones. Totes aquestes diferències pel que fa a corpus com el ESDLA o el BDP-UAB fan que el càlcul dels paràmetres acústics sigui menys robust. Per aquest motiu, a més de la parametrització acústica proposada en el capítol 3, s'incorpora una sèrie de paràmetres lingüístics extrets a partir de les transcripcions de les locucions del corpus, els quals no es veuen afectats pels inconvenients que introdueixen les condicions d'enregistrament. Cal assenyalar que en aquesta tesi s'ha treballat amb les transcripcions exactes i no les procedents d'un sistema de reconeixement automàtic de veu, que seria el desitjable en un sistema completament automàtic. Així doncs, els resultats expressats han de considerar-se com una cota màxima, partint de la premissa que el reconeixedor funcionaria sense error. Malgrat la baixa taxa de reconeixement aconseguida pels paràmetres lingüístics de forma individual, la seva fusió amb els paràmetres acústics millora notablement els resultats que aquests obtenen per si sols. D'altra banda, la reducció de paràmetres aconseguida mitjançant la seva selecció aconsegueix mantenir o fins i tot millorar les taxes d'identificació aconseguides amb els conjunts de dades complets.

Definició d'un entorn de treball que permeti avaluar adequadament els mètodes proposats i comparar els resultats amb altres estudis. Una de les dificultats que es planteja amb freqüència en els estudis de reconeixement afectiu automàtic és la varietat de mètodes d'avaluació i mètriques emprades, la qual cosa sovint impedeix realitzar una comparació correcta dels resultats. Schuller *et al.* (2009) proposen un entorn de treball basat en la mètrica UAR, convenient per a l'estudi d'un corpus poc equilibrat com el FAU Aibo, així com un mètode d'avaluació basat en dos conjunts independents. Tots dos conjunts contenen informació de locutors diferents pel que no existeix el risc d'entrenar i provar el sistema de reconeixement amb la informació dels mateixos locutors. Els estudis realitzats sobre aquest corpus per diferents grups de recerca arran de la proposta inicial han seguit aquests postulats, així com la present tesi, per la qual cosa la comparació de resultats pot realitzar-se de forma correcta. Per completar les anàlisis realitzades en aquesta tesi, el mètode d'avaluació s'ha estès a una validació creuada de dues iteracions, intercanviant els conjunts d'entrenament i prova en la segona iteració i calculant el valor mitjà dels resultats obtinguts en tots dos passos. No obstant això, la comparació final sempre es realitza sobre la base del resultat obtingut en la primera iteració.

Selecció dels esquemes de classificació i fusió més adequats. En aquesta tesi s'ha tractat el reconeixement afectiu automàtic com un problema de classificació categòrica en el qual els elements a classificar són locucions parametritzades de veu i la classe és l'etiqueta descriptiva de l'emoció associada. Com tal s'han estudiat algorismes de classificació individuals i també formant estructures més complexes, considerant fins i tot més d'una classe per a cada locució en el cas de la classificació basada en el model bidimensional proposat en la secció 6.2.2.2. No obstant això, els esquemes que han proporcionat els millors resultats no han estat algorismes complexos sinó algorismes més senzills emprant conjunts de dades convenientment seleccionades. En aquest sentit, la fusió dels mateixos

ha resultat un punt clau, arribant-se a determinar que la fusió a nivell de paràmetres per concatenació és més adequada que la fusió a nivell de decisió malgrat la senzillesa del seu plantejament. Així mateix, la selecció de paràmetres realitzada mitjançant un algorisme genètic a partir dels paràmetres prèviament fusionats per concatenació és la que proporciona la millor taxa de reconeixement, la qual cosa posa de manifest la importància de la correcta selecció dels paràmetres comentada més amunt.

7.2 Reconeixement afectiu en corpus actuats

El reconeixement afectiu basat en corpus actuats suposa treballar en un escenari bastant allunyat del paradigma de les aplicacions reals ja que, en el marc d'una interacció natural, la comunicació persona-màquina es realitzarà de forma espontània, en condicions d'entorn variables i no controlades, en les quals l'expressivitat emocional no sempre es produirà amb la màxima intensitat ni de forma prototípica i en la qual, a més, aquestes emocions probablement distaran de pertànyer al conjunt de les emocions bàsiques. A més, els resultats obtinguts en un escenari d'aquest tipus són excessivament bons, fins i tot per sobre de la classificació que es realitzaria de forma subjectiva. Efectivament, reconèixer emocions de forma molt precisa quan es corresponen amb emocions plenes, gravades en condicions excel·lents i de forma molt controlada, manca de poc sentit en un escenari d'interacció natural. És per aquest motiu que, a més de l'estudi de reconeixement basat en els corpus ESDLA i BDP-UAB, s'ha plantejat un segon escenari en el qual aquest tipus de reconeixement sí pot ser d'utilitat: la validació automàtica d'un corpus destinat a síntesi de veu expressiva.

El procés de validació automàtica de corpus consisteix en la classificació d'un corpus actuat (i per tant prèviament etiquetat) que mapegi el criteri subjectiu de les persones que l'escolten i, per tant, decideixen si la locució es correspon adequadament amb l'emoció que se suposa que l'actor havia de transmetre en el moment del seu enregistrament. Per aconseguir-ho es realitza una avaluació subjectiva d'un fragment del corpus, el resultat del qual serveix per ajustar el sistema de reconeixement automàtic. L'ajust es realitza mitjançant un procés iteratiu que selecciona els paràmetres que maximitzen la mesura F1 computada sobre la base de la classificació subjectiva i l'objectiva. Finalment, mitjançant la combinació de classificadors gràcies a una tècnica de *stacking* s'aconsegueix refinar el resultat final.

L'anàlisi de resultats demostra que existeix un grau elevat de correspondència entre les decisions del sistema de validació automàtica i la percepció subjectiva dels participants en la prova d'escolta. Malgrat que aquest sistema encara podria incloure millores, el seu desenvolupament i presentació en aquesta tesi queda com a il·lustrador de l'aplicació real que un reconeixement afectiu d'aquestes característiques podria tenir. No obstant això el focus d'atenció se situa en un escenari d'emocions espontànies, les conclusions de l'experimentació de les quals s'aborden en la secció següent.

7.3 Reconeixement afectiu en corpus espontanis

El reconeixement afectiu basat en corpus de veu espontània és molt diferent al que es realitza sobre la base de corpus de veu actuada donades les pròpies característiques dels corpus. La presència d'emocions no prototípiques i locucions de baixa expressivitat afectiva fa que les condicions siguin molt específiques, com per exemple la classe R del corpus FAU Aibo emprat en aquesta tesi, la qual aglutina una sèrie de locucions sota una mateixa etiqueta malgrat que no es corresponen amb un estat afectiu ben definit. A més, s'ha treballat en un entorn de diversos locutors, entorn més d'acord amb un escenari de propòsit general.

Seguint la metodologia més adequada per poder comparar resultats amb altres recerques que empren el corpus FAU Aibo s'ha adoptat la mètrica UAR i un sistema d'avaluació basat en dos conjunts independents. La mesura UAR permet ponderar les classes majoritàries i minoritàries per igual atès que aquest corpus és altament no equilibrat. Els conjunts independents permeten mantenir aïllats els conjunts d'entrenament i prova durant tot l'experiment.

En primer lloc, el corpus s'ha analitzat a nivell acústic i a nivell lingüístic de forma independent. Les principals conclusions en referència a això es detallen a continuació:

- Un primer experiment preliminar posa de manifest que l'algorisme Naïve-Bayes, amb el processament previ del conjunt de dades acústiques mitjançant una discretització supervisada dels mateixos, és el que obté el millor valor de UAR, per sobre fins i tot d'altres esquemes més complexos.

- El classificador SVM de kernel lineal és també adequat quan s'entrena amb el conjunt de dades acústiques d'entrenament prèviament normalitzat i remostrejat per aconseguir una distribució uniforme de classes. Una fusió mitjançant una estratègia de *stacking* d'aquest classificador amb l'esquema Naïve-Bayes abans descrit millora mínimament aquest últim en termes de UAR però aconsegueix una millor classificació dels elements de la classe R.

- Abordar la classificació del corpus FAU Aibo des d'una perspectiva acústica mitjançant un plantejament bidimensional en termes d'activació i avaluació no ofereix resultats superiors als anteriors. La pròpia definició de les classes del corpus fa complicat el mapatge de les mateixes en aquest espai bidimensional, especialment per la definició de la classe R.

- Les reduccions automàtiques de paràmetres basades en cerques voraces de tipus incremental i decremental aconsegueixen reduir eficaçment el volum de dades, especialment en el cas de la cerca incremental. Per a aquest últim conjunt reduït de dades es mantenen les taxes de classificació en algorismes com SVM i SL però es redueix dràsticament en el cas de l'esquema Naïve-Bayes, ja que aquest mètode de

selecció no garanteix la independència dels paràmetres escollits, requisit d'aquest classificador.

- La modalitat lingüística analitzada de forma individual proporciona resultats inferiors als de la modalitat acústica i fins i tot per sota dels valors escollits com a referència durant l'experimentació. No obstant això, bigrames a nivell de paraula i trigrames a nivell POS es revelen com els més rellevants, per sobre de l'anàlisi lingüística de les paraules considerades individualment. No obstant això, no poden considerar-se suficients per abordar per si mateixos un problema de reconeixement afectiu automàtic.

En segon lloc, les modalitats acústica i lingüística són fusionades a diversos nivells en una fase posterior, la qual cosa proporciona les següents conclusions principals:

- En general, la fusió tant a nivell de decisió com a nivell de paràmetres de la informació acústica i lingüística millora els resultats de UAR obtinguts per ambdues modalitats de forma individual.

- En l'anàlisi conjunta de les modalitats acústica i lingüística es comprova que el millor esquema de fusió és el de concatenació a nivell de paràmetres, malgrat la seva senzillesa, i que la reducció acústica realitzada mitjançant una cerca voraç incremental és la més convenient per a la seva fusió amb la informació lingüística, malgrat proporcionar un conjunt de dades molt reduït, de només 28 paràmetres.

- Tal com es desprenia de l'anàlisi individual de la modalitat lingüística, bigrames a nivell de paraula i trigrames a nivell POS són els paràmetres lingüístics més rellevants. Fusionats per concatenació amb el conjunt reduït de 28 paràmetres acústics maximitzen el valor de UAR de l'algorisme SL.

- El millor esquema de classificació és el que empra un classificador SL i el conjunt de dades *Data8*, consistent en una selecció realitzada per un algorisme genètic a partir del conjunt complet de dades acústiques i lingüístiques. Dels 159 paràmetres que en formen part, 149 són acústics (davant dels 28 seleccionats per l'algorisme de cerca voraç incremental en la selecció prèvia) i 10 són lingüístics, 8 dels quals fan referència als bigrames a nivell de paraula i als trigrames a nivell POS. El seu resultat és el millor obtingut amb aquest corpus publicat fins avui en les condicions de treball plantejades en aquesta tesi.

- Amb el conjunt de dades *Data8* l'algorisme Naïve-Bayes obté un valor de UAR millor que amb altres conjunts reduïts de dades. Això és així perquè l'esquema de reducció basat en l'algorisme genètic sí està orientat a aconseguir una baixa correlació creuada entre els paràmetres escollits.

7.4 Línies de futur

Malgrat que els resultats obtinguts en analitzar el corpus FAU Aibo mitjançant els esquemes aquí proposats superen els de treballs realitzats en el mateix context, cal assenyalar que encara hi ha un llarg camí per recórrer en l'anàlisi d'aquest corpus per millorarlos notablement. No en va, el valor màxim de UAR aconseguit se situa per sota del 50% considerant l'anàlisi del corpus segons es defineix en (Schuller *et al.*, 2009) per a un escenari de cinc etiquetes afectives.

En aquest sentit de millora cabria l'ampliació de la parametrització lingüística la qual, actualment, presenta un volum de dades molt inferior al de la parametrització acústica. Tal com s'ha comprovat, els paràmetres lingüístics actuals no són capaços de realitzar per si sols una adequada classificació però sí milloren notablement els resultats de classificació obtinguts pels paràmetres acústics analitzats de forma individual.

En referència també als paràmetres lingüístics, la parametrització realitzada s'ha fet sobre la base de la transcripció manual de les locucions. Això suposa que aquesta transcripció està lliure d'errors. Aquesta suposició podria no ser realista en un context d'interacció natural en el qual el sistema d'anàlisi afectiva comptés amb un reconeixedor automàtic de veu. Les limitacions d'aquest reconeixedor comportarien la introducció d'errors en les transcripcions el que podria suposar una degradació del rendiment aquí assenyalat. En qualsevol cas, els resultats aquí exposats referents a això han de considerarse com una cota màxima suposant un reconeixedor ideal de veu i, per tant, sense errors.

Així mateix, seria convenient realitzar una anàlisi en altres idiomes. El corpus FAU Aibo emprat en aquesta tesi està gravat en alemany. No obstant això, podria resultar interessant provar aquestes mateixes tècniques en corpus similars gravats en altres idiomes per comprovar la validesa dels mètodes proposats i la seva robustesa davant de les variacions de l'expressió de les emocions en diferents llenguatges i cultures.

Finalment, la definició emprada del corpus FAU Aibo presenta detalls que haurien de ser tinguts en compte en el moment de dissenyar una aplicació per a un escenari real. El principal seria la reconsideració de la classe R. Aquesta classe no està ben definida i suposa una font important d'errors de classificació. En aquest sentit tal vegada hauria de ser ignorada com a tal i agrupar les seves locucions en diverses classes bé d'entre les ja existents o ben creant altres noves.

Bibliografia

Aha, D. W., Kibler, D., i Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, 6(1):37–66.

Alías, F. i Iriondo, I. (2002). La evolución de la síntesis del habla en Ingeniería y Arquitectura La Salle. En *Actas de las II Jornadas en Tecnología del Habla*, Granada, Spain.

Alías, F., Monzo, C., i Socoró, J. C. (2006). A pitch marks filtering algorithm based on restricted dynamic programming. En *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH 2006 ICSLP)*, p. 1698–1701, Pittsburgh, PA, USA.

Alías, F., Sevillano, X., Socoró, J. C., i Gonzalvo, X. (2008). Towards high-quality next-generation text-to-speech synthesis: A multidomain approach by automatic domain classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(7):1340–1354.

Ambady, N. i Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. *Psychological Bulletin*, 111(2):256–274.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., i Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. En Hansen, J. H. L. i Pellom, B. L., editors, *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, p. 2037–2040, Denver, Colorado, USA. ISCA.

Arnfield, S., Roach, P., Setter, J., Greasley, P., i Horton, D. (1995). Emotional stress and speech tempo variation. En *Proceedings of the ESCA-NATO Workshop on Speech Under Stress*, p. 13–15, Lisbon, Portugal.

Austermann, A., Esau, N., Kleinjohann, L., i Kleinjohann, B. (2005). Prosody based emotion recognition for MEXI. En *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005)*, p. 1138–1144, Alberta, Canada.

Averill, J. R. (1980). A constructivist view of emotion. *Emotion: Theory, research and experience*, 1:305–339.

Bachorowski, J. A. i Owren, M. J. (1995). Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science*, 6(4):219–224.

Banse, R. i Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614–636.

Barra, R. (2011). *Contributions to the analysis, design and evaluation of strategies for corpus-based emotional speech synthesis*. Tesi doctoral, Universidad Politécnica de Madrid.

Barra, R., Fernández, F., Lutfi, S., Lucas-Cuesta, J. M., Macías-Guarasa, J., Montero, J. M., San Segundo, R., i Pardo, J. M. (2009). Acoustic emotion recognition usin dynamic bayesian networks and multi-space distributions. En *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, p. 336–339, Brighton, UK.

Barra, R., Macías-Guarasa, J., Montero, J. M., Rincón, C., Fernández, F., i Córdoba, R. (2007). In search of primary rubrics for language independent emotional speech identification. En *Proceedings of the IEEE International Symposium on Intelligent Signal Processing (WISP 2007)*, Madrid, Spain.

Bartlett, M. S., Littlewort, G., Braathen, B., Sejnowski, T. J., i Movellan, J. R. (2003). A prototype for automatic recognition of spontaneous facial actions. En Becker, S., Thrun, S., i Obermayer, K., editors, *Advances in Neural Information Processing Systems*, volum 15, p. 1271–1278. MIT Press.

Bartneck, C. (2000). Affective expressions of machines. Tesi de màster, Stan Ackerman Institute, Eindhoven.

Batliner, A., Fischer, K., Huber, R., Spilker, J., i Nöth, E. (2000). Desperately seeking emotions: Actors, wizards and human beings. En *Proceedings of the ISCA Workshop on Speech and Emotion*, p. 195–200, Newcastle, Northern Ireland, UK.

Batliner, A., Fischer, K., Huber, R., Spilker, J., i Nöth, E. (2003). How to find trouble in communication. *Speech Communication*, 40(1-2):117–143.

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M. J., i Wong, M. (2004). "You stupid tin box"— Children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. En *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal. European Language Resources Association.

Batliner, A., Schuller, B., Seppi, D., Steidl, S., Devillers, L., Vidrascu, L., Vogt, T., Aharonson, V., i Amir, N. (2011). The automatic recognition of emotions in speech. En Cowie, R., Pelachaud, C., i Petta, P., editors, *Emotion-Oriented Systems*, Cognitive Technologies, p. 71–99. Springer Berlin Heidelberg.

Batliner, A., Steidl, S., Hacker, C., i Nöth, E. (2008). Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech. *User Modeling and User-Adapted Interaction*, 18:175–206.

Beller, G. i Marty, A. (2006). Talkapillar: outil d'analyse de corpus oraux. En *Rencontres Jeunes Chercheurs de l'Ecole Doctorale 268*, Paris, France.

Bezdek, J. C., Pal, M. R., Keller, J., i Krisnapuram, R. (1999). *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Kluwer Academic Publishers, Norwell, MA, USA.

Bhutekar, S. i Chandak, M. (2012). Corpus based emotion extraction to implement prosody feature in speech synthesis systems. *International Journal of Computer and Electronics Research*, 1(2).

Bozkurt, E., Erzin, E., Erdem, C. E., i Erdem, A. T. (2009). Improving automatic emotion recognition from speech signals. En *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, p. 324–327, Brighton, UK.

Bozkurt, E., Erzin, E., Erdem, C. E., i Erdem, A. T. (2011). Formant position based weighted spectral features for emotion recognition. *Speech Communication*, 53(9-10):1186–1197.

Bradley, M. M. i Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy & Experimental Psychiatry*, 25(1):49–59.

Bradley, M. M. i Lang, P. J. (1999). Affective Norms for English Words (ANEW): Stimuli, instruction manual, and affective ratings. Technical report, Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, USA.

Breazeal, C. L. (2002). *Designing Sociable Robots*. Intelligent Robotics and Autonomous Agents Series. MIT Press, Cambridge, MA, USA.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Broekens, J., Jonker, C. M., i Meyer, J.-J. C. (2010). Affective negotiation support systems. *Journal of Ambient Intelligence and Smart Environments*, 2(2):121–144.

Buck, R. (1999). The biological affects: a typology. *Psychological Review*, 106(2):301–336.

Bullington, J. (2005). 'Affective' computing and emotion recognition systems: the future of biometric surveillance? En *Proceedings of the 2nd Annual Conference on Information Security Curriculum Development (InfoSecCD 2005)*, InfoSecCD '05, p. 95–99, New York, NY, USA. ACM.

Burkhardt, F., Paeschke, A., Rolfes, M., i Sendlmeier, W. (2005). A database of german emotional speech. En *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH 2005)*, p. 1517–1520, Lisbon, Portugal. ISCA.

Burkhardt, F. i Sendlmeier, W. (2000). Verification of acoustical correlates of emotional speech using fromant-synthesis. En *Proceedings of the ISCA Workshop on Speech and Emotion*, p. 151–156, Belfast, North Ireland.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., i Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. En *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI 2004)*, p. 205–211, New York, NY, USA. ACM.

Busso, C., Lee, S., i Narayanan, S. (2009). Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech and Language Processing*, 17(4):582–596.

Busso, C. i Narayanan, S. (2008). Recording audio-visual emotional databases from actors: A closer look. En *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, p. 17–22, Marrakech, Morocco.

Cañamero, L. i Fredslund, J. (2001). I show you how I like you — Can you read it in my face? *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 31:454–459.

Cahn, J. (1989). Generating expression in synthesized speech. Tesi de màster, Massachusetts Institute of Technology.

Calzada, À. i Socoró, J. C. (2012). Voice quality modification using a harmonics plus noise model. *Cognitive Computation*.

Campbell, N. (2000). Databases of emotional speech. En *Proceedings of the ISCA Workshop on Speech and Emotion*, p. 34–38, Newcastle, Northern Ireland, UK.

Campbell, N. (2002). Recording techniques for capturing natural every-day speech. En *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain.

Canny, J. (2006). The future of human-computer interaction. *Queue*, 4(6):24–32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., i Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16(1):321–357.

Chen, L., Tao, H., Huang, T., Miyasato, T., i Nakatsu, R. (1998). Emotion recognition from audiovisual information. En *Proceedings of the IEEE 2nd Workshop on Multimedia Signal Processing*, p. 83 –88, California, USA.

Clavel, C., Vasilescu, I., Devillers, L., Richard, G., i Ehrette, T. (2008). Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50(6):487–503.

Cleary, J. G. i Trigg, L. E. (1995). K*: An instance-based learner using an entropic distance measure. En *12th International Conference on Machine Learning*, p. 108–114.

Cohn, J., Reed, L., Ambadar, Z., Xiao, J., i Moriyama, T. (2004). Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. En *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC 2004)*, volum 1, p. 610–616, The Hague, The Netherlands.

Cornelius, R. R. (2000). Theoretical approaches to emotion. En *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, p. 3–10, Belfast, Northern Ireland, UK.

Cowie, R. i Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1-2):5–32.

Cowie, R., Douglas-Cowie, E., Apolloni, B., Taylor, J., Romano, A., i Fellenz, W. (1999). What a neural net needs to know about emotion words. *Journal Computational Intelligence and Applications*, p. 109–114.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., i Schröder, M. (2000). FEELTRACE: An instrument for recording perceived emotion in real time. En *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, p. 19–24, Newcastle, Northern Ireland, UK.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., i Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80.

Cowie, R. i Schröder, M. (2005). Piecing together the emotion jigsaw. En Bengio, S. i Bourlard, H., editors, *Machine Learning for Multimodal Interaction*, volum 3361 de *Lecture Notes in Computer Science*, p. 305–317. Springer Berlin Heidelberg.

Darwin, C. R. (1872). *La expresión de las emociones en los animales y en el hombre*. Alianza Editorial. Traduït per Eusebio Heras.

de Melo, C. M., Carnevale, P., i Gratch, J. (2010). The influence of emotions in embodied agents on human decision-making. En *Proceedings of the 10th International Conference on Intelligent Virtual Agents (IVA 2010)*, p. 357–370, Philadelphia, PA.

de Melo, C. M., Carnevale, P., i Gratch, J. (2011). The effect of expression of anger and happiness in computer agents on negotiations with humans. En *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (ICAAMS 2011)*, volum 3, p. 937–944, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Decety, J. i Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, 3(2):71–100.

Devillers, L., Vidrascu, L., i Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422.

Dick, P. K. (1968). *¿Sueñan los androides con ovejas eléctricas?* Planeta DeAgostini. Traduït per César Terrón.

Douglas-Cowie, E., Campbell, N., Cowie, R., i Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, 40:33–60.

Douglas-Cowie, E., Devillers, L., Martin, J. C., Cowie, R., Savvidou, S., Abrilian, S., i Cox, C. (2005). Multimodal databases of everyday emotion: Facing up to complexity. En *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2005)*, p. 813–816, Lisbon, Portugal.

Drioli, C., Tisato, G., Cosi, P., i Tesser, F. (2003). Emotions and voice quality: experiments with sinusoidal modeling. En *Proceedings of Voice Quality: Functions, Analysis and Synthesis (VOQUAL 2003), ISCA Tutorial and Research Workshop*, p. 127–132, Geneva, Switzerland.

Dumouchel, P., Dehak, N., Attabi, Y., Dehak, R., i Boufaden, N. (2009). Cepstral and long-term features for emotion recognition. En *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, p. 344–347, Brighton, UK.

Ekman, P. (1999). Basic emotions. En Dalgleish, T. i Power, T., editors, *Handbook of Cognition and Emotion*, p. 45–60. John Wiley & Sons Ltd., Sussex, UK.

Ekman, P. i Friesen, W. V. (1969). The repertoire of nonverbal behaviour: Categories, origins, usage and coding. *Semiotica*, 1(1):49–97.

Empirisoft (2011). Medialab research software. En línea. Darrera visita el 4 de juny de 2013. Disponible en <http://www.empirisoft.com/medialab.aspx>.

Enos, F. (2009). *Detecting deception in speech*. Tesi doctoral, Columbia University, New York, USA.

Eriksson, A. i Lacerda, F. (2007). Charlatanry in forensic speech science: A problem to be taken seriously. *International Journal of Speech, Language and the Law*, 14(2):169–193.

Escudero, D. (2003). *Modelado Estadístico de Entonación con Funciones de Bézier: Aplicaciones a la Conversion Texto-Voz en Español*. Tesi doctoral, Universidad de Valladolid.

Eyben, F., Wöllmer, M., i Schuller, B. (2009). openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. En *Proceedings of the 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, p. 576–581, Amsterdam, The Netherlands.

Fayyad, U. M. i Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. En *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI 1993)*, p. 1022–1029, Chambéry, France.

Fernández, R. i Picard, R. W. (2003). Modeling drivers' speech under stress. *Speech Communication*, 40:145–159.

Formiga, L. (2010). *Optimització perceptiva dels sistemes de síntesi de la parla basats en selecció d'unitats mitjançant algorismes genètics interactius actius*. Tesi doctoral, La Salle, Universitat Ramon Llull.

Formiga, L., Trilla, A., Alías, F., Iriondo, I., i Socoró, J. C. (2010). Adaptation of the URL-TTS system to the 2010 Albayzín Evaluation Campaign. En *Proceedings of the VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop*, p. 363–370, Vigo, Spain.

Fragopanagos, N. i Taylor, J. (2005). Emotion recognition in human-computer interaction. *Neural Networks*, 18(4):389–405.

Francisco, V. (2008). *Identificación automática del contenido afectivo de un texto y su papel en la presentación de información*. Tesi doctoral, Universidad Complutense de Madrid, Madrid, Spain.

Francisco, V. i Gervás, P. (2006). Exploring the compositionality of emotions in text: Word emotions, sentence emotions and automated tagging. En *Proceedings of 21st National Conference on Artificial Intelligence (AAAI 2006). Workshop on Computational Aesthetics: AI Approaches to Beauty and Happiness*, Boston, MA, USA.

Francisco, V., Hervás, R., Peinado, F., i Gervás, P. (2012). EmoTales: creating a corpus of folk tales with emotional annotations. *Language Resources and Evaluation*, 46(3):341–381.

Frank, E. i Witten, I. H. (1998). Generating accurate rule sets without global optimization. En *Proceedings of the 15th International Conference on Machine Learning (ICML 1998)*, p. 144–151, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

François, H. i Boëffard, O. (2002). The greedy algorithm and its application to the construction of a continuous speech database. En *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, volum 5, p. 1420–1426, Las Palmas de Gran Canaria, España.

Freund, Y. i Schapire, R. E. (1996). Experiments with a new boosting algorithm. En *Proceedings of the 13th International Conference on Machine Learning (ICML 1996)*, p. 148–156, Bari, Italy.

Furnas, G. W., Landauer, T. K., Gómez, L. M., i Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971.

Gabrys, B. i Ruta, D. (2006). Genetic algorithms in classifier fusion. *Applied Soft Computing*, 6(4):337–347.

Gadallah, M. E., Matar, M. A., i Algezawi, A. F. (1999). Speech based automatic lie detection. En *Proceedings of the 16th National Radio Science Conference (NRSC 1999)*, p. C33/1 –C33/8, Cairo, Egypt.

Garrido, J. M. (1991). *Modelización de patrones melódicos del español para la síntesis y el reconocimiento del habla*. Universitat Autònoma de Barcelona, Bellaterra, España.

Gerrards-Hesse, A., Spies, K., i Hesse, F. W. (1994). Experimental inductions of emotional states and their effectiveness: a review. *British Journal of Psychology*, 85:55–78.

Giakoumis, D., Tzovaras, D., Moustakas, K., i Hassapis, G. (2011). Automatic recognition of boredom in video games using novel biosignal moment-based features. *IEEE Transactions on Affective Computing*, 2:119–133.

Gil, J. (2007). *Fonética para Profesores de Español: De la Teoría a la Práctica*. Manuales de formación de profesores de español 2/L. Arco Libros.

Giles, H. i Smith, P. (1979). Accommodation theory: Optimal levels of convergence. En Giles, H. i St. Clair, R. N., editors, *Language and Social Psychology*, p. 45–65. Blackwell, Oxford, UK.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1ª edició.

Goleman, D. (1995). *Inteligencia Emocional*. Editorial Kairós, Barcelona, Spain, 44ª edició. Traduït per David González i Fernando Mora.

Gonzalvo, J. (2010). *Síntesi basada en models ocults de Markov aplicada a l'espanyol i a l'anglès, les seves aplicacions i una proposta híbrida*. Tesi doctoral, La Salle, Universitat Ramon Llull.

Gonzalvo, X., Iriondo, I., Socoró, J. C., Alías, F., i Monzo, C. (2007a). Mixing HMM-based spanish speech synthesis with a CBR for prosody estimation. En Chetouani, M., Hussain, A., Gas, B., Milgram, M., i Zarader, J.-L., editors, *Advances in Nonlinear Speech Processing*, volum 4885 de *Lecture Notes in Computer Science*, p. 78–85. Springer Berlin Heidelberg.

Gonzalvo, X., Morán, J. A., Monzo, C., i Planet, S. (2005). Entorno para el aprendizaje automático de las estrategias de diálogo. En *Actas del XX Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2005)*, Gandia, Spain.

Gonzalvo, X., Socoró, J. C., Iriondo, I., C.Monzo, i Martinez, E. (2007b). Linguistic and mixed excitation improvements on a HMM-based speech synthesis for Castilian Spanish. En *Proceedings of the 6th ISCA Workshop on Speech Synthesis*, Bonn, Germany.

Gouizi, K., Bereksi, R. F., i Maaoui, C. (2011). Emotion recognition from physiological signals. *Journal of Medical Engineering and Technology*, 35(6-7):300–307.

Gratch, J. i Marsella, S. (2001). Tears and fears: modeling emotions and emotional behaviors in synthetic agents. En Müller, J. P., Andre, E., Sen, S., i Frasson, C., editors, *Proceedings of the 5th International Conference on Autonomous Agents (ICAA 2001)*, p. 278–285, Montreal, Canada. ACM Press.

Grimm, M., Kroschel, K., Mower, E., i Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10-11):787–800.

Gupta, P. i Rajput, N. (2007). Two-stream emotion recognition for call center monitoring. En *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007)*, p. 2241–2244, Antwerp, Belgium. ISCA.

Guyon, I. i Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. Tesi doctoral, University of Waikato, Hamilton, New Zealand.

Hassan, A. i Damper, R. I. (2009). Emotion recognition from speech using extended feature selection and a simple classifier. En *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, p. 2043–2046, Brighton, UK.

Hastie, T. i Tibshirani, R. (1998). Classification by pairwise coupling. *Annals of Statistics*, 26(2):451–471.

Heinemann, P. (1980). *Pedagogía de la comunicación no verbal*. Herder, Barcelona, Spain.

Hinde, R. A., editor (1972). *Non-verbal communication*. Cambridge University Press.

Hofer, G., Richmond, K., i Clark, R. (2005). Informed blending of databases for emotional speech synthesis. En *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH 2005)*, Lisbon, Portugal.

Hollingsed, T. K. i Ward, N. G. (2007). A combined method for discovering short-term affect-based response rules for spoken tutorial dialog. En *Proceedings of the ISCA ITRW Speech and Language Technology in Education (SLaTE 2007)*, p. 61–64, Pennsylvania, USA.

Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., i Nogueiras, A. (2002). Interface databases: Design and collection of a multilingual emotional speech database. En *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas de Gran Canaria, España.

Ioannou, S., Raouzaiou, A., Tzouvaras, V., Mailis, T. P., Karpouzis, K., i Kollias, S. D. (2005). Emotion recognition through facial expression analysis based on a neurofuzzy network. *Neural Networks*, 18:423–435.

Iriondo, I. (2008). *Producción de un corpus oral y modelado prosódico para la síntesis del habla expresiva*. Tesi doctoral, La Salle, Universitat Ramon Llull.

Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J., Bernadas, D., Oliver, J., Tena, D., i Longhi, L. (2000). Validation of an acoustical modelling of emotional expression in spanish using speech synthesis techniques. En *Proceedings of the ISCA Workshop on Speech and Emotion*, p. 161–166, Newcastle, Northern Ireland, UK.

Iriondo, I., Planet, S., Alías, F., Socoró, J. C., i Martínez, E. (2007a). Validation of an expressive speech corpus by mapping automatic classification to subjective evaluation. En Sandoval, F., Prieto, A., Cabestany, J., i Graña, M., editors, *Computational and Ambient Intelligence, 9th International Work-Conference on Artificial Neural Networks, IWANN 2007, San Sebastián, Spain, June 2007, Proceedings*, volum 4507 de *Lecture Notes in Computer Science*, p. 646–653. Springer.

Iriondo, I., Planet, S., Alías, F., Socoró, J. C., i Martínez, E. (2008). Emulating subjective criteria in corpus validation. En Rabuñal, J. R., Dorado, J., i Pazos, A., editors, *Encyclopedia of Artificial Intelligence*, p. 541–546. Information Science Reference.

Iriondo, I., Planet, S., Alías, F., Socoró, J. C., Monzo, C., i Martínez, E. (2007b). Expressive speech corpus validation by mapping subjective perception to automatic classification based on prosody and voice quality. En *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007)*, p. 2125–2128, Saarbrücken, Germany.

Iriondo, I., Planet, S., Socoró, J. C., i Alías, F. (2007c). Objective and subjective evaluation of an expressive speech corpus. En Chetouani, M., Hussain, A., Gas, B., Milgram, M., i Zarader, J. L., editors, *Advances in Nonlinear Speech Processing, International Conference on Nonlinear Speech Processing, NOLISP 2007, Paris, France, May 2007, Revised Selected Papers*, volum 4885 de *Lecture Notes in Artificial Intelligence*, p. 86–94. Springer.

Iriondo, I., Planet, S., Socoró, J. C., Martínez, E., Alías, F., i Monzo, C. (2009). Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification. *Speech Communication*, 51(9):744–758. Special issue on non-linear and conventional speech processing - NOLISP 2007.

Iriondo, I., Socoró, J. C., i Alías, F. (2007d). Prosody modelling of spanish for expressive speech synthesis. En *Proceedings of the 32nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, volum 4, p. 821–824, Honolulu, HI, USA.

Irtel, H. (2007). PXLab: The psychological experiments laboratory. En línea. Darrera visita el 4 de juny de 2013. Disponible en <http://www.pxlab.de>.

ITU-T (1996). *Methods for Subjective Determination of Transmission Quality. ITU-T Recommendation P.800*. ITU-T recommendation: Series P. International Telecommunication Union. Previously CCITT Recommendation.

James, W. (1884). What is an emotion? *Mind*, 9(34):188–205.

John, G. H. i Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. En *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (CUAI 1995)*, p. 338–345, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jovičić, S. T., Kašić, Z., Đorđević, M., i Rajković, M. (2004). Serbian emotional speech database: design, processing and evaluation. En *Proceedings of the 9th Conference on Speech and Computer (SPECOM 2004)*, p. 77–81, St. Petersburg, Russia.

Juslin, P. N. (1997). Perceived emotional expression in synthesized performances of a short melody: Capturing the listener's judgment policy. *Musicae Scientiae*, 1(2):225–256.

Juslin, P. N. i Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129(5):770–814.

Juslin, P. N. i Scherer, K. R. (2005). *The New Handbook of Methods in Nonverbal Behavior Research*, capítol Vocal expression of affect, p. 65–135. Oxford University Press.

Kapoor, A., Burleson, W., i Picard, R. W. (2007). Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8):724–736.

Kim, J. i André, E. (2006). Emotion recognition using physiological and speech signal in short-term observation. En André, E., Dybkjær, L., Minker, W., Neumann, H., i Weber, M., editors, *Perception and Interactive Technologies*, volum 4021 de *Lecture Notes in Computer Science*, p. 53–64. Springer Berlin Heidelberg.

Kockmann, M., Burget, L., i Černocký, J. (2009). Brno University of Technology system for Interspeech 2009 Emotion Challenge. En *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, p. 348–351, Brighton, UK.

Kohavi, R. (1995). The power of decision tables. En *Proceedings of the 8th European Conference on Machine Learning (ECML 1995)*, p. 174–189, Heraclion, Crete, Greece. Springer Verlag.

Kollias, S. i Piat, F. (1999). Principled hybrid systems: theory and applications (Physta). En *Proceedings of the IEEE International Conference on Multimedia Computing and Systems (ICMCS 1999)*, volum 2, p. 1089–1091, Florence, Italy.

Koolagudi, S. G. i Rao, K. (2012). Emotion recognition from speech: a review. *International Journal of Speech Technology*, 15(2):99–117.

Kostoulas, T., Mporas, I., Kocsis, O., Ganchev, T., Katsaounos, N., Santamaría, J. J., Jiménez-Murcia, S., Fernández-Aranda, F., i Fakotakis, N. (2012). Affective speech interface in serious games for supporting therapy of mental disorders. *Expert Systems with Applications*, 39(12):11072–11079.

Kotti, M., Paternò, F., i Kotropoulos, C. (2010). Speaker-independent negative emotion recognition. En *Proceedings of the 2nd International Workshop on Cognitive Information Processing (CIP 2010)*, p. 417–422, Elba Island, Italy.

Kwon, O. W., Chan, K., Hao, J., i Lee, T. W. (2003). Emotion recognition by speech signals. En *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003)*, p. 125–128, Geneva, Switzerland. ISCA.

Labov, W. (1972). *Sociolinguistic Patterns*. Conduct and Communication. University of Pennsylvania Press.

Landwehr, N., Hall, M., i Frank, E. (2005). Logistic model trees. *Machine Learning*, 59(1–2):161–205.

Lang, P. J. (1980). Behavioral treatment and bio-behavioral assessment: computer applications. En Sidowski, J. B., Johnson, J. H., i Williams, T. H., editors, *Technology in Mental Health Care Delivery Systems*, p. 119–137. Ablex, Norwood, NJ.

Langley, P. (1994). Selection of relevant features in machine learning. En *Proceedings of the Association for the Advancement of Artificial Intelligence Fall Symposium on Relevance (AAAI 1994)*, volum 97, p. 127–131, New Orleans, Louisiana, USA. AAAI Press.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge University Press.

Lazarus, R. S. (1999). *Stress and Emotion: A New Synthesis*. Springer Publishing Company.

Lazarus, R. S. (2001). *Appraisal processes in emotion: Theory, methods, research*, capítol Relational meaning and discrete emotions, p. 37–67. Series in Affective Science. Oxford University Press.

Lee, C. C., Mower, E., Busso, C., Lee, S., i Narayanan, S. (2009). Emotion recognition using a hierarchical binary decision tree approach. En *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, p. 320–323, Brighton, UK.

Lee, C. M. i Narayanan, S. (2005). Towards detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303.

Lee, C. M., Narayanan, S., i Pieraccini, R. (2001). Recognition of negative emotions from the speech signal. En *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2001)*, p. 240–243, Trento, Italy.

Leventhal, H. (1980). Toward a comprehensive theory of emotion. En Berkowitz, L., editor, *Advances in Experimental Social Psychology*, volum 13, p. 139–207. Academic Press.

Lewis, M. i Haviland, J. M. (1993). *Handbook of Emotions*. Guilford Publications.

Li, L. i Chen, J. H. (2006). Emotion recognition using physiological signals. En Pan, Z., Cheok, A., Haller, M., Lau, R., Saito, H., i Liang, R., editors, *Advances in Artificial Reality and Tele-Existence*, volum 4282 de *Lecture Notes in Computer Science*, p. 437–446. Springer Berlin Heidelberg.

Liberman, M., Davis, K., Grossman, M., Martey, N., i Bell, J. (2002). Emotional prosody speech and transcripts.

Lin, Y. P., Wang, C. H., Jung, T. P., Wu, T. L., Jeng, S. K., Duann, J. R., i Chen, J. H. (2010). EEG-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering*, 57(7):1798–1806.

Litman, D. i Forbes-Riley, K. (2003). Recognizing emotions from student speech in tutoring dialogues. En *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2003)*, p. 25–30, St. Thomas, VI, USA.

Litman, D. i Forbes-Riley, K. (2004). Predicting student emotions in computer-human tutoring dialogues. En *Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics (ACL 2004)*, Barcelona, Spain.

Litman, D. i Forbes-Riley, K. (2006). Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, 48(5):559–590.

Litman, D., Friedberg, H., i Forbes-Riley, K. (2012). Prosodic cues to disengagement and uncertainty in physics tutorial dialogues. En *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012)*, Portland, Oregon.

Llisterri, J., Carbó, C., Machuca, M. J., de la Mota, C., Riera, M., i Ríos, A. (2004). *Tecnologías del texto y del habla*, capítol La conversión de texto en habla: aspectos lingüísticos, p. 145–186. Edicions de la Universitat de Barcelona y Fundación Duques de Soria, Barcelona.

Luengo, I., Navas, E., i Hernáez, I. (2009). Combining spectral and prosodic information for emotion recognition in the Interspeech 2009 Emotion Challenge. En *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, p. 332–335, Brighton, UK.

Luengo, I., Navas, E., Hernáez, I., i Sánchez, J. (2005). Automatic emotion recognition using prosodic parameters. En *Proceedings of the 6th Annual Conference of the International Speech Communication Association (INTERSPEECH 2005)*, p. 493–496, Lisbon, Portugal.

Maria, K. A. i Zitar, R. A. (2007). Emotional agents: A modeling and an application. *Information and Software Technology*, 49(7):695–716.

Matsumoto, D. i Willingham, B. (2009). Spontaneous facial expressions of emotion of congenitally and noncongenitally blind individuals. *Journal of Personality and Social Psychology*, 96(1):1–10.

McGilloway, S., Cooper, S. J., i Douglas-Cowie, E. (2003). Can patients with chronic schizophrenia express emotion? a speech analysis. *Schizophrenia Research*, 64(2-3):189–190.

McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., i Stroeve, S. (2000). Approaching automatic recognition of emotion from voice: a rough benchmark. En *Proceedings of the ISCA Workshop on Speech and Emotion*, Belfast, North Ireland.

Mehrabian, A. i Russell, J. (1974). *An approach to environmental psychology*. MIT Press, Cambridge, MA.

Melenchón, J. (2007). *Síntesis audiovisual realista personalizable*. Tesi doctoral, Enginyeria i Arquitectura La Salle, Universitat Ramon Llull, Barcelona, Spain.

Michaelis, D., Gramss, T., i Strube, H. (1997). Glottal to noise excitation ratio — A new measure for describing pathological voices. *Acustica / Acta acustica*, 83:800–806.

Mihalcea, R. i Strapparava, C. (2009). The lie detector: explorations in the automatic recognition of deceptive language. En *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009)*, p. 309–312, Stroudsburg, PA, USA. Association for Computational Linguistics.

Montero, J. A., Alías, F., Garriga, C., Vicent, L., i Iriondo, I. (2007). Assessing students' teamwork performance by means of fuzzy logic. En *Proceedings of the 9th International Work-Conference on Artificial Neural Networks (IWANN 2007)*, San Sebastián, Spain.

Montero, J. M., Gutiérrez-Arriola, J., Palazuelos, S., Enríquez, E., Aguilera, S., i Pardo, J. M. (1998a). Emotional speech synthesis: From speech database to TTS. En *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 1998)*, p. 923–926, Sydney, Australia.

Montero, J. M., Gutiérrez-Arriola, J., Palazuelos, S., Enríquez, E., i Pardo, J. M. (1998b). Spanish emotional speech from database to TTS. En *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 1998)*, p. 923–925, Sydney, Australia.

Montoya, N. (1998). El papel de la voz en la publicidad audiovisual dirigida a los niños. *Zer. Revista de estudios de comunicación*, (4):161–177.

Monzo, C. (2010). *Modelado de la cualidad de la voz para la síntesis del habla expresiva*. Tesi doctoral, La Salle, Universitat Ramon Llull.

Monzo, C., Alías, F., Iriondo, I., Gonzalvo, X., i Planet, S. (2007). Discriminating expressive speech styles by voice quality parameterization. En *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007)*, p. 2081–2084, Saarbrücken, Germany.

Monzo, C., Morán, J. A., Planet, S., i Gonzalvo, X. (2004). Protocolo DS-CDMA orientado a la gestión de un auditorio. En *Actas del XIX Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2004)*, Barcelona, Spain.

Morrison, D., Wang, R., i De Silva, L. C. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49(2):98–112.

Mozziconacci, S. (1998). *Speech variability and emotion: Production and perception*. Tesi doctoral, Technical University of Eindhoven, Eindhoven, The Netherlands.

Murray, I. R. i Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion. *Journal of the Acoustic Society of America*, 93(2):1097–1108.

Nasoz, F., Alvarez, K., Lisetti, C. L., i Finkelstein, N. (2004). Emotion recognition from physiological signals using wireless sensors for presence technologies. *International Journal of Cognition, Technology and Work*, 6(1):4–14.

Navas, E., Hernáez, I., i Luengo, I. (2006). An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1117–1127.

NBS (2013). Presentation, precise, powerful stimulus delivery. En línea. Darrera visita el 4 de juny de 2013. Disponible en <http://www.neurobs.com/>.

Neiberg, D., Elenius, K., i Laskowski, K. (2006). Emotion recognition in spontaneous speech using GMM. En *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2006)*, p. 809–812, Pittsburgh, Pennsylvania, USA.

Nöth, E., Batliner, A., Niemann, H., Stemmer, G., Gallwitz, F., i Spilker, J. (2001). Language models beyond word strings. En *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2001)*, p. 167–176, Italy.

Ochs, M. i Prendinger, H. (2010). A virtual character's emotional persuasiveness. En *Proceedings of the International Conference on Kansei Engineering and Emotion Research (KEER 2010)*, Paris.

Osherenko, A. (2008). Towards semantic affect sensing in sentences. En *Proceedings of Annual Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB 2008)*, Aberdeen, Scotland, UK.

Osherenko, A., André, E., i Vogt, T. (2009). Affect sensing in speech: Studying fusion of linguistic and acoustic features. En *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII 2009)*, Amsterdam, The Netherlands.

Oudeyer, P. Y. (2003). The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, 59(1-2):157–183. Special issue on Affective Computing.

Oviatt, S. (2002). *Handbook of Human-Computer Interaction*. Lawrence Erlbaum.

Pammi, S., Charfuelan, M., i Schröder, M. (2010). Multilingual voice creation toolkit for the MARY TTS platform. En Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., i Tapias, D., editors, *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta. European Language Resources Association.

Pantic, M. i Rothkrantz, L. J. M. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390.

Payrató, L. (2005). Aspectos del discurso multimodal: Relaciones ente elementos verbales, prosódicos y gestuales en el discurso oral. En *Curso de Tecnologías Lingüísticas: El Futuro de los Sistemas de Diálogo*. Fundación Duques de Soria, Soria.

Peleg, G., Katzir, G., Peleg, O., Kamara, M., Brodsky, L., Or, H. H., Keren, D., i Nevo, E. (2006). Hereditary family signature of facial expression. *Proceedings of the National Academy of Sciences of the United States of America*, 103(43):15921–15926.

Peng, H., Long, F., i 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):1226–1238.

Pérez-Espinosa, H., Reyes-García, C. A., i Villaseñor-Pineda, L. (2011). EmoWisconsin: an emotional children speech database in Mexican Spanish. En *Affective Computing and Intelligent Interaction*, p. 62–71. Springer Berlin Heidelberg.

Picard, R. W. (2000). *Affective Computing*. MIT Press.

Picard, R. W., Vyzas, E., i Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191.

Planet, S. i Iriondo, I. (2011a). Improving spontaneous children's emotion recognition by acoustic feature selection and feature-level fusion of acoustic and linguistic parameters. En Travieso, C. i Alonso, J., editors, *Advances in Nonlinear Speech Processing, 5th International Conference on Nonlinear Speech Processing, NOLISP 2011, Las Palmas de Gran Canaria, Spain, November 2011, Proceedings*, volum 7015 de *Lecture Notes in Computer Science*, p. 88–95. Springer Berlin/Heidelberg.

Planet, S. i Iriondo, I. (2011b). Spontaneous children's emotion recognition by categorical classification of acoustic features. En Rocha, A., Gonçalves, R., Pérez Cota, M., i Reis, L. P., editors, *Actas da 6a Conferência Ibérica de Sistemas e Tecnologias de Informação (CISTI 2011)*, volum 1, p. 594–599, Chaves, Portugal.

Planet, S. i Iriondo, I. (2012a). Children's emotion recognition from spontaneous speech using a reduced set of acoustic and linguistic features. *Cognitive Computation*.

Planet, S. i Iriondo, I. (2012b). Comparative study on feature selection and fusion schemes for emotion recognition from speech. *International Journal of Interactive Multimedia and Artificial Intelligence*, 1(6):44–51. Special Issue on Intelligent Systems and Applications.

Planet, S. i Iriondo, I. (2012c). Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition. En Rocha, A., Calvo Manzano, J. A., Reis, L. P., i Pérez Cota, M., editors, *Actas de la 7a Conferencia Ibérica de Sistemas y Tecnologías de la Información (CISTI 2012)*, volum 1, p. 199–204, Madrid, Spain.

Planet, S., Iriondo, I., Martínez, E., i Montero, J. A. (2008). TRUE: an online Testing platfoRm for mUltimedia Evaluation. En *Proceedings of the 2nd International Workshop on EMOTION: Corpora for Research on Emotion and Affect at the 6th Conference on Language Resources & Evaluation (LREC 2008)*, p. 61–65, Marrakech, Morocco.

Planet, S., Iriondo, I., Socoró, J. C., Monzo, C., i Adell, J. (2009). GTM-URL contribution to the Interspeech 2009 Emotion Challenge. En *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, p. 316–319, Brighton, UK.

Planet, S., Morán, J. A., i Formiga, L. (2006). Reconocimiento de emociones basado en el análisis de la señal de voz parametrizada. En Cunha, M. M. i Rocha, A., editors, *Actas da I Conferência Ibérica de Sistemas e Tecnologias de Informação (CISTI 2006)*, volum 2, p. 837–854, Ofir, Portugal.

Planet, S., Morán, J. A., Monzo, C., i Gonzalvo, X. (2004). Asistencia al seguimiento docente basada en técnicas avanzadas de análisis. En *Actas del XIX Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2004)*, Barcelona, Spain.

Platt, J. C. (1998). Fast training of support vector machines using sequential minimal optimization. En Schoelkopf, B., Burges, C., i Smola, A., editors, *Advances in Kernel Methods-Support Vector Learning*, p. 41–65. MIT Press.

Plutchik, R. (1980). *Emotion: A Psychoevolutionary Synthesis*. Harper and Row, New York.

Plutchik, R. (1994). *The psychology and biology of emotion*. HarperCollins Publishers.

Plutchik, R. (2001). The nature of emotions. *American Scientist*, 89(4):344–350.

Polzehl, T., Sundaram, S., Ketabdar, H., Wagner, M., i Metze, F. (2009). Emotion classification in children's speech using fusion of acoustic and linguistic features. En *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, p. 340–343, Brighton, UK.

Polzin, T. i Waibel, A. (2000). Emotion-sensitive human-computer interface. En *Proceedings of the ISCA Workshop on Speech and Emotion*, Belfast, North Ireland.

Pous, M. i Ceccaroni, L. (2010). Multimodal interaction in distributed and ubiquitous computing. En *Proceedings of the 5th International Conference on Internet and Web Applications and Services (ICIW 2010)*, p. 457–462, Barcelona, Spain.

Prieto, M. A. (2008). *Blade Runner. Colección Making Of.* T&B Editores, Madrid.

Puigví, D., Jiménez, D., i Fernández, J. M. (1994). Parametrización de las pausas ortográficas en castellano. Aplicación a un conversor de texto a habla. *Procesamiento del Lenguaje Natural*, 15.

Přibil, J. i Přibilová, A. (2009). Spectral flatness analysis for emotional speech synthesis and transformation. En Esposito, A. i Vích, R., editors, *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions*, volum 5641 de *Lecture Notes in Computer Science*, p. 106–115. Springer Berlin Heidelberg.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Series in Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1ª edició.

Quinlan, R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

Rapcan, V., D'Arcy, S., Yeap, S., Afzal, N., Thakore, J., i Reilly, R. B. (2010). Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia. *Medical Engineering and Physics*, 32(9):1074–1079.

Reips, U. D. (2002). Standards for internet-based experimenting. *Experimental Psychology*, 49(4):243–256.

Ringeval, F. i Chetouani, M. (2007). Exploiting a vowel based approach for acted emotion recognition. En Esposito, A., Bourbakis, N. G., Avouris, N. M., i Hatzilygeroudis, I., editors, *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction, COST Action 2102 International Conference, Patras, Greece, October 29-31, 2007. Revised Papers*, volum 5042 de *Lecture Notes in Computer Science*, p. 243–254. Springer.

Rish, I. (2001). An empirical study of the Naïve-Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 3(22):41–46.

Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J., Bernadas, D., Oliver, J., i Longhi, L. (1999). Modelización acústica de la expresión emocional en el español. *Procesamiento del Lenguaje Natural*, 25:159–166.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Russell, J. A., Bachorowski, J., i Fernández-Dols, J. (2003). Facial and vocal expressions of emotion. *Annual Reviews of Psychology*, 54:329–349.

Russell, J. A., Lewicka, M., i Niit, T. (1989). A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology*, 57(5):848–856.

Ruvolo, P., Fasel, I., i Movellan, J. (2008). Auditory mood detection for social and educational robots. En *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2008)*, p. 3551–3556, California, USA.

Scherer, K. R. (1979). *Social markers in speech*, capítol Personality markers in speech, p. 147–209. Cambridge University Press, Cambridge.

Scherer, K. R. (1984). *Review of Personality and Social Psychology*, volum 5, capítol Emotion as a Multicomponent Process: A model and some cross-cultural data, p. 37–63. Sage, Beverly Hills, CA.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99:143–165.

Scherer, K. R. (1988). *Facets of Emotion: Recent Research*. Lawrence Erlbaum Associates Publishers, New Jersey.

Scherer, K. R. (1999). Appraisal theory. En Dalgleish, T. i Power, M., editors, *Handbook of Cognition and Emotion*, p. 637–663. John Wiley & Sons Ltd., Sussex, UK.

Scherer, K. R. (2000). *Introduction to social psychology: A European perspective*, capítol Emotion, p. 151–191. Blackwell, Oxford, 3ª edició.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227–256.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729.

Scherer, K. R. i Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, 1(4):331–346.

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological Review*, 61(2):81–88.

Schröder, M. (2004). *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis*. Tesi doctoral, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University.

Schröder, M. (2013). Rating test, Java software for designing and carrying out listening tests. En línea. Darrera visita el 4 de juny de 2013. Disponible en <http://sourceforge.net/projects/ratingtest/>.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., i Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. En *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH 2001)*, volum 1, p. 87–90, Aalborg, Denmark.

Schuller, B., Batliner, A., Steidl, S., i Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, In Press, Corrected Proof.

Schuller, B., Steidl, S., i Batliner, A. (2009). The Interspeech 2009 Emotion Challenge. En *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, p. 312–315, Brighton, UK.

Schuller, B., Villar, R., Rigoll, G., i Lang, M. (2005). Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. En *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, p. 325–328, Pennsylvania, USA.

Schweitzer, A. i Möbius, B. (2003). On the structure of internal prosodic models. En *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)*, Barcelona, Spain.

Sebe, N., Cohen, I., i Huang, T. S. (2005). *Handbook of Pattern Recognition and Computer Vision*, capítol Multimodal emotion recognition. World Scientific.

Sevillano, X., Melenchón, J., i Socoró, J. C. (2006). Análisis y síntesis audiovisual para interfaces multimodales ordenador-persona. En *Proceedings of the 7th Congreso Internacional de Interacción Persona-Ordenador (Interacción 2006)*, Puertollano, Ciudad Real, Spain.

Sezgin, M., Gunsel, B., i Kurt, G. (2012). Perceptual audio features for emotion detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2012:1–21.

Shami, M. i Verhelst, W. (2007). An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication*, 49(3):201–212.

Shaukat, A. i Chen, K. (2008). Towards automatic emotional state categorisation from speech signals. En *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008)*, p. 2771–2774, Brisbane, Australia.

Slaney, M. i McRoberts, G. (1998). Baby Ears: a recognition system for affective vocalizations. *1998 IEEE International Conference on Acoustics Speech and Signal Processing*, p. 985–988.

Steidl, S. (2009). *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Tesi doctoral, University of Erlangen-Nuremberg.

Stevenson, R. A., Mikels, J. A., i James, T. W. (2007). Characterization of the affective norms for english words by discrete emotional categories. *Behavior Research Methods*, 39(4):1020–1024.

Suwa, M., Sugie, N., i Fujimora, K. (1978). A preliminary note on pattern recognition of human emotional expression. En *Proceedings of the International Joint Conference on Pattern Recognition (IJCPR 1978)*, p. 408–410, Florida, USA.

Tolkien, J. R. R. (1954). *El Señor de los Anillos. La Comunidad del Anillo*. Ediciones Minotauro, 24ª edició. Traduït per Luis Domènech.

Trilla, A. i Alías, F. (2009). Sentiment classification in English from sentence-level annotations of emotions regarding models of affect. En *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, p. 516–519, Brighton, UK.

Truong, K. P. i Raaijmakers, S. (2008). Automatic recognition of spontaneous emotions in speech using acoustic and lexical features. En *Proceedings of the 5th International Workshop on Machine Learning for Multimodal Interaction (MLMI 2008)*, p. 161–172, Utrecht, The Netherlands.

Truong, K. P. i van Leeuwen, D. A. (2007). An open-set detection evaluation methodology applied to language and emotion recognition. En *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007)*, p. 338–341, Antwerp, Belgium. ISCA.

Valero, X. i Alías, F. (2012). Hierarchical classification of environmental noise sources considering the acoustic signature of vehicle pass-bys. *Archives of Acoustics*, 37(4):423–434.

Valstar, M. F., Gunes, H., i Pantic, M. (2007). How to distinguish posed from spontaneous smiles using geometric features. En *Proceedings of the 9th International Conference on Multimodal Interfaces (ICMI 2007)*, p. 38–45, New York, NY, USA. ACM.

Ververidis, D. i Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181.

Vlasenko, B., Schuller, B., Wendemuth, A., i Rigoll, G. (2007). Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing. En *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction (ACII 2007)*, p. 139–147, Berlin, Heidelberg. Springer-Verlag.

Vlasenko, B. i Wendemuth, A. (2009). Processing affected speech within human machine interaction. En *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, p. 2039–2042, Brighton, UK.

Vogt, T. i André, E. (2009). Exploring the benefits of discretization of acoustic features for speech emotion recognition. En *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, p. 328–331, Brighton, UK.

Vogt, T., André, E., i Bee, N. (2008). EmoVoice — A framework for online recognition of emotions from voice. En *Proceedings of the Workshop on Perception and Interactive Technologies for Speech-Based Systems*.

Wang, R. i Fang, B. (2008). Affective computing and biometrics based HCI surveillance system. En *Proceedings of the International Symposium on Information Science and Engineering (ISISE 2008)*, volum 1, p. 192–195, Shanghai, China.

Whissell, C. M. (1989). The dictionary of affect in language. *Emotion: Theory, Research, and Experience*, 4:113–131.

Whissell, C. M. (2008). A comparison of two lists providing emotional norms for english words (ANEW and the DAL). *Psychological Reports*, (102):597–600.

Williams, C. E. i Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America*, 52(4):1238–1250.

Witten, I. H. i Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, CA, 2ª edició.

Yildirim, S., Narayanan, S., i Potamianos, A. (2011). Detecting emotional state of a child in a conversational computer game. *Computer Speech and Language*, 25:29–44.

Zara, A., Maffiolo, V., Martin, J. C., i Devillers, L. (2007). Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics. En Paiva, A., Prada, R., i Picard, R. W., editors, *Affective Computing and Intelligent Interaction*, volum 4738 de *Lecture Notes in Computer Science*, p. 464–475. Springer Berlin Heidelberg.

Zeng, Z., Hu, Y., Roisman, G., Wen, Z., Fu, Y., i Huang, T. S. (2007). Audio-visual spontaneous emotion recognition. En Huang, T. S., Nijholt, A., Pantic, M., i Pentland, A., editors, *Artifical Intelligence for Human Computing*, volum 4451 de *Lecture Notes in Computer Science*, p. 72–90. Springer Berlin Heidelberg.

Zeng, Z., Pantic, M., Roisman, G. I., i Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58.

Zhang, T., Hasegawa-Johnson, M., i Levinson, S. E. (2006). Cognitive state classification in a spoken tutorial dialogue system. *Speech Communication*, 48(6):616–632.

Apèndix A

Aportacions

En aquest annex es resumeix la divulgació científica associada al present treball de tesi i la participació del seu autor en projectes de recerca i desenvolupament.

A.1 Publicacions científiques

El present treball de recerca ha proporcionat diferents aportacions d'interès per a la comunitat científica. Les principals idees, mètodes i resultats que són fruit de l'activitat investigadora presentada en aquesta tesi s'han exposat en diferents congressos i publicat en revistes d'àmbit internacional.

L'impacte del treball de recerca en la comunitat científica es pot resumir en les següents publicacions:

Congressos i conferències

1. Planet, S., Morán, J. A., i Formiga, L. (2006). Reconocimiento de emociones basado en el análisis de la señal de voz parametrizada. En Cunha, M. M. i Rocha, A., editors, *Actas da I Conferência Ibérica de Sistemas e Tecnologias de Informação (CISTI 2006)*, volum 2, p. 837–854, Ofir, Portugal

2. Iriondo, I., Planet, S., Socoró, J. C., i Alías, F. (2007c). Objective and subjective evaluation of an expressive speech corpus. En Chetouani, M., Hussain, A., Gas, B., Milgram, M., i Zarader, J. L., editors, *Advances in Nonlinear Speech Processing, International Conference on Nonlinear Speech Processing, NOLISP 2007, Paris, France, May 2007, Revised Selected Papers*, volum 4885 de *Lecture Notes in Artificial Intelligence*, p. 86–94. Springer

3. Iriondo, I., Planet, S., Alías, F., Socoró, J. C., i Martínez, E. (2007a). Validation of an expressive speech corpus by mapping automatic classification to subjective evaluation. En Sandoval, F., Prieto, A., Cabestany, J., i Graña, M., editors, *Computational and Ambient Intelligence, 9th International Work-Conference on Artificial Neural Networks, IWANN 2007, San Sebastián, Spain, June 2007, Proceedings*, volum 4507 de *Lecture Notes in Computer Science*, p. 646–653. Springer

4. Monzo, C., Alías, F., Iriondo, I., Gonzalvo, X., i Planet, S. (2007). Discriminating expressive speech styles by voice quality parameterization. En *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007)*, p. 2081–2084, Saarbrücken, Germany

5. Iriondo, I., Planet, S., Alías, F., Socoró, J. C., Monzo, C., i Martínez, E. (2007b). Expressive speech corpus validation by mapping subjective perception to automatic classification based on prosody and voice quality. En *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007)*, p. 2125–2128, Saarbrücken, Germany

6. Planet, S., Iriondo, I., Martínez, E., i Montero, J. A. (2008). TRUE: an online Testing platfoRm for mUltimedia Evaluation. En *Proceedings of the 2nd International Workshop on EMOTION: Corpora for Research on Emotion and Affect at the 6th Conference on Language Resources & Evaluation (LREC 2008)*, p. 61–65, Marrakech, Morocco

7. Planet, S., Iriondo, I., Socoró, J. C., Monzo, C., i Adell, J. (2009). GTM-URL contribution to the Interspeech 2009 Emotion Challenge. En *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, p. 316–319, Brighton, UK

8. Planet, S. i Iriondo, I. (2011b). Spontaneous children's emotion recognition by categorical classification of acoustic features. En Rocha, A., Gonçalves, R., Pérez Cota, M., i Reis, L. P., editors, *Actas da 6a Conferência Ibérica de Sistemas e Tecnologias de Informação (CISTI 2011)*, volum 1, p. 594–599, Chaves, Portugal

9. Planet, S. i Iriondo, I. (2011a). Improving spontaneous children's emotion recognition by acoustic feature selection and feature-level fusion of acoustic and linguistic parameters. En Travieso, C. i Alonso, J., editors, *Advances in Nonlinear Speech Processing, 5th International Conference on Nonlinear Speech Processing, NOLISP 2011, Las Palmas de Gran Canaria, Spain, November 2011, Proceedings*, volum 7015 de *Lecture Notes in Computer Science*, p. 88–95. Springer Berlin/Heidelberg

10. Planet, S. i Iriondo, I. (2012c). Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition. En Rocha, A., Calvo Manzano, J. A., Reis, L. P., i Pérez Cota, M., editors, *Actas de la 7a Conferencia Ibérica de Sistemas y Tecnologías de la Información (CISTI 2012)*, volum 1, p. 199–204, Madrid, Spain

Capítols de llibre

1. Iriondo, I., Planet, S., Alías, F., Socoró, J. C., i Martínez, E. (2008). Emulating subjective criteria in corpus validation. En Rabuñal, J. R., Dorado, J., i Pazos, A., editors, *Encyclopedia of Artificial Intelligence*, p. 541–546. Information Science Reference

Revistes

1. Iriondo, I., Planet, S., Socoró, J. C., Martínez, E., Alías, F., i Monzo, C. (2009). Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification. *Speech Communication*, 51(9):744–758. Special issue on non-linear and conventional speech processing - NOLISP 2007

2. Planet, S. i Iriondo, I. (2012a). Children's emotion recognition from spontaneous speech using a reduced set of acoustic and linguistic features. *Cognitive Computation*

3. Planet, S. i Iriondo, I. (2012b). Comparative study on feature selection and fusion schemes for emotion recognition from speech. *International Journal of Interactive Multimedia and Artificial Intelligence*, 1(6):44–51. Special Issue on Intelligent Systems and Applications

(Planet *et al.*, 2006) és la primera aportació relacionada amb la línia de recerca en reconeixement afectiu automàtic basada en el corpus de veu actuada ESDLA, tal com s'exposa en la secció 5.1.

(Iriondo *et al.*, 2007c), (Iriondo *et al.*, 2007a), (Iriondo *et al.*, 2007b), (Iriondo *et al.*, 2008) i (Iriondo *et al.*, 2009) són publicacions relacionades amb la validació subjectiva automàtica del corpus de veu actuada BDP-UAB, tal com s'exposa en la secció 5.2. Així mateix, la publicació (Iriondo *et al.*, 2009) va ser tercera finalista del premi al millor article l'any 2009 de la *Red Temática en Tecnologías del Habla*[1].

(Monzo *et al.*, 2007) és una col·laboració en l'anàlisi de paràmetres de VoQ com a únics elements discriminants per abordar un estudi de reconeixement afectiu automàtic.

(Planet *et al.*, 2008) presenta una versió inicial de la plataforma de test subjectius TRUE, l'última versió del qual es descriu en l'apèndix B.

(Planet *et al.*, 2009) detalla la primera aproximació al reconeixement afectiu automàtic basat en el corpus de veu espontània FAU Aibo en el marc de l'*Interspeech Emotion Challenge 2009* (Schuller *et al.*, 2009), tal com s'explica en la secció 6.2.1. Els resultats presentats van obtenir la primera posició en el desafiament de paràmetres i en segona posició en el desafiament de classificadors de l'*Emotion Challenge*. Per la seva banda, la publicació (Planet i Iriondo, 2011b), presenta un treball derivat directament de l'anterior emprant una parametrització millorada i aportant noves propostes de classificació, tal com es recull en la secció 6.2.2.

(Planet i Iriondo, 2011a), (Planet i Iriondo, 2012c), (Planet i Iriondo, 2012a) i (Planet i Iriondo, 2012b) recullen els estudis realitzats a nivell acústic i lingüístic per al reconeixement afectiu automàtic en veu espontània, així com els experiments de fusió a tant a nivell de paràmetres com a nivell de classificadors, detallats en les seccions 6.3 i 6.4.

A.2 Altres publicacions

En els inicis de l'etapa investigadora es van dur a terme treballs que finalment no van formar part de la línia de recerca principal, però que van permetre tenir una primera presa de contacte amb la metodologia científica necessària per a la realització de la tesi i algunes nocions de temes vinculats. A continuació es presenten, en ordre cronològic, les publicacions que es van derivar.

[1] http://lorien.die.upm.es/ lapiz/rtth

Congressos i conferències

1. Planet, S., Morán, J. A., Monzo, C., i Gonzalvo, X. (2004). Asistencia al seguimiento docente basada en técnicas avanzadas de análisis. En *Actas del XIX Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2004)*, Barcelona, Spain

2. Monzo, C., Morán, J. A., Planet, S., i Gonzalvo, X. (2004). Protocolo DS-CDMA orientado a la gestión de un auditorio. En *Actas del XIX Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2004)*, Barcelona, Spain

3. Gonzalvo, X., Morán, J. A., Monzo, C., i Planet, S. (2005). Entorno para el aprendizaje automático de las estrategias de diálogo. En *Actas del XX Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2005)*, Gandia, Spain

(Planet *et al.*, 2004) és una primera publicació relacionada amb tècniques d'aprenentatge artificial aplicades en l'àmbit docent. Va suposar una primera presa de contacte amb aquestes tècniques, les quals van ser aplicades posteriorment en aquesta tesi tot i que en un àmbit diferent.

Les publicacions (Monzo *et al.*, 2004) i (Gonzalvo *et al.*, 2005) es corresponen amb col·laboracions inicials amb membres de el GPMM.

A.3 Comitès tècnics

A més de les aportacions realitzades com a autor, s'ha format part del programa del comitè tècnic de les següents publicacions i esdeveniments científics:

1. ICME 2011. 2011 International Conference on Multimedia and Expo.

2. ICME 2012. 2012 International Conference on Multimedia and Expo.

3. CISTI 2012. 7th Iberian Conference on Information Systems and Technologies.

4. Revisor del *Special Issue Sensing Emotion and Affect, Facing Realism in Speech Processing*, de la revista *Speech Communication*, volum 53, issues 9-10, pàg. 1059–1228 (novembre-desembre de 2011), editat per Björn Schuller, Stefan Steidl i Anton Batliner.

5. ICME 2013. 2013 International Conference on Multimedia and Expo.

A.4 Projectes de recerca i desenvolupament

En aquesta secció es descriuen els projectes de recerca i desenvolupament en els quals aquesta tesi es va veure emmarcada. Aquests projectes es van realitzar com a membre de el GTM.

SALERO — Semantic AudiovisuaL Entertainment Reusable Objects (FP6/2004/IST/4)

Projecte finançat pel VI Programa Marco de la Unió Europea (2006-2009)[2]. Van participar tretze socis entre empreses i centres de recerca. El seu objectiu era facilitar la creació de nous productes multimèdia com a jocs, pel·lícules o programes de televisió, fent-la millor, més ràpida i amb menys costos gràcies a la combinació de gràfics per ordinador, tecnologia de la parla i el llenguatge, web semàntica i cerques basades en contingut.

SAVE — Síntesis Audiovisual Expresiva (TEC2006-08043/TCM)

El projecte *Síntesis Audiovisual Expresiva* (SAVE) és un projecte de R+D cofinançat pel *Ministerio de Educación y Ciencia* espanyol amb l'objectiu de generar caps parlants d'avatars virtuals capaços de transmetre estats d'ànim mentre parlen.

INREDIS — INterfaces de Relación entre el Entorno y las personas con DIScapacidad (CEN-20072011)

El projecte INREDIS és un projecte de *Consorcios Estratégicos Nacionales de Investigación Técnica* (CENIT) de R+D+i que s'inscriu en la iniciativa del govern espanyol INGENIO 2010 i que és gestionada pel *Centro de Desarrollo Tecnológico Industrial* (CDTI)[3]. Va suposar el desenvolupament de la recerca en l'àmbit de les tecnologies accessibles i interoperables durant el període 2007-2010. El seu objectiu era el desenvolupament de tecnologies de base que permetessin crear canals de comunicació i interacció entre les persones amb algun tipus de necessitat especial i el seu entorn. Va ser liderat per Technosite, empresa tecnològica del Grup Fundosa, dependent de la Fundació ONCE.

evMIC — Entornos virtuales Multimodales, Inmersivos y Colaborativos (TSI-020301-2009-25)

El projecte *entornos virtuales Multimodales, Inmersivos y Colaborativos* (evMIC) és un projecte de R+D+i cofinançat pel *Ministerio de Industria, Turismo y Comercio* (MITYC). Segons la presentació oficial del propi projecte[4], aquest s'emmarca dins de l'*Internet del futur* i pretén investigar i generar solucions principalment en l'*Internet del Coneixement i dels Continguts*, així com en l'*Internet de les Persones* i, en menor mesura, en l'*Internet dels Serveis*. El seu objectiu principal és crear una plataforma interoperable, centrada en l'usuari, capaç de permetre la creació d'entorns virtuals d'aprenentatge.

[2] http://www.salero.info
[3] http://www.inredis.es
[4] http://www.evmic.es

A.5 Participació en esdeveniments

El futuro de los sistemas de diálogo (Soria, 11 de juliol de 2005)

Exposició dels primers avanços realitzats en la línia de recerca de reconeixement afectiu automàtic, dins del marc del curs de Tecnologies Lingüístiques dirigit pel Dr. Joaquim Llisterri i organitzat per la *Fundación Duques de Soria*.

Apèndix B

Plataforma en línia de tests per a l'avaluació d'estímuls multimèdia

Testing platfoRm for mUltimedia Evaluation (TRUE) és una plataforma en línia de tests per a l'avaluació d'estímuls multimèdia desenvolupada dins de l'àmbit d'aquesta tesi, orientada a l'avaluació d'estímuls audiovisuals, gràfics i/o de text. El focus d'atenció se centra en l'avaluació de corpus multimèdia si bé pot emprar-se per analitzar estímuls multimèdia en general. Ofereix una gran flexibilitat per crear tests destinats a identificació emocional o avaluació de la qualitat de sistemes de síntesi audiovisual, per exemple. Els resultats poden emprar-se amb diferents finalitats depenent dels objectius del test, per exemple per validar el contingut emocional d'un arxiu multimèdia o per mesurar l'expressivitat d'elements audiovisuals sintetitzats. TRUE abasta totes les etapes relacionades amb el cicle de vida de les proves subjectives, des de la seva creació fins a la recuperació dels seus resultats, i permet als avaluadors realitzar els tests des de qualsevol dispositiu amb connexió a Internet a través d'un navegador web. La simplificació de la realització de les proves ajuda a minimitzar els efectes de la fatiga sobre les mateixes i també permet als investigadors focalitzar els seus esforços en l'anàlisi dels resultats en lloc de destinar esforços a la seva supervisió. Els detalls de la plataforma TRUE es van publicar per primera vegada en (Planet *et al.*, 2008).

B.1 Descripció

Els tests subjectius poden aplicar-se en diversos àmbits tals com la validació expressiva de corpus audiovisuals emocionals, l'etiquetatge d'elements de corpus audiovisuals, gràfics o de text, i l'avaluació de sistemes de síntesis puntuant estímuls individuals o comparant elements generats mitjançant diferents tècniques. No obstant això, els test subjectius poden ser útils en qualsevol altre tipus d'estudis en els quals es requereixi una avaluació sobre la base del criteri propi d'un conjunt de persones.

B.1.1 Objectius

Els tests subjectius en general poden presentar diversos inconvenients, entre d'altres els següents:

- Alta especifitat. Els tests solen dissenyar-se per adequar-se a les característiques particulars d'estudis concrets cosa que els fa poc generalitzables per a estudis posteriors.
- Baixa reutilitzabilitat. El disseny *ad hoc* dels tests pot dificultar que puguin emprar-se en estudis posteriors.
- Fatiga en els usuaris. Els tests poden resultar sovint llargs i tediosos, la qual cosa pot influir en els resultats dels mateixos.

- Lloc i condicions de realització dels tests. Solen suposar el desplaçament de l'avaluador a un lloc concret per realitzar la prova de forma individual o col·lectiva, la qual cosa sol requerir de l'adequació d'un lloc físic per a la seva realització i la supervisió d'un responsable.
- Problemes de procediment. Si els tests es realitzen en paper la recopilació de dades pot resultar laboriosa i incórrer en errors.
- Audiència. Sol resultar complicat reunir un elevat nombre de subjectes disposats a realitzar el test de forma adequada i que responguin al perfil convenient, ja sigui homogeni o heterogeni.

TRUE és una eina en línia que a través d'una interfície web proporciona una plataforma única per dissenyar i realitzar tests subjectius, la qual cosa aporta:

- Generalització. Es poden crear tests completament personalitzats i específics que poden modificar-se posteriorment per adaptar-los a nous experiments.
- Reutilitzabilitat. TRUE proporciona plantilles per a tests estàndard així com la possibilitat de generar nous tests a partir d'uns altres ja realitzats.
- Sistema orientat a l'usuari. L'usuari pot detenir en qualsevol moment el test i continuar-ho més endavant per evitar que la fatiga pugui repercutir negativament en els resultats.
- Sistema en línia. Els usuaris poden realitzar els tests des de qualsevol dispositiu amb connexió a Internet a través d'un navegador web. No és necessari un supervisor que controli la prova i les instruccions es poden proporcionar a l'inici de la mateixa de forma automàtica.
- Procediment automàtic. TRUE permet dissenyar els tests i recuperar els resultats en qualsevol moment des de la mateixa eina. Ja que la informació es conserva en una base de dades informàtica les dades poden recuperar-se en qualsevol moment amb la màxima fiabilitat.
- Fàcil accés per a una àmplia audiència. Donades les facilitats per als usuaris resulta més senzill aconseguir un major nombre de participants en qualsevol lloc geogràfic.

B.1.2 Esquema bàsic de funcionament

TRUE proporciona una eina web que permet crear i configurar tests a través de formularis perquè usuaris remots puguin avaluar diversos estímuls. Les respostes dels avaluadors s'emmagatzemen en una base de dades i poden ser recuperades en qualsevol moment mitjançant una descàrrega directa dels mateixos en diferents formats. Els estímuls són arxius multimèdia que poden emmagatzemar-se en el mateix servidor web en

el qual s'allotja TRUE o en un altre extern ja que TRUE enllaça aquests arxius a través de la definició del test i els mostra de forma adequada i ordenada. Aquest esquema de funcionament es mostra en la figura B.1.

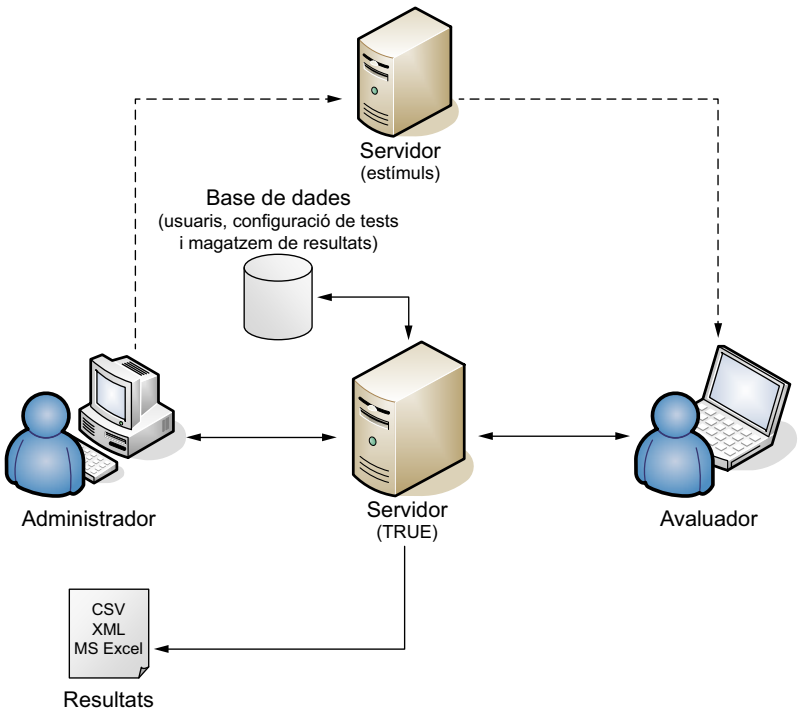


Figura B.1: Esquema bàsic de funcionament de la plataforma TRUE.

La figura B.1 mostra dos actors diferents dins de l'esquema de funcionament de TRUE. L'usuari *Administrador* és el que té la capacitat de crear i gestionar els seus propis tests. Aquest usuari ha de registrar-se prèviament en el sistema. L'usuari *Avaluador* és el que realitza els tests, avaluant els diferents estímuls que es mostren segons la configuració dels mateixos. Existeix una tercera figura que és la que té la possibilitat de gestionar tots els tests i tots els usuaris *Administrador* registrats en TRUE. Aquesta tercera figura és l'administrador general del sistema i és qui instal·la i configura tots els detalls de funcionament de la plataforma.

La figura B.2 mostra el menú principal de la plataforma TRUE d'un administrador general del sistema. Amb l'excepció de la secció de gestió del sistema (TRUE System Manager), aquesta pantalla és igual per a la resta d'usuaris de la plataforma, la qual dóna accés a les eines de gestió dels tests (creació, edició, recuperació de resultats i eliminació) i a la gestió de les dades pròpies de l'usuari.

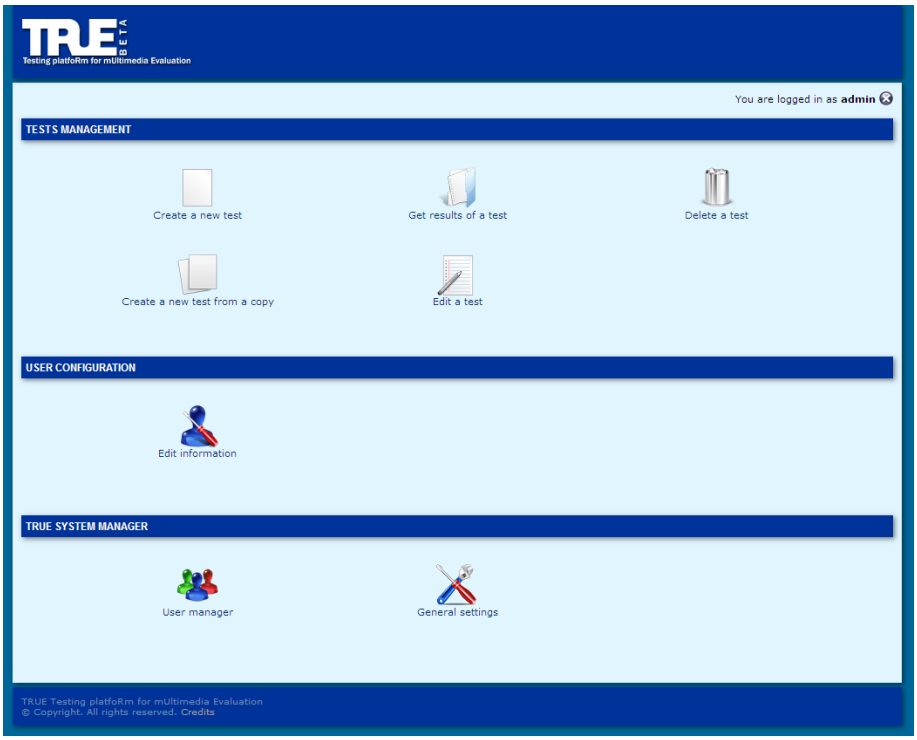


Figura B.2: Menú principal de la plataforma TRUE per a un administrador general del sistema.

B.1.3 Característiques generals

TRUE permet considerar diferents modalitats (àudio, vídeo, gràfics i text) com a estímuls per ser avaluats a través de la seva adequada representació en el navegador web de l'usuari. Els estímuls gràfics i de text s'incrusten a la pròpia pàgina web com un element HTML. Els estímuls d'àudio i de vídeo s'enllacen per ser reproduïts per un reproductor multimèdia adequat: o bé el propi del dispositiu de l'usuari o bé s'ofereix la possibilitat d'incrustar un reproductor multiplataforma[1]. El disseny dels elements en pantalla és configurable durant la creació del test, podent escollir el nombre d'estímuls simultanis a mostrar, la possibilitat de mostrar un estímul de referència en pantalla, el nombre de qüestions a realitzar a l'avaluador i les opcions de resposta i podent determinar, també, la disposició d'aquests elements. L'usuari pot introduir la resposta a través de camps d'elecció múltiple, entrada lliure de text i/o a través d'interfícies gràfiques d'usuari interactives.

La figura B.3 mostra quatre exemples de test diferents, considerant diferents tipus d'estímuls: gràfics, d'àudio, de vídeo i de text. Els estímuls audiovisuals apareixen incrustats mitjançant un reproductor multiplataforma, mentre que els estímuls gràfics i de text s'incrusten directament com a elements HTML. En aquests quatre exemples es mostren diferents tipus d'entrades d'avaluacions a través de camps d'elecció múltiple amb disposició horitzontal i vertical de les opcions i camps d'entrada de text lliure.

[1]En la versió actual aquest reproductor és *JW Player* (http://www.longtailvideo.com/jw-player).

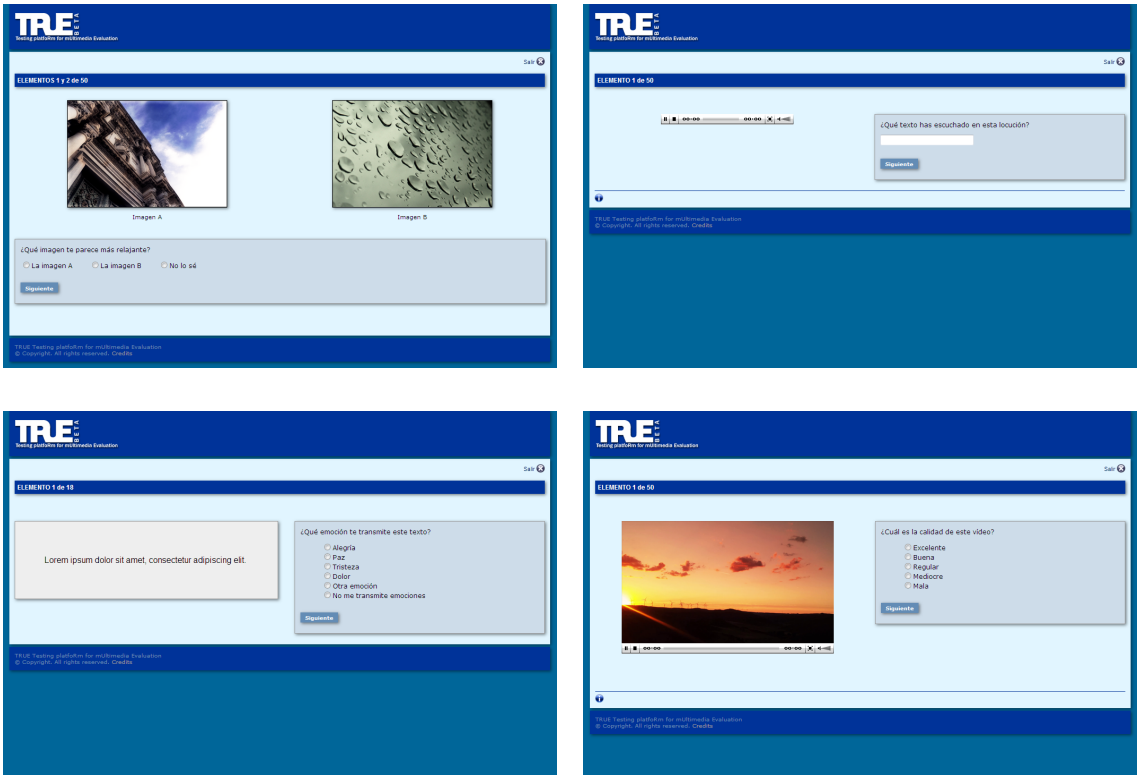


Figura B.3: Quatre tests de mostra creats amb la plataforma TRUE. En sentit horari, des de la cantonada superior esquerra: un test gràfic de dos estímuls, un test d'un estímul d'àudio, un test de vídeo i un test d'estímul textual.

De forma general, un administrador pot seleccionar diferents opcions comunes per a tots els tests, incloent les següents:

- Idioma de la interfície. Per defecte, TRUE inclou l'espanyol, català i anglès com a idiomes predeterminats, si bé es poden incloure fàcilment nous idiomes mitjançant arxius de traducció.
- Aparença. Determinació del full d'estils més adequat per a cada test, escollint la millor combinació de colors i tipografies.
- Disposició d'elements en pantalla. Estímul/s i camp/s d'entrada d'avaluacions poden disposar-se de forma horitzontal o vertical.
- Grandària de la font. Els textos poden mostrar-se amb grandària petita, normal o gran.
- Durada del test. Defineix el temps durant el qual el test estarà actiu i disponible per a la seva avaluació. Després el test passarà a estar inactiu.

Per evitar errors inicials en l'avaluació dels estímuls s'ofereix la possibilitat de crear,

juntament amb les instruccions del test, diferents tipus de demostracions a l'inici del mateix. L'objectiu d'aquestes demostracions és mostrar una sèrie d'estímuls als avaluadors abans del començament de test perquè es familiaritzin amb ells. Aquestes demostracions poden ser completes o cegues i consisteixen a mostrar alguns estímuls concrets especificant o no, respectivament, la resposta que l'usuari hauria d'indicar en cada cas (o algun altre tipus d'indicació). Les instruccions i la demostració poden incloure's a través de la plantilla predeterminada proporcionada per TRUE o bé poden crear-se de forma personalitzada a través de l'editor HTML proporcionat en el formulari de creació. La figura B.4 mostra un exemple de demostració inicial completa en la qual, juntament amb les instruccions del test, l'avaluador pot visualitzar cinc imatges que estan associades a un text descriptiu. Una demostració cega no inclouria els textos descriptius associats. En aquest cas es tracta d'un test on els estímuls són de tipus gràfic i l'idioma configurat és l'anglès.

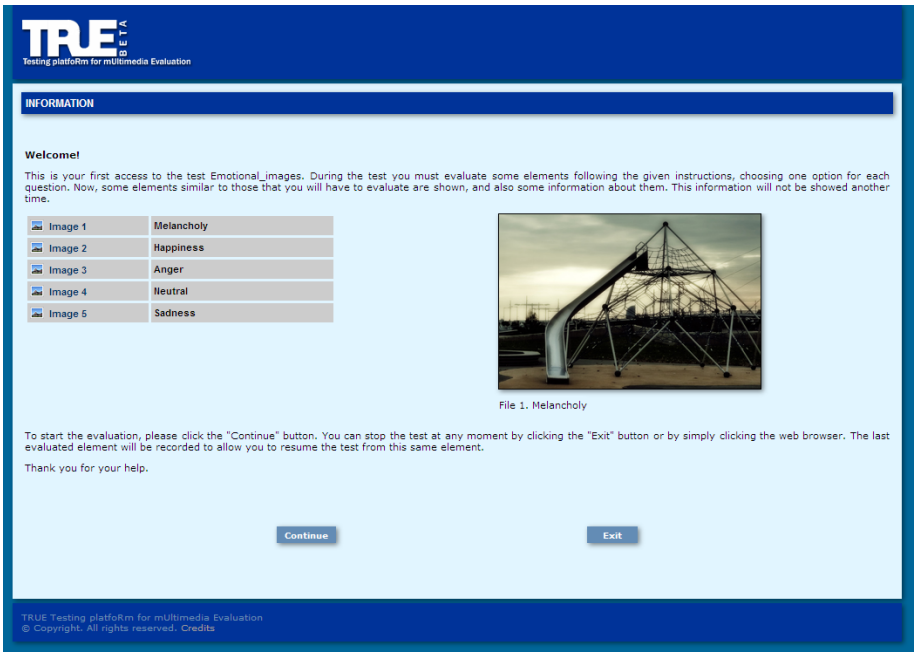


Figura B.4: Exemple de demostració inicial completa en un test gràfic en anglès.

D'altra banda, al final del test pot incloure's una enquesta totalment personalitzable en la qual el creador del test pot recaptar informació sobre l'avaluador. La figura B.5 mostra un exemple d'enquesta al final d'un test, la qual inclou quatre camps personalitzats d'entrada de dades. En aquest cas les entrades són de tipus de text lliure, botons de radi i un àrea de text. Cadascun dels camps es poden configurar per permetre diverses longituds màximes permeses de text o múltiples opcions. També existeix la possibilitat d'incloure una llista desplegable d'opcions. La informació introduïda apareixerà lligada a les respostes dels avaluadors en el moment de descarregar els resultats del test. Tant les respostes del test, la informació dels avaluadors introduïda en l'enquesta, l'últim estímul avaluat així com el temps emprat per cada avaluador per finalitzar-ho són dades que poden ser obtingudes en qualsevol moment en format *Comma-Separated Values* (CSV), XML o com a full de càlcul de Microsoft Excel.

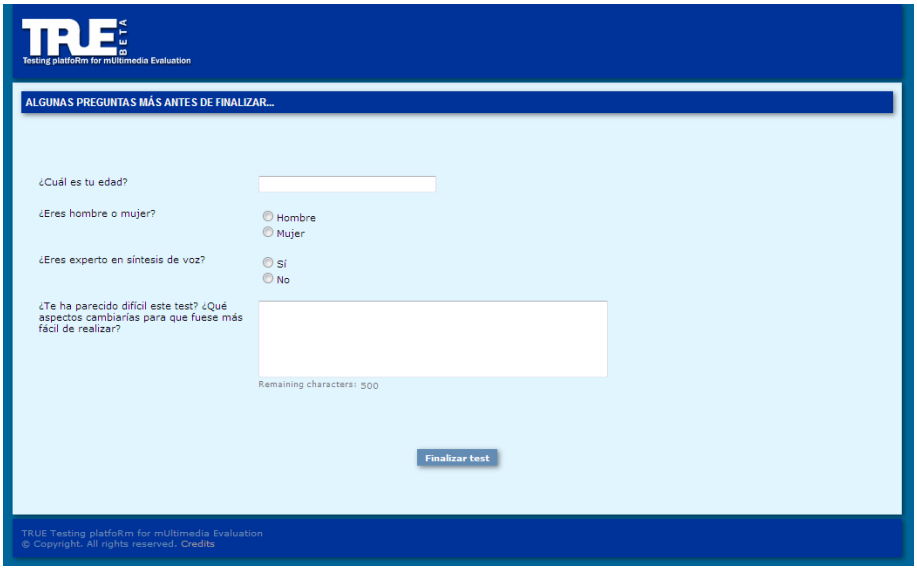


Figura B.5: Exemple d'enquesta al final d'un test.

B.1.4 Tipus de tests implementats

TRUE permet crear tests de diversos estímuls i diverses consultes a l'avaluador (entrades d'avaluació). Aquestes entrades d'avaluació poden configurar-se creant una llista d'elecció múltiple o bé introduint un camp de text lliure, donant resposta a la consulta que el dissenyador del test consideri necessària.

Amb l'objectiu d'aportar eines estàndard per als dissenyadors dels tests subjectius, TRUE inclou plantilles per crear alguns tests específics. Aquestes plantilles permeten crear tests *Mean Opinion Score* (MOS), *Comparison Mean Opinion Score* (CMOS) i *Degradation Mean Opinion Score* (DMOS), tal com defineix ITU-T (1996). Els tests MOS es refereixen a la valoració de la qualitat percebuda d'un estímul mitjançant una escala numèrica; els tests CMOS realitzen una comparació entre estímuls, a nivell de qualitat dels mateixos; i els tests DMOS són similars als CMOS però realitzen una avaluació de la degradació d'un estímul pel que fa a un altre. Aquests tests estan vinculats amb la qualitat de sistemes de transmissió i la seva capacitat per no degradar un senyal.

Tanmateix, TRUE permet incrustar interfícies gràfiques interactives per avaluar els estímuls. Aquestes interfícies estan programades mitjançant tecnologia Flash i permeten configurar tests específics. La versió actual de TRUE inclou una implementació de la interfície SAM, segons la descripció gràfica de Bradley i Lang (1994), destinada a l'avaluació emocional dels estímuls. Així mateix inclou una llista de 121 termes afectius en anglès en forma de llista i una botonera de cinc botons configurables d'ús general.

La figura B.6 mostra un esquema de configuració general d'un test. Es representa la possibilitat de configurar un test mostrant els elements en vertical o en horitzontal, incloent entrades d'avaluació d'elecció múltiple, text lliure i interfícies gràfiques avançades.

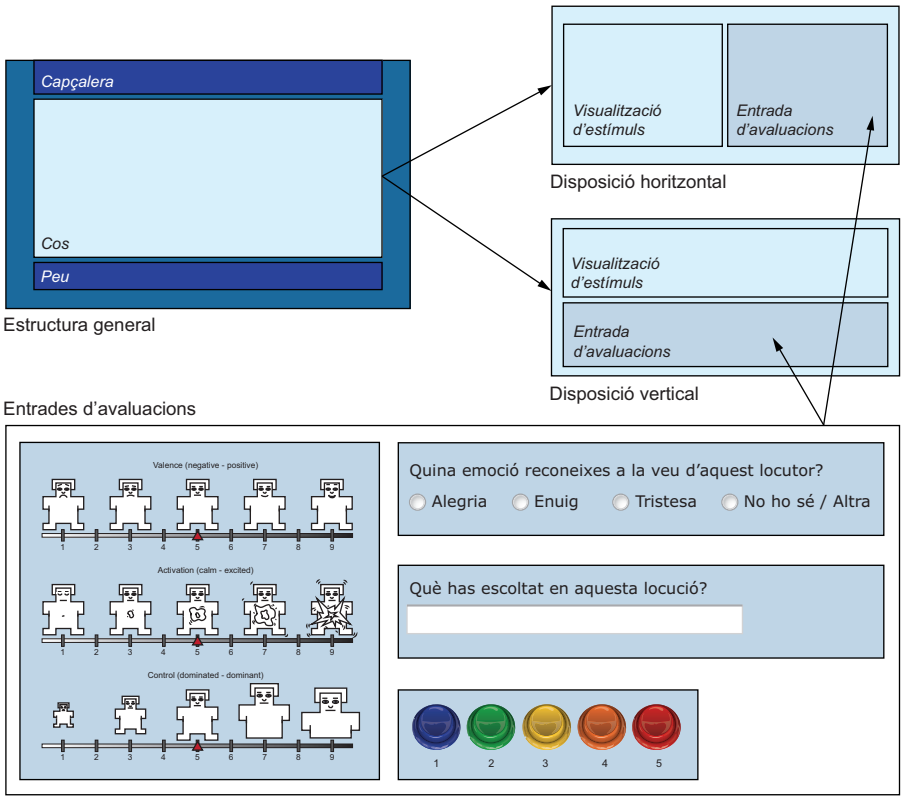


Figura B.6: Esquema de configuració d'un test. La plantilla general permet ser adaptada per mostrar els estímuls i les entrades d'avaluació en diferents disposicions. Es mostren alguns exemples d'entrades d'avaluació (interfície SAM, entrada d'elecció múltiple, entrada de text lliure i botons configurables d'ús general).

B.1.5 Diferències amb altres plataformes existents

TRUE difereix d'altres sistemes de tests en què no està dissenyat per a un tipus específic d'estímul, sinó que admet diferents tipus de formats multimèdia, i en la seva flexibilitat.

A diferència de (Irtel, 2007) i (NBS, 2013), TRUE no abasta un ampli ventall d'àrees en psicologia sinó que se centra en l'avaluació d'elements multimèdia procedents de corpus. Donada la seva focalització i que s'orienta en la creació de tests en línia, els seus procediments són més senzills que els de d'eines com (Empirisoft, 2011) i, a més, permet arribar a una audiència més àmplia que amb els tests que es realitzen mitjançant formularis en paper o a través d'un ordinador local, com (Schröder, 2013).

Les possibles inconsistències que poden derivar-se per l'ús d'una eina en línia poden minimitzar-se introduint elements de control en els tests, els quals poden permetre avaluar la coherència dels avaluadors. No obstant això, els tests en línia, igual que altres mètodes d'experimentació a través d'Internet, no estan exempts de discussió referent a la seva metodologia (Reips, 2002), la qual cosa ha de ser tinguda en consideració pels au-

tors dels tests en el moment de dissenyar-los per minimitzar els inconvenients que poden derivar-se d'ells i aprofitar al màxim els seus avantatges.

B.2 Implementació

TRUE és un servei web implementat en Java que fa ús de les tecnologies Java Servlets i *Java Server Pages* (JSP). La implementació en Java fa que el codi pugui ser executat sobre qualsevol sistema operatiu donada la seva característica de codi multiplataforma i facilita la seva distribució en forma d'un únic arxiu d'aplicació web (WAR). Aquest arxiu únic inclou totes les classes relatives a la gestió de test i usuaris així com tots els elements de la interfície web.

L'execució de TRUE pot realitzar-se en qualsevol ordinador que disposi d'un servidor web compatible amb la tecnologia Java com, per exemple, Apache Tomcat[2], simplement instal·lant l'arxiu d'aplicació web i accedint a l'aplicació a través d'un navegador web per a la seva correcta configuració inicial.

Per emmagatzemar totes les dades que fa servir TRUE, incloent les referents a la informació dels usuaris registrats, la configuració dels tests i els seus resultats, l'aplicació requereix d'accés a un servidor MySQL[3] que pot estar instal·lat en el mateix ordinador en el qual s'ha instal·lat TRUE o bé en un de remot. Els resultats dels test, disponibles per a la descàrrega per part dels seus creadors, estan disponibles en els formats CSV, XML i Microsoft Excel, i s'obtenen directament de la consulta i traducció (de forma transparent per a l'usuari final) de la informació emmagatzemada en la base de dades en MySQL.

La interfície web està programada mitjançant codi HTML i Javascript, incrustat a les pàgines JSP, cuidant que la seva visualització sigui correcta en diferents navegadors web. Per unificar l'aparença de tota la interfície i permetre la seva fàcil modificació i adaptació s'empren fulls d'estil en cascada. Així mateix, per millorar la interacció i la velocitat d'algunes funcions de la plataforma TRUE, alguns elements estan implementats mitjançant tecnologia *Asynchronous JavaScript And XML* (AJAX).

B.3 Treballs relacionats amb TRUE

TRUE va néixer amb la finalitat de validar el corpus de veu actuada ESDLA (vegeu la secció 3.4.1) dins d'un experiment de reconeixement afectiu automàtic (Planet *et al.*, 2006). L'objectiu d'aquesta validació era comprovar la taxa de reconeixement de les emocions preteses en les locucions del corpus per part dels avaluadors humans per, posteriorment, contrastar-la amb la taxa de reconeixement aconseguida pel sistema de reconei-

[2] http://tomcat.apache.org/
[3] http://www.mysql.com/

xement automàtic (més detalls d'aquest experiment en la secció 5.1.1).

En un sentit més ampli, TRUE s'empra en els treballs realitzats per Iriondo *et al.* (2007c,a,b) els quals pretenen la validació automàtica del corpus BDP-UAP (vegeu la secció 3.4.2) mitjançant el mapatge dels criteris subjectius en algorismes de classificació automàtica. Aquests criteris subjectius s'extreuen a partir de l'avaluació d'un fragment del corpus BPD-UAB mitjançant TRUE.

A diferència dels treballs anteriors, en els quals TRUE s'empra per avaluar corpus de veu, en (Iriondo *et al.*, 2007d) els tests estan relacionats amb la qualitat de locucions expressives de veu sintètica, així com també en els treballs realitzats per Gonzalvo *et al.* (2007a,b) en els quals s'empra amb la mateixa finalitat. En els tres casos, l'avaluació subjectiva es duu a terme mitjançant tests MOS.

TRUE també ha estat utilitzat per a tasques d'anàlisis i síntesi audiovisual per Sevillano *et al.* (2006) i per Melenchón (2007).

Montero *et al.* (2007) mostren la flexibilitat de la plataforma TRUE emprant-la en la recerca de noves metodologies docents. En el seu treball, els tests subjectius s'utilitzen per recollir coneixement expert de professors que serveix per modelar un sistema de lògica difusa capaç d'avaluar el rendiment en equip d'un grup d'estudiants d'enginyeria.

Sense que aquesta sigui una llista exhaustiva, altres treballs en els quals s'ha emprat la plataforma TRUE són els realitzats per Francisco (2008), Formiga *et al.* (2010), Pérez-Espinosa *et al.* (2011), Francisco *et al.* (2012), Calzada i Socoró (2012) i Valero i Alías (2012).

Apèndix C

Aplicacions de reconeixement afectiu automàtic

Aquest apèndix descriu dues aplicacions de programari dissenyades en el marc de la present tesi doctoral que implementen eines de reconeixement afectiu automàtic. Ambdues es basen en la creació de models incrementals, és a dir, que poden ser entrenats de forma iterativa per crear models cada vegada més precisos. No obstant això, la primera aplicació permet realitzar aquest entrenament de forma interactiva mitjançant una interfície gràfica mentre que la segona s'implementa com un mòdul que pot ser fàcilment incrustat en una aplicació externa una vegada el model ha estat entrenat convenientment. Així mateix, la segona aplicació estén els processos de la primera creant models de reconeixement basats tant en el senyal de veu com en l'anàlisi de l'expressivitat facial de l'usuari.

C.1 Aplicació d'anàlisi afectiva basada en aprenentatge interactiu incremental

L'aplicació detallada en aquesta secció implementa un sistema que permet a l'usuari integrar en una mateixa interfície gràfica tots els mecanismes necessaris per realitzar un reconeixement afectiu basat en senyal de veu. Aquest procés és interactiu i incremental mentre que el model de reconeixement es crea sobre la base de les indicacions de l'usuari, al que s'adapta de forma progressiva aprenent les característiques de la seva expressivitat afectiva.

Actualment aquesta aplicació és completament operativa si bé es tracta d'una versió preliminar de desenvolupament i per tant de característiques limitades. No obstant això, el disseny de la mateixa, permet que pugui actualitzar-se en versions posteriors. Actualment està plantejada per estudiar el reconeixement afectiu automàtic i no per ser usada com una aplicació de reconeixement genèric.

C.1.1 Descripció

L'aplicació d'anàlisi afectiva automàtica basada en aprenentatge interactiu incremental és una aplicació destinada a crear models de reconeixement afectiu a partir del senyal de veu d'un usuari, podent-se crear models diferents per a diferents usuaris. Aquests models es creen mitjançant un procés d'aprenentatge interactiu a través del qual són capaços d'assimilar les característiques de l'expressivitat afectiva de l'usuari. El procés d'aprenentatge és interactiu i incremental de manera que el model és més precís a mesura que l'usuari utilitza l'aplicació. Al mateix temps, aquesta aplicació permet la creació d'un corpus de veu i un model de reconeixement en paral·lel.

El model creat consisteix, bàsicament, en un classificador entrenat sobre la base de diverses locucions d'àudio enregistrades per l'usuari. Aquestes locucions, no obstant això, no formen part d'un corpus sinó que s'afegeixen seqüencialment a mesura que van sent

gravades i analitzades. Així, el model original manca de coneixement i va sent entrenat amb cada nova frase que es grava i analitza a través de l'aplicació.

Una vegada el model ha estat creat, l'usuari pot activar el mode de prova de l'aplicació de manera que el model ja no és modificat sinó que s'empra únicament amb la finalitat d'identificar els estats afectius de les noves locucions.

L'aplicació posseeix una interfície gràfica que centralitza cadascun dels passos de creació del model, oferint llibertat a l'usuari per intervenir en cadascun d'ells segons ho cregui convenient, tal com es descriu a continuació.

C.1.2 Funcionament de l'aplicació

El funcionament d'aquesta aplicació està inspirat en el proposat per Oudeyer (2003). La figura C.1 mostra el diagrama de flux de l'aplicació. El procediment consisteix en l'enregistrament d'un arxiu d'àudio a través del micròfon connectat a la targeta de so de l'ordinador. L'usuari ha d'escriure el text de la frase que ha gravat[1] perquè l'aplicació pugui realitzar la transcripció fonètica i associar-la a l'arxiu d'àudio.

Una vegada l'arxiu d'àudio i la transcripció fonètica han estat emmagatzemats, es realitza la parametrització de forma automàtica. Els paràmetres calculats són visualitzats en la interfície gràfica i poden ser modificats manualment per l'usuari si ho considera necessari[2]. Els paràmetres formen una instància incompleta (perquè no té associada una etiqueta afectiva) que és avaluada pel model de reconeixement creat en anteriors execucions, assignant-li una etiqueta afectiva[3]. La secció C.1.3 descriu els detalls d'aquest procés.

L'etiqueta afectiva assignada a la instància en el pas de classificació anterior és mostrada a l'usuari, el qual pot determinar llavors si és correcta o no. Si ho ha estat, aquesta etiqueta s'uneix a la instància anterior creant una instància completa i aquesta instància s'afegeix al conjunt de dades format per les instàncies anteriors. A continuació es realitza un entrenament del model amb el conjunt de dades actualitzat. Si no ho ha estat, l'usuari té l'oportunitat d'indicar l'etiqueta correcta i es procedeix igual que en el cas anterior. En cas que l'usuari estigui avaluant el model de reconeixement en mode de prova només es visualitzarà l'etiqueta afectiva, però no es procedirà, en cap cas, a actualitzar el conjunt de dades ni a entrenar de nou el model.

[1] En la versió actual de l'aplicació aquest procés és manual. La inclusió d'un mòdul de reconeixement de veu automatitzaria aquest pas.

[2] Això és així atès que aquesta és una aplicació per a l'estudi del reconeixement afectiu. Aquesta opció no tindria lloc en una aplicació per explotar aquest reconeixement.

[3] En cas de tractar-se de la primera execució l'etiqueta assignada és, per defecte, la corresponent a l'estat afectiu alegre

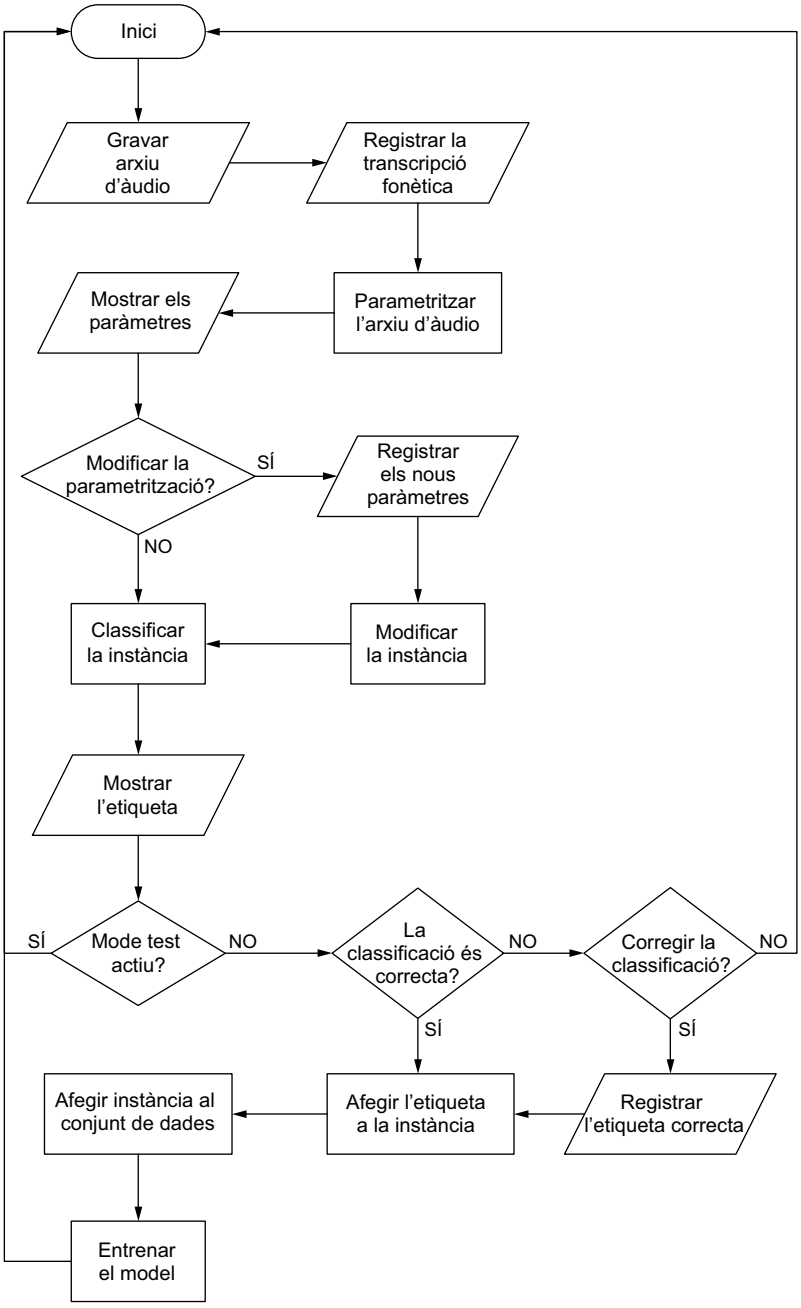


Figura C.1: Diagrama de flux de l'aplicació d'anàlisi afectiva automàtica basada en aprenentatge interactiu incremental.

C.1.3 Implementació del mòdul de reconeixement afectiu

La versió actual d'aquesta aplicació implementa un mòdul de reconeixement afectiu basat únicament en l'anàlisi del senyal de veu. Donat el seu caràcter preliminar de desenvolupament, la parametrització és senzilla i consta d'un total de 17 paràmetres ba-

sats en una segmentació en fonemes. Aquests paràmetres es refereixen a l'energia, F0 i durada dels fonemes i dels silencis. Per a cadascun d'ells es calculen 4 funcionals: valor mitjà, desviació estàndard, valor màxim i valor mínim, i per al cas dels silencis s'afegeix també el nombre de trames de silenci detectades.

El ventall d'emocions pel qual es crea el conjunt de dades per defecte en el moment de crear el model abasta 4 emocions bàsiques: alegria, tristesa, enuig i por. No obstant això la definició del nombre i quines haurien de ser aquestes emocions podria ser modificable en versions posteriors.

L'algorisme de classificació per defecte és l'arbre de decisió J4.8 si bé la versió actual del programa ja ofereix la possibilitat de poder escollir un altre esquema d'aprenentatge en el moment de crear el model de reconeixement.

C.1.4 Implementació de l'aplicació

L'aplicació d'anàlisi afectiva automàtica basada en aprenentatge interactiu incremental és una aplicació independent implementada en Java, emprant la biblioteca gràfica *Swing* per implementar la interfície gràfica d'usuari. La figura C.2 mostra el disseny d'aquesta interfície gràfica en la qual es poden distingir els diferents mòduls que la formen:

- Enregistrament d'àudio. Gestiona la funcionalitat d'enregistrament dels arxius de veu i la introducció de la seva transcripció.
- Paràmetres. Visualitza els paràmetres calculats i permet la seva edició.
- Classificació. Visualitza l'etiqueta afectiva assignada a l'arxiu d'àudio i permet indicar la validesa de la mateixa.
- Correcció de la classificació. Permet corregir l'etiqueta afectiva assignada.
- Informació. Mostra informació rellevant del model en ús.

Aquesta aplicació fa ús d'un altre programari per implementar algunes funcionalitats específiques. Quant al processament de l'arxiu d'àudio per a l'extracció de paràmetres basada en fonemes s'empra l'aplicació ITPcommmand. ITPcommand és una aplicació basada en el programari ITP (Alías i Iriondo, 2002) desenvolupada específicament per a aquest cas i permet accedir a diferents funcionalitats del programari ITP a través de crides independents mitjançant la línia de comandes o des d'una altra aplicació. Aquesta aplicació genera una sèrie d'arxius que han de ser processats posteriorment per extreure els paràmetres requerits. Donada la necessitat de la transcripció fonètica del text corresponent a l'arxiu d'àudio, s'empra l'eina RST[4] la qual permet la seva extracció de forma

[4] *Research Speech Toolkit* (Gonzalvo, 2010).

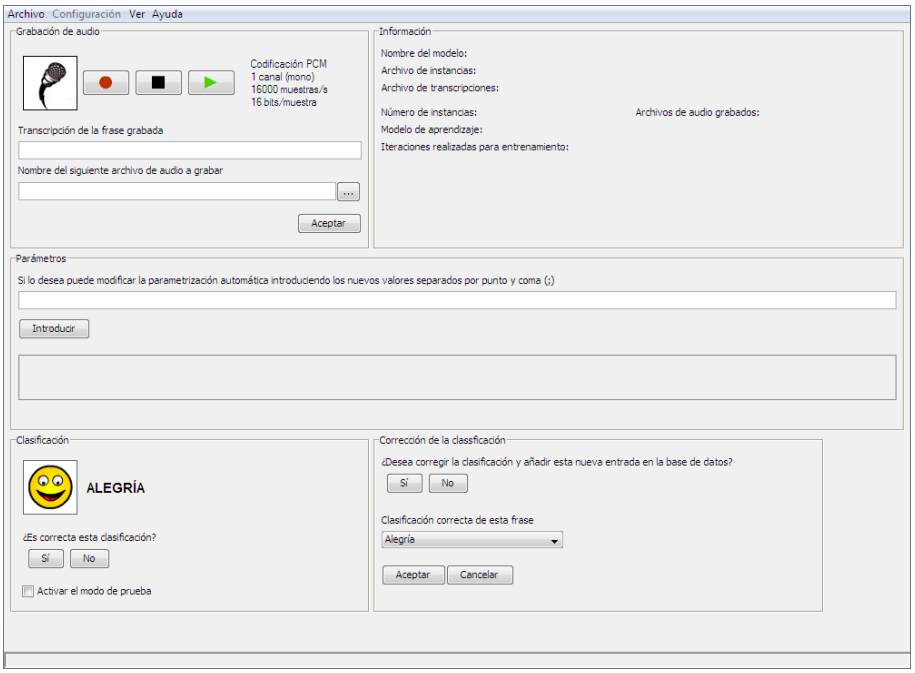


Figura C.2: Interfície gràfica de l'aplicació d'anàlisi afectiva automàtica basada en aprenentatge interactiu incremental.

automàtica a partir de l'arxiu de text. ITPcommand realitza la segmentació fonètica a través de la transcripció i l'arxiu d'àudio així com també extreu els paràmetres necessaris.

L'anàlisi dels arxius de dades generades per l'aplicació ITPcommand i l'extracció dels paràmetres necessaris es realitza mitjançant el programa MATLAB, de Mathworks, que ha d'estar instal·lat en l'ordinador de l'usuari. Tot aquest procés és completament transparent a l'usuari.

Quant als algorismes de classificació, aquesta aplicació incorpora el codi de la llibreria WEKA. WEKA és de fàcil distribució[5] i flexible en la incorporació dels classificadors en el codi font de l'aplicació principal, ja que també està implementada en Java. A més, treballar amb aquesta llibreria afavoreix l'ús d'un conjunt de dades en un format compatible amb ella el que facilita la posterior anàlisi de dit conjunt de dades directament en l'aplicació WEKA de forma externa a l'aplicació principal.

Cal destacar que els mòduls són independents i es comuniquen entre si a través d'arxius d'àudio, les instàncies i el conjunt de dades. Per aquest motiu cada mòdul pot ser fàcilment reemplaçat per un altre que realitzi la mateixa tasca però seguint un altre algorisme. Així doncs, podria modificar-se la parametrització (tant a nivell de quins paràmetres es calculen com a nivell del programari que s'empra per a la mateixa) simplement canviant la part de codi que afecta a aquest mòdul sense variar les altres, la qual cosa permet que la plataforma sigui fàcilment actualitzable.

[5]WEKA es distribueix sota Llicència Pública General de GNU.

L'aplicació es distribueix a través d'un arxiu d'instal·lació que copia en l'ordinador de l'usuari tot el codi executable i les llibreries necessàries. En aquesta versió de desenvolupament es fa necessari, tal com s'ha comentat, que l'usuari disposi del programa MATLAB instal·lat en l'equip. L'instal·lador detecta automàticament la seva configuració i copia els arxius necessaris. En versions posteriors d'explotació el càlcul dels paràmetres haurà de ser realitzat mitjançant una llibreria pròpia del programa, sense necessitat de recórrer a instal·lacions de programari propietari independents.

C.2 Mòdul de reconeixement afectiu audiovisual

Un dels objectius del projecte INREDIS era la creació d'un mòdul de reconeixement afectiu que pogués ser utilitzat per una aplicació externa. Amb aquesta finalitat es crea el mòdul exposat en aquesta secció. Aquest mòdul es basa en les modalitats de veu i expressivitat facial, tal com es detalla a continuació.

C.2.1 Descripció

El Mòdul de Reconeixement Afectiu Audiovisual (MRAA) és una llibreria Java que incorpora les funcionalitats necessàries per processar un arxiu amb contingut audiovisual i assignar-li una etiqueta afectiva sobre la base d'un model de reconeixement prèviament entrenat. Aquest mòdul permet crear el model en cas de no existir, així com entrenar-ho de forma incremental a través de l'anàlisi d'arxius audiovisuals successius. És a dir, el sistema posseeix la capacitat d'adaptar els models emocionals als usuaris mitjançant un sistema d'aprenentatge semiautomàtic que requereix de la supervisió del propi usuari. La llibreria pot funcionar com a aplicació independent o bé utilitzar-se dins del codi font d'una altra aplicació.

El MRAA es desenvolupa originalment com a part de la tasca 4.5 (Tecnologia multimodal afectiva) del paquet de treball PT4 (Tecnologies d'interacció persona-màquina) del projecte INREDIS. El MRAA forma part del prototip denominat 4.1+4.3+4.5 i té com a principal funcionalitat la detecció automàtica de l'estat afectiu de l'usuari a partir de les modalitats d'expressivitat facial i de la veu del mateix. La detecció automàtica de l'estat afectiu permet que la interfície de sortida del prototip[6] pugui adaptar la seva expressivitat o emoció aconseguint un gran nivell d'empatia entre la persona i l'ordinador i reduir, d'aquesta manera, el distanciament o la barrera que suposa per a persones no expertes en l'ús d'aquestes tecnologies interactuar amb aquest tipus de sistemes. Posteriorment, el MRAA és ampliat i adaptat per a la seva inclusió com a part del projecte evMIC[7], que té com objectiu principal la creació d'una plataforma interoperable centrada en l'usuari

[6] La interfície de sortida compta amb un motor de diàleg per generar respostes en llenguatge natural així com un avatar virtual.

[7] Projecte TSI-020301-2009-25 cofinançat pel *Ministerio de Industria, Turismo y Comercio de España*.

capaç de permetre la creació d'entorns virtuals d'aprenentatge.

La concepció del MRAA és similar a la de l'aplicació descrita en la secció C.1, ja que es planteja la creació d'un model de reconeixement capaç d'adaptar-se a l'expressivitat afectiva pròpia de l'usuari. Cal puntualitzar que aquest mòdul es dissenya inicialment per a un cas d'ús específic definit en el projecte INREDIS com a "Usuari perdut". Aquest cas d'ús planteja un actor que s'ha perdut a la ciutat i que precisa d'un servei d'assistència perquè li proporcioni ajuda. Aquest cas d'ús planteja només tres estats afectius: positiu (emoció d'alegria o acceptació del missatge), negatiu (emoció d'enuig o rebuig del missatge) i neutre (absència d'emoció), si bé el mòdul podria ser adaptat per admetre noves etiquetes. Mitjançant aquesta codificació, l'usuari del MRAA pot crear un model de reconeixement i entrenar-ho a partir d'interaccions gravades en arxius d'àudio i vídeo. El sistema pot ser entrenat amb tants senyals com es desitgi, diversificant els exemples subministrats sent recomanable aplicar un nombre d'entrenaments similar per a cadascuna de les tres categories afectives anteriors. No obstant això, la programació original del sistema es flexibilitza posteriorment per permetre la seva adaptació al projecte evMIC que presenta unes condicions menys restrictives.

El reconeixement afectiu es realitza a partir del processament dels senyals d'àudio i de vídeo[8], i el mòdul proporciona etiquetes afectives per a ambdues modalitats. D'altra banda, el mòdul proporciona una sortida única com a resultat de la fusió d'ambdues modalitats. Aquesta etiqueta afectiva addicional proporciona robustesa davant de possibles problemes en les modalitats anteriors. Així, davant de l'absència del canal de vídeo o d'àudio, la sortida multimodal ignora la modalitat que ha fallat. Davant d'un empat entre la decisió presa per ambdues modalitats, el sistema prioritza la decisió del canal de vídeo perquè es considera més robusta[9]. Si es produeix un error que afecta a ambdues modalitats simultàniament o si la seqüència no pot ser processada correctament (a causa de problemes amb el propi arxiu, per exemple) el sistema genera una etiqueta adequada informant d'aquest error, sense que això ocasioni una fallada general de l'aplicació.

C.2.2 Característiques generals

El sistema assigna una etiqueta afectiva a cada seqüència audiovisual per a cada modalitat i una addicional corresponent a la fusió d'ambdues modalitats. El sistema pot generar una etiqueta emocional fins i tot en l'absència d'alguna de les modalitats d'entrada. No obstant això, si la fallada afecta a ambdues modalitats simultàniament o si la seqüència no pot ser processada correctament (a causa de problemes amb el propi arxiu, per exemple) el sistema generarà una etiqueta adequada informant d'aquest error, sense que això ocasioni una fallada general de l'aplicació.

[8] El processament de vídeo consisteix, realment, en l'anàlisi de tres fotogrames de la seqüència de vídeo completa. L'etiqueta afectiva assignada a aquesta modalitat consisteix en una combinació de les decisions preses pel MRAA per a cadascun d'aquests fotogrames.

[9] Es considera més robusta perquè que la seva etiqueta procedeix de tres decisions independents.

Es poden produir errors en el processament de la seqüència de vídeo en el moment de realitzar la detecció de la cara i el seu seguiment. Aquests errors poden ser freqüents en funció de la qualitat de l'enregistrament. Això desembocarà en errors en l'assignació de l'etiqueta afectiva però no ocasionarà fallades generals de l'aplicació. Per evitar l'aparició d'aquests errors de processament del vídeo serà necessari que la càmera enfoqui la cara de l'usuari obtenint una imatge frontal (sense rotacions), ben il·luminada i sense ombres ni focus de llum massa vius, i l'usuari haurà de romandre ben enfocat i sense realitzar moviments bruscs, ni rotacions del cap. També és important que la resolució de la cara sigui prou elevada com per poder extreure correctament els atributs facials que més denoten l'estat afectiu de l'usuari[10]. Quant a l'àudio, el micròfon ha d'estar ben calibrat de manera que la veu tingui bona presència sense arribar a provocar saturació o retallada del senyal. El sistema no és robust davant de sorolls o interferències que es puguin produir durant la interacció.

Inicialment el MRAA proporcionarà respostes poc fiables i el nivell de fiabilitat aconseguit dependrà de la similitud entre la cara i veu de l'usuari i la base de dades de cares i veus prèvia disponible. Això succeirà mentre no s'hagi aconseguit un nivell d'entrenament personalitzat suficientment satisfactori. El nivell de fiabilitat en la detecció de l'emoció de l'usuari millorarà a mesura que el sistema s'entreni amb més casos (exemples de locució emocional) de l'usuari, dotant al sistema d'un nivell de personalització més gran. Aquest entrenament es realitzarà amb la participació activa de l'usuari.

El sistema compta amb la capacitat d'actualitzar els models de reconeixement durant la seva explotació la qual cosa li confereix l'habilitat d'aprendre o reforçar el reconeixement dels estats afectius. El sistema pot garantir la capacitat d'aprenentatge per adequar el reconeixement a un únic usuari, si bé pot ser entrenat i explotat per diversos usuaris amb la finalitat d'intentar convergir cap a un model de reconeixement multiusuari.

El reconeixement dut a terme pel sistema se suposa independent de l'idioma emprat pels locutors, si bé es poden produir diferències en l'expressió de l'estat afectiu entre locutors de contextos culturals diferents.

El sistema proporciona una resposta quan s'ha completat l'extracció de paràmetres audiovisuals de la seqüència i aquests han estat classificats per l'algorisme d'aprenentatge. En funció de la longitud de la seqüència aquest procés pot arribar a ser costós computacionalment (segons el processador de què es disposi i de la resolució de les imatges i del grau d'optimització del codi que es pugui aconseguir) pel que pot existir un retard considerable que dificulti la seva aplicació en un entorn a temps real. No obstant això, aquests temps poden veure's reduïts si es limita la durada dels enregistraments a locucions curtes. Aquest aspecte pot condicionar que el reconeixement emocional sigui usat més com una eina d'anàlisi de les interaccions en comptes d'usar l'estat afectiu reconegut com a part influent de la resposta immediata en un sistema d'interacció en temps real.

[10] Es requereix que el rostre ocupi més del 60% de la imatge en una imatge amb resolució de 640×480 píxels.

C.2.3 Funcionament del MRAA

En aquesta secció es detalla el funcionament general del MRAA i es mostren exemples del seu funcionament com a aplicació independent i com a llibreria que pot ser cridada des del codi font d'una altra aplicació. Els detalls de la implementació dels mòduls que formen el MRAA poden consultar-se en la secció C.2.4.

C.2.3.1 Funcionament general del MRAA

La figura C.3 mostra el diagrama de blocs del MRAA. Principalment, el mòdul té dos components de parametrització dels senyals d'entrada (un per la de vídeo i un altre per la d'àudio), tres components de classificació (un per a la modalitat de vídeo, un altre per a la modalitat d'àudio, i un altre per a la classificació audiovisual que té en compte ambdues modalitats), i finalment un mòdul per a l'entrenament dels classificadors. Els mòduls de parametrització generen a la seva sortida un vector de paràmetres rellevants des d'un punt de vista de la seva funció per a la classificació a partir dels senyals d'entrada corresponents. Els mòduls de classificació reben aquest vector com a entrada per realitzar la funció pròpia del reconeixement de l'estat afectiu de l'usuari.

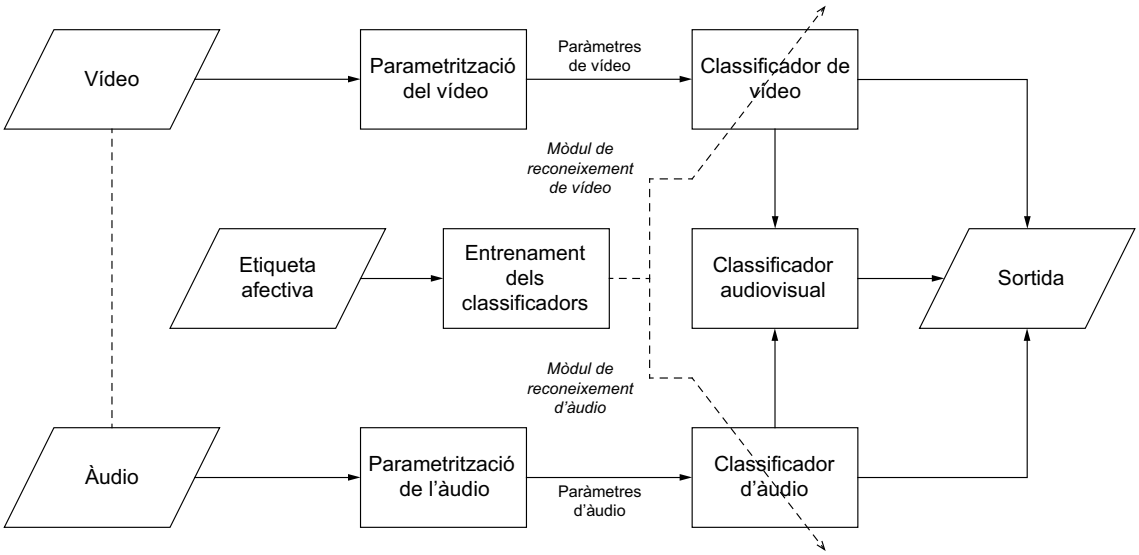


Figura C.3: Diagrama de blocs del MRAA.

El MRAA té dos modes de funcionament:

- Mode d'entrenament d'un model de reconeixement.
- Mode d'avaluació de la interacció d'un usuari.

En el mode d'entrenament, el MRAA considera tres entrades: els senyals de ví-

deo i d'àudio i l'etiqueta afectiva associada. Una vegada s'han parametritzat els senyals de vídeo i d'àudio el component d'entrenament dels classificadors modifica els models de reconeixement mitjançant un cicle d'entrenament complet[11]. Perquè els classificadors funcionin correctament és necessari realitzar entrenaments amb diferents exemples suficientment significatius per a cada emoció a reconèixer. Es requereix la intervenció de, com a mínim, un usuari que supervisi l'etiqueta afectiva a la qual s'associa cada interacció usada per a l'entrenament.

En el mode d'avaluació no es precisa l'etiqueta afectiva ja que aquesta serà assignada pel MRAA al final de la classificació. Una vegada s'ha parametritzat el vídeo i l'àudio, els classificadors respectius realitzen la tasca de classificació usant els models de reconeixement disponibles, donant a la seva sortida una etiqueta identificativa de l'emoció reconeguda. Les etiquetes assignades per aquests dos classificadors es prenen com a entrada del classificador audiovisual. Aquesta tercera etiqueta generada per aquest classificador també segueix el mateix conveni que les altres dues (emocions neutra, negativa i positiva), però permet obtenir decisions més estables enfortint la decisió final gràcies a la combinació de les modalitats anteriors.

El MRAA, tal com s'ha comentat, pot funcionar com a aplicació independent o com una llibreria importable al codi font d'una aplicació Java. A continuació es detallen exemples de tots dos funcionaments.

C.2.3.2 Execució del MRAA com a aplicació independent

Per crear i entrenar un model de reconeixement n'hi ha prou amb crear, inicialment, una base de dades d'àudio i vídeo buides a partir de les bases de dades de mostra que es distribueixen amb l'aplicació. Partint d'una interacció d'un usuari amb el sistema de reconeixement, recollida en un arxiu d'àudio que contingui l'enregistrament de la seva veu i un altre de vídeo amb la seva expressió facial, ha d'executar-se la següent instrucció[12]:

```
java -jar mraa.jar
<dir-configuracion>/mraa-config.properties
<dir-audio>/<archivo-audio.wav>
<dir-video>/<archivo-video.wmv>
<dir-modelos>/<modeloAudio.mod>
<dir-modelos>/<modeloVideo.mod>
<dir-baseDatos>/<datasetAudio.arff>
<dir-baseDatos>/<datasetVideo.arff>
train <etiquetaAfectiva>
```

[11] És a dir, els paràmetres d'àudio i de vídeo s'uneixen a l'etiqueta proporcionada per crear una instància que s'afegeix al conjunt de dades. Aquest nou conjunt de dades serveix per entrenar de nou el classificador.

[12] Aquesta és una línia única però s'han afegit salts de línia per millorar la seva visualització impresa. Cada salt de línia equival a un espai en blanc.

On:

- <dir-configuracion> és el directori en el qual està l'arxiu de configuració mraa-config.properties.
- <dir-audio> és el directori en el qual està l'arxiu d'àudio a analitzar (archivo-audio.wav).
- <dir-video> és el directori en el qual està la seqüència de vídeo a analitzar (archivo-video.wmv).
- <dir-models> és el directori en el qual es crearan els models de reconeixement d'àudio i de vídeo (modeloAudio.mod i modeloVideo.mod, respectivament).
- <dir-baseDatos> és el directori en el qual estan les bases de dades d'àudio i de vídeo que s'han creat anteriorment (datasetAudio.arff i datasetVideo.arff, respectivament).
- La clàusula "train" indica que s'està realitzant un entrenament d'aquests models amb aquesta interacció.
- <etiquetaAfectiva> indica l'emoció associada a aquesta interacció. Les opcions vàlides són "neg" per als estats negatius, "pos" per als positius i "neu" per als neutres.

La propera vegada que es realitzi una crida com l'anterior, usant els mateixos arxius de model i de base de dades, aquests arxius seran actualitzats.

Una vegada que ja es disposa d'un model creat i entrenat pot avaluar-se la interacció d'un usuari la qual, com abans, ha d'estar emmagatzemada en un arxiu d'àudio i un altre de vídeo. La instrucció per a l'avaluació és[13]:

```
java -jar mraa.jar <dir-configuracion>/mraa-config.properties
<dir-audio>/<archivo-audio.wav> <dir-video>/<archivo-video.wmv>
<dir-modelos>/<modeloAudio.mod> <dir-modelos>/<modeloVideo.mod>
<dir-baseDatos>/<datasetAudio.arff> <dir-baseDatos>/<datasetVideo.arff>
test
```

En aquest cas la clàusula "test" indica que s'està avaluant aquesta interacció. L'aplicació retornarà la informació relativa al resultat de l'avaluació realitzada pel mòdul acústic, el mòdul visual i el mode audiovisual conjunt. Aquesta informació seguirà el format descrit pel següent exemple:

[13] Aquesta és una línia única però s'han afegit salts de línia per millorar la seva visualització impresa. Cada salt de línia equival a un espai en blanc.

```
Detalle de resultados
=====================
 Audio 1: neg
 Video 1: neg
 Video 2: pos
 Video 3: neg
Resultado final del reconocimiento afectivo
=========================================
 <EmocionAudio value="neg">,<EmocionVideo value="neg">,
 <EmocionAudiovisual value="neg">
```

C.2.3.3 Execució de les funcions del MRAA des del codi d'una aplicació Java

El MRAA pot ser emprat en el codi font d'una aplicació Java, incloent la llibreria *mraa.jar* en el projecte de l'aplicació. No obstant això, el MRAA encara es troba en una versió inicial pel que les funcionalitats que ofereix com a llibreria són molt limitades així com la documentació proporcionada. Bàsicament compta amb un constructor que permet determinar la manera de funcionament del mòdul de reconeixement i una funció destinada a l'anàlisi d'una interacció a partir d'un arxiu d'àudio i de vídeo. Aquesta funció és la següent:

```
public static java.lang.String etiqueta = mraa.AnalisiAfectiu.getEmotion(
    String sPathProp,
    String sPathAudio,
    String sPathVideo,
    String sPathModelAudio,
    String sPathModelVideo,
    String sPathDataAudio,
    String sPathDataVideo)
  throws java.lang.Exception
```

Aquesta funció retorna una cadena de text amb el següent format:

```
<EmocionAudio value="xxx">,
<EmocionVideo value="yyy">,
<EmocionAudiovisual="zzz">
```

On *xxx*, *yyy* i *zzz* són les etiquetes emocionals detectades dins de l'àmbit {*pos, neu, neg, dkn*}, corresponents a positiu, neutre, negatiu i desconegut, respectivament. L'etiqueta *dkn* es reserva per al cas en què fallés el procés de reconeixement. Aquesta fallada pot ser a causa d'un problema en el procés de parametrització dels senyals així com per l'absència d'ambdues modalitats de forma simultània.

C.2.4 Implementació del mòdul de reconeixement afectiu

En aquesta secció es realitza una breu descripció del disseny dels components usats per el MRAA i que han estat introduïts en l'apartat anterior.

C.2.4.1 Component de parametrització de l'àudio

Aquest component genera un vector de 46 paràmetres associats al senyal d'àudio d'entrada proporcionat al mòdul com un arxiu en format WAV. Els paràmetres considerats són característiques mesurades sobre la corba de F0, sobre la durada dels fonemes (diferenciant entre sons sords i sons sonors) i també sobre l'energia del senyal de veu. D'aquestes característiques es calculen 11 estadístics (valor mitjà, màxim, mínim, desviació estàndard, rang, quartils, rang interquartílic, asimetria i curtosis). Així mateix s'inclou la mesura dels amples de banda mesurats sobre les corbes de variació de F0 i de l'energia del senyal de veu.

Per obtenir les corbes de variació de la F0, de l'energia i les durades dels sons sonors i sords, s'aplica un modelatge harmònic estocàstic del senyal de veu *Harmonic plus Noise Model* (HNM). A partir dels paràmetres d'aquest model es poden realitzar estimacions de les corbes i les durades anteriors.

C.2.4.2 Component de parametrització del vídeo

Aquest component genera un vector de paràmetres associats a tres fotogrames del vídeo d'entrada. Aquests tres fotogrames són seleccionats a partir de la inicialització temporal sobre la meitat de tres zones equidistants: la zona inicial, la zona mitjana i la zona final del vídeo[14]. A partir d'aquests punts inicials d'anàlisi, es busquen els fotogrames més propers que permeten detectar amb èxit la zona d'interès, a través d'un procés de prova i error. La zona d'interès està formada per la regió que inclou els ulls i les celles de la persona, ja que és a partir de la qual que es pot mesurar informació correlada amb l'emoció o afectivitat de la persona.

En primer lloc, aquest component calcula la luminància de cada fotograma analitzat i, posteriorment, realitza un seguiment de regions basat en aparença que consisteix en la localització i seguiment dels ulls i del nas de la persona. Una vegada que s'ha localitzat la zona de la cara de l'usuari, s'aplica un filtre logístic per compensar possibles problemes d'il·luminació en la imatge. Seguidament s'aplica una màscara que aïlla la zona dels ulls. Després de realitzar aquest procés de detecció de la zona ocular s'obté una imatge de 27×48 píxels, la qual s'obté mitjançant un procés de remostreig a partir de la zona de la imatge original marcada.

[14] Entenent aquestes tres zones com una divisió temporal.

La parametrització de cadascun dels tres fotogrames considerats en la parametrització del vídeo es basa en l'aplicació d'un banc de filtres Gabor amb diferents angles i freqüències espacials[15]. A les corresponents sortides complexes de cada filtre se'ls aplica l'operador mòdul. Així doncs, el vector de paràmetres per a cadascun dels tres fotogrames analitzats està format per 40 subvectors de 1.296 valors[16] cadascun.

C.2.4.3 Components d'entrenament i classificació de les modalitats d'àudio i vídeo

Els mòduls de classificació d'àudio i de vídeo són classificadors de tipus SVM. No obstant això, les característiques de la programació del MRAA permeten que aquests esquemes puguin ser fàcilment modificats per uns altres.

C.2.4.4 Component de classificació audiovisual

Aquest component segueix una regla heurística basada en vot majoritari per combinar els resultats dels mòduls independents de reconeixement basats en àudio i vídeo. La modalitat d'àudio compta amb un vot. La modalitat de vídeo compta amb tres, atès que es processen tres fotogrames per decidir l'emoció final. La combinació de les quatre decisions resulta en la decisió final del mòdul conjunt.

C.2.5 Implementació del MRAA

El MRAA ha estat desenvolupat en llenguatge de programació Java, la qual cosa permet facilitar la portabilitat a múltiples plataformes. La seva implementació no considera una interfície gràfica d'usuari, a diferència de l'aplicació d'anàlisi afectiva automàtica basada en aprenentatge interactiu incremental desenvolupada anteriorment, ja que aquest mòdul es concep originalment per formar part d'un altre sistema, si bé pot ser executada com a aplicació final.

Com l'aplicació anterior, el MRAA empra un altre programari per implementar algunes funcions. Així, incorpora la llibreria gratuïta de codi obert Xuggler[17] per realitzar la descodificació dels arxius de vídeo. Així mateix, incorpora el compilador en temps d'execució MATLAB[18] per realitzar tasques de parametrització i processament. La incorporació d'aquest programari evita que l'usuari final hagi de tenir instal·lat el programa MATLAB. Finalment, inclou la implementació dels algorismes de classificació procedents de la seva implementació en la llibreria WEKA.

[15] Concretament, es tracta d'una anàlisi multiresolutiva amb 40 kernels diferents.
[16] Corresponents a la quantitat de píxels de cada imatge (27×48).
[17] http://www.xuggle.com/
[18] MCR, de l'anglès *MATLAB Compiler Runtime*, implementat per Mathworks.

Universitat Ramon Llull

Aquesta Tesi Doctoral ha estat defensada el dia ____ d _____ de 20

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
Santiago Planet García

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