UNIVERSITAT POLITÈCNICA
DE CATALUNYA

Ph.D. Dissertation

Grapheme-to-Phoneme Conversion in the Era of Globalization

Tatyana V. Polyàkova

Thesis advisor: *Antonio Bonafonte*

TALP Research Center, Speech Processing Group
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya

Barcelona, November 2014

Language is the most massive and inclusive art we know, a mountainous and anonymous work of unconscious generation.

Edward Sapir

Abstract

Only ten years ago, applications of Text-to-speech (TTS) systems were much more limited even though such a recent past seems so far away due to the massive invasion of our lives by smart technologies. Service automatization process has also reached a new level. So what defines a good TTS system nowadays? The market requires it to be highly adaptive to any kind of framework. High level of reliability is also a must since a simple mistake by a TTS can lead to unpleasant if not serious consequences in our daily lives. Our agenda becomes more and more demanding and we need to cope with more information in less time. We delegate our everyday tasks to our newest devices that help us read while we are doing something else, choose products, find a place, etc. Besides we travel more and more everyday. We learn to speak new languages, we mix them, we become globalized. A TTS system that is not able to cope with multilingual entries will not be able to sustain competition. TTS systems have to be multilingual. Phonetic transcription is the first TTS module which makes its precision fundamental.

This thesis focuses on improving adaptability, reliability and multilingual support in the phonetic module of our TTS system. Phonetic transcription module in TTS switched from being rule- or dictionary-based based to being automatic, data-driven. The language is constantly evolving just like all living organisms. That is why adaptability is one of major concerns in phonetic transcription. Therefore, a well-functioning data-driven method is needed to drive the pronunciation of out-of-vocabulary words. This thesis compares different data-driven Grapheme-to-phoneme (G2P) methods using the same training and test data and proposes improvements. Several data-driven classifiers such as Decision Trees (DT), Finite State Transducers (FST), Hidden Markov Models (HMM) are applied to the G2P task and their performance is compared. The obtained results are further improved by means of application of the Transformation-based error-driven learning (TBL) algorithm, which allows to capture and correct error patterns. Significant improvements are obtained especially for classifiers with higher error rates. Best results are obtained by the best classifier FST enhanced by TBL, the word accuracy improves by 2-4 percentage points depending on the lexicon. Other classifiers enhanced by TBL show improvements between 8 and 83 percentage points in word accuracy. The improvements that are obtained by

application of the TBL to the output of the simplest classifier based only letter-phoneme correspondences in the training corpora Most-likely phoneme assignation (ML) is huge (77-83 percentage points depending on the lexicon), which proves the effectiveness of the TBL by itself. The success of the TBL algorithm proves the effectiveness of learning from errors which is quite similar to human language learning. Another technique that humans use on a regular basis when learning languages is pronunciation by analogy. This is even more true for languages with deep orthography, where the correspondence between the written and spoken forms is quite ambiguous. To further improve adaptability of our phonetic module pronunciation by analogy algorithm was developed. This algorithm finds letter arcs with the same pronunciation and calculates their frequency. The pronunciation of new words is built from largest arcs that constitute the shortest path through the graph of all available pronunciations for this particular word. Based on such parameters as arc frequency, position in the word, etc., the resulting shortest pronunciations (if several) are given a rank and the scoring strategies choose the best one. New scoring strategies are proposed and promising results are obtained. One of the newly proposed strategies clearly outperforms the others. Proposed strategies also appear in top strategy combinations. The best results for Pronunciation by Analogy (PbA) are between 63 and 88% words correct depending on the lexicon. The G2P results are given for English and several other European languages. The case of connected speech is also considered. Pronunciation adaptation for connected speech is carried out using weak forms. The overall results show that adaptability of phonetic module has been improved . Next, steps are taking towards increasing reliability of the output of the phonetic module. Although, as the adaptability experiments show the G2P results are quite good, even a few mistakes can seriously hamper the intelligibility of certain words and , therefore, the overall speech quality. It is proposed to achieve a higher level of reliability through dictionary fusion. The ways the pronunciations are represented in different lexical depend on many factors such as: expert's opinion, local accent specifications, phonetic alphabet chosen, assimilation level (for proper names), etc. There are often discrepancies between pronunciations of the same word found in different lexica. Usually these discrepancies, although sometimes significant, do not seriously hamper the overall pronunciation of the word since all lexicon pronunciations have been previously validated by an expert linguist. These discrepancies can be present in vowel or diphthong transcription. Substitution of a vowel by a similar one does not affect the intelligibility and therefore speech quality. The fusion system is a Phoneme-to-phoneme (P2P) system that transforms pronunciations from the source lexicon into pronunciations from the target lexicon (the system is trained to learn these transformations). To train the classifier, common entries from both source and target lexica are selected. The experiments are carried out both for common words and proper names. Improvements ranging from 13 to 50 percentage points are obtained for different lexicon pairs. The results are obtained with Decision Trees (DT) and FST show important compatibility improvements. These results

show that the overall speech quality can be significantly improved given the already low G2P error rates and extensive system dictionary coverage.

An adaptable and reliable TTS system needs to be ready to face the challenge of multilingualism, the phenomenon becoming a usual part of our daily lives. This thesis considers mixed-language contexts where language can change unexpectedly. Such contexts are widely present in social media, forums, etc. A multilingual G2P scheme including the *nativization* proposal is presented. The first component of a multilingual TTS is the language identification module. N-gram based language identifier is developed and good results are obtained. Mixed-language contexts are to be treated with special delicacy. Usually each utterance or paragraph have a main language and the foreign words present in it are considered to be inclusions. How should mixed-language utterances be pronounced? Two scenarios can be considered here: 1) to apply different G2P classifiers depending on the language of foreign inclusion leading to harsh phonetic changes (such changes would be very unnatural to languages like Spanish); 2) to apply the G2P converter for the main language of the utterance assuming that this pronunciation would be more acceptable than the foreign-sounding one. What if it is neither? For countries like Spain, where population's proficiency in foreign languages is rather limited we propose to *nativize* pronunciation of foreign words in Spanish utterances. Which criteria should be used given the significant differences in the phoneme inventory? Our goal is to obtain pronunciations that are not totally unfamiliar either to a native or proficient speakers of the language to be adapted, or to speakers of this language with average to low proficiency. Nativization is carried out for both English and Catalan inclusions in Spanish utterances. When there are significant differences in phoneme inventories between the languages, nativization presents additional challenges. In order to quickly validate the idea of nativization, tables mapping foreign phonemes to the nativized ones are created and a perceptual evaluation is held. Nativization using mapping tables shows a much higher level of acceptance by the audience than synthesis without any nativization.

In order to further improve nativization results an efficient data-driven method is needed. As a great part of foreign pronunciations are learned by analogy what can be better than using the PbA classifier for this task, especially since it has already shown great performance for the G2P task. Analogy both in orthographic and phonetic domains is believed to help achieve a more successful nativization. All data-driven methods require training corpora, PbA, of course, is not an exception. Since no available nativization corpora suitable for the task was found it was decided to create training and test corpora to train our data-driven classifier. These corpora were created for nativization of English and Catalan inclusions in Spanish utterances. Both training corpora contain 1000 words and are orthographically balanced. Grapheme-to-phoneme Nativization (G2Pnat) and Phoneme-to-phoneme Nativization (P2Pnat) are applied in order to nativize English proper names and

common words and Catalan common words in Spanish utterances. The results obtained show that phonetic analogy gives better performance than analogy in the orthographic domain for both proper names and common nouns. However, the results for English proper names are about 12 percentage points lower than those obtained for English common words. This is due to the fact that proper names pronunciation is influenced by more complex factors, as even for humans it presents important challenges. TBL algorithm is also applied to enhance nativization results for English inclusions. Top scoring P2Pnat results were further improved as well as results obtained by nativization tables. Good results obtained by TBL algorithm applied to the ML prediction proves the effectiveness of learning from errors for this task as well. In the perceptual evaluation carried out for English the listeners were asked to vote best and words out of three available methods (Spanish G2P, NatTAB and P2Pnat).

P2Pnat is voted best in 50% of the cases while Spanish G2P obtains the most negative votes (45% of the cases). These perceptual results and as well as encouraging objective results prove the suitability of nativization for multilingual TTS systems.

Resum

Fa tan sols uns deu anys les aplicacions de sistemes TTS eren molt més limitades, encara que un passat tan recent sembla més llunyà a causa dels canvis produïts en les nostres vides per la invasió massiva de les tecnologies intel·ligents. Els processos d'automatització de serveis també han assolit nous nivells. Què és el que defineix un bon sistema TTS avui dia? El mercat exigeix que aquest sigui molt adaptable a qualsevol tipus d'àmbit. També és imprescindible un alt nivell de fiabilitat ja que un simple error d'un TTS pot causar problemes seriosos en el nostre dia a dia. La nostra agenda és cada vegada més exigent i hem de fer front a més volums d'informació en menys temps. Deleguem les nostres tasques quotidianes als nostres dispositius intel·ligents que ens ajuden a llegir llibres, triar productes, trobar un lloc al mapa, etc. A més viatgem més i més cada dia. Aprenem a parlar noves llengües, les barregem, en un món més i més globalitzat. Un sistema TTS que no és capaç de fer front a les entrades multilingües no serà capaç de sostenir la competència. Els sistemes TTS moderns han de ser multilingües. La transcripció fonètica és el primer mòdul del TTS per la qual cosa el seu correcte funcionament és fonamental. Aquesta tesi se centra en la millora de l'adaptabilitat, fiabilitat i suport multilingüe del mòdul fonètic del nostre sistema TTS. El mòdul de transcripció fonètica del TTS va passar de ser basat en regles o diccionaris a ser automàtic, derivat de dades. La llengua està en constant evolució, igual que tots els organismes vius. És per això que l'adaptabilitat és un dels principals problemes de la transcripció fonètica. Per millorar-la es necessita un mètode basat en dades que funcioni bé per a derivar la pronunciació de paraules no trobades al lèxic del sistema. En aquesta tesi es comparen diferents mètodes G2P impulsats per dades que utilitzen les mateixes dades d'entrenament i test i es proposen millores. S'han aplicat diversos classificadors basats en dades, com ara arbres de decisió, traductors d'estats finits i models de Markov, a la tasca de transcripció fonètica, analitzant i comparant els resultats.

L'algorisme TBL, basat en aprenentatge dels errors proporciona millores adicionals als classificadors esmentats. Aquest mètode permet capturar patrons d'errors i corregir-los. Les millores més significatives s'obtenen per classificadors amb taxes d'errors més gran. Els millors resultats s'obtenen mitjançant l'aplicació del millor classificador FST amb posterior correcció dels errors pel TBL. Els resultats obtingut per altres classificadors i corregits

pel TBL mostren millores entre 2-4 punts percentuals en la taxa d'error de les paraules. La millora que s'obté mitjançant l'aplicació del TBL per als resultats del classificador més simple basat només en correspondències lletra-fonema presents en el corpus d'entrenament, ML, és enorme (77-83 punts percentuals depenent del lèxic), el que demostra l'eficàcia del TBL per si sol. L'èxit de l'algorisme TBL demostra l'eficàcia de l'aprenentatge basat en els errors, que és bastant similar a l'aprenentatge de llengües pels humans.

Una altra tècnica que els éssers humans utilitzen de forma regular en l'aprenentatge d'idiomes és la pronunciació per analogia. Això és encara més cert per a llengües amb ortografia profunda, on la correspondència entre la forma escrita i parlada és bastant ambigua. Per millorar encara més la capacitat d'adaptació del nostre mòdul de pronunciació fonètica, es va desenvolupar un algorisme de pronunciació per analogia. Aquest algorisme troba arcs de lletres als quals correspon la mateixa pronunciació i calcula la seva freqüència. La pronunciació d'una nova paraula es construeix amb els arcs més llargs que constitueixen el camí més curt a través del graf de totes les pronunciacions disponibles per a aquesta paraula. Es basa en paràmetres com ara la freqüència d'arc, posició en la paraula, etc. Les pronunciacions que contenen el menor nombre d'arcs (si hi ha més d'una) es donen un rang i les estratègies de puntuació escullen la millor opció. En aquest treball s'han proposat noves estratègies de puntuació i s'han obtingut resultats prometedors. Una de les noves estratègies propostes clarament supera a les altres. Les noves estratègies propostes també apareixen a la llista de les millors combinacions d'estratègies. Els millors resultats per al PbA són entre 63 i 88 % paraules correctes segons el lèxic. S'han avaluat els G2P no solament per a l'anglès, si no també per altres idiomes europeus. També s'ha considerat el cas de la parla contínua. Per L'anglès, La adaptació de la pronunciació a la parla contínua considera les formes febles. Els resultats generals mostren que la capacitat d'adaptació del mòdul fonètic ha estat millorada.

També s'ha actuat en línies que permeten augmentar la fiabilitat del mòdul fonètic. Tot i que els resultats experimentals per al G2P són bastant bons, encara hi ha errors que poden impedir que la intel·ligibilitat de certes paraules i, per tant, reduir la qualitat de la parla en general. Es proposa aconseguir un major nivell de fiabilitat a través de fusió de diccionaris. Les pronunciació de les paraules presents en els diccionaris depèn de molts factors, per exemple: opinió experta, especificacions de l'accent local, alfabet fonètic triat, nivell d'assimilació (per a noms propis), etc. Sovint hi ha discrepàncies entre la pronunciació de la mateixa paraula en diferents lèxics. En general, aquestes discrepàncies, encara que de vegades significatives, no obstaculitzen greument la pronunciació global de la paraula ja que totes les pronunciacions lèxic han estat prèviament validades per un lingüista expert. Aquestes discrepàncies normalment es troben a la pronunciació de vocals i diftongs. La substitució de vocals per similars no es considera un error greu perquè no afecta la intel·ligibilitat i per tant la qualitat de veu. El sistema de fusió proposat es basa

en el mètode P2P, que transforma les pronunciacions del lèxic d'origen a les pronunciacions del lèxic de destí (el sistema està capacitat per aprendre aquestes transformacions). Per entrenar el classificador, es seleccionen les entrades comunes entre el lèxic font i destí. Els experiments es duen a terme tant per paraules comuns com per a noms propis. Els experiment realitzat s'han basat en les tècniques DT i FST. Els resultats mostren que la qualitat de la parla en general es pot millorar significativament donadas les baixes taxes d'error de G2P i una àmplia cobertura del diccionari del sistema. El sistema TTS final és més adaptable i fiable, més preparat per afrontar el repte del multilingüisme, el fenomen que ja forma part habitual de les nostres vides quotidianes.

Aquesta tesi considera contextos que contenen la barreja de llengües, on la llengua pot canviar de forma inesperada. Aquestes situacions abunden en les xarxes socials, fòrums, etc. Es proposa un esquema de G2P multilingüe incloent la *nativització*. El primer component d'un TTS multilingüe és el mòdul d'identificació d'idioma. S'ha desenvolupat un identificador d'idioma basat en n -gramas (de lletres) obtenint bons resultats. Els contextos amb llengües mixtes han de ser tractats amb especial delicadesa. En general, cada frase o paràgraf tenen una llengua principal i les paraules estrangeres presents s'hi consideren inclusions. A l'hora de decidir com pronunciar frases en diverses llengües es poden considerar dos escenaris: 1) aplicar, per cada llengua el diferents G2P classificadors propis de la llengua (es produiria canvis fonètics bruscs que sonarien molt poc natural); 2) aplicar el classificador G2P per a l'idioma principal de la frase suposant que aquesta pronunciació seria més acceptable que la que conté fonemes estrangers. I si cap de les propostes anteriors es acceptada? Per països com Espanya, on el domini de llengües estrangeres per la població general és bastant limitat, proposem *nativitzar* la pronunciació de paraules estrangeres en frases espanyoles. Quins criteris s'han d'utilitzar tenint en compte les significatives diferències en l'inventari de fonemes? El nostre objectiu és obtenir pronunciacions que no són del tot desconegudes i que siguin acceptades tant per parlants nadius o amb alt domini de l'idioma estranger com per parlants d'aquesta llengua amb nivell mitjà o baix. En aquest treball la nativització es porta a terme per a les inclusions angleses i catalanes en frases en castellà. Quan hi ha diferències significatives en els inventaris de fonemes entre les llengües nativització presenta reptes addicionals. Per tal de validar ràpidament la idea de nativització es van crear taules de mapeig de fonemes estrangers als nativizats, també es va dur a terme una avaluació perceptual. La nativització basada en taules mostra un major nivell d'acceptació per part del públic que la síntesi sense cap nativiztació.

Per tal de millorar encara més els resultats de nativització de forma eficaç es necessita un mètode basat en dades. Com a gran part de pronunciacions estrangeres s'aprenen per analogia, l'aplicació del PbA a aquesta tasca és idoni, sobretot perquè ja ha demostrat excel·lents resultats per a la tasca de transcripció fonètica. Per a això s'explora l'analogia tant en el domini ortogràfic com fonètic. Tots els mètodes basats en dades requereixen

un corpus d'entrenament i PbA, per descomptat, no és una excepció. Ja que cap corpus de nativització adequat per a la tasca estava disponible es va prendre la decisió de crear un corpus d'entrenament i test per entrenar i validar el nostre classificador per inclusions angleses en castellà, i un altre joc per a les catalanes. Tots els dos corpus d'entrenament contenen 1.000 paraules i són ortogràficament equilibrats. S'aplica la nativització per analogia basada en la forma ortogràfica de la paraula G2Pnat i també basada en la forma fonètica acs ppnat per tal d'nativitzar paraules comunes i noms propis en anglès i paraules comunes en català en frases en castellà. Els resultats obtinguts mostren que l'analogia fonètica dóna un millor rendiment que l'analogia en el domini ortogràfic pels noms propis i paraules comunes. No obstant això, els resultats obtinguts per als noms propis anglesos es troben uns 12 punts percentuals per sota dels obtinguts per a les paraules comunes en anglès. Això és degut al fet que la pronunciació noms propis està influenciada per factors més complexos i fins i tot per als éssers humans presenta importants reptes. L'algorisme TBL també s'ha aplicat per millorar els resultats de nativización per inclusions angleses. S'obtenen millores per als resultats obtinguts per P2Pnat, així com per als resultats obtinguts per les taules de nativiztació. Els bons resultats obtinguts per l'algorisme TBL aplicat a la predicció del mètode ML demostra l'eficàcia del mètode d'aprenentatge a partir d'errors, també per a aquesta tasca. A l'avaluació perceptual duta a terme per inclusions angleses en castellà, es va demanar als oients que votessin el millor dels tres mètodes disponibles: G2P (per castellà), NatTAB i P2Pnat. P2Pnat és triat com el millor en el 50 % dels casos mentre que el G2P per a espanyol obté la majoria de vots negatius (45 % dels casos). Aquests resultats perceptuals i els encoratjadors resultats objectius demostren la idoneïtat de nativització per sistemes TTS multilingües.

Resumen

Hace tan sólo unos diez años, las aplicaciones de sistemas TTS estaban mucho más limitadas, aunque un pasado tan reciente parece más lejano debido a los cambios producidos en nuestras vidas por la invasión masiva de las tecnologías inteligentes. Los procesos de automatización de los servicios han alcanzado a nuevos niveles. ¿Qué es lo que define un buen sistema TTS hoy en día? El mercado exige que éste sea muy adaptable a cualquier tipo de ámbito. También es imprescindible un alto nivel de fiabilidad, ya que un simple error de un TTS puede causar problemas serios en nuestro día a día. Nuestra agenda es cada vez más exigente y tenemos que hacer frente a un volumen cada vez mayor de información en menos tiempo. Delegamos nuestras tareas cotidianas a nuestros dispositivos inteligentes que nos ayudan a leer libros, elegir productos, encontrar un lugar en el mapa, etc.

Además, cada día viajamos más, aprendemos a hablar nuevas lenguas, las mezclamos, volviéndonos más y más globalizados. Un sistema TTS que no sea capaz de hacer frente a las entradas multilngües no será capaz de sostener la competencia. Los sistemas TTS modernos tienen que ser multilngües. La transcripción fonética es el primer módulo del TTS por lo cual su correcto funcionamiento es fundamental.

Esta tesis se centra en la mejora de la adaptabilidad, fiabilidad y soporte del módulo fonético de nuestro sistema TTS. El módulo de transcripción fonética del TTS pasó de ser basado en reglas o diccionarios a ser automática, basada en datos. La lengua está en constante evolución al igual que todos los organismos vivos. Es por eso que la adaptabilidad es uno de los principales problemas de la transcripción fonética. Para mejorarla se necesita un método basado en datos que funcione bien para derivar la pronunciación de palabras no encontradas en el léxico del sistema. En esta tesis se comparan diferentes métodos G2P basados en datos, utilizando los mismos datos de entrenamiento y test y se proponen mejoras. Se han estudiado clasificadores basados en datos, tales como árboles de decisión, traductores de estados finitos y modelos de Markov, aplicados a la tarea de transcripción fonética y comparando los resultados.

El algoritmo TBL, basado en aprendizaje de los errores y que permite capturar patrones de errores y corregirlos ha aportado nuevas mejoras, que han sido especialmente

significativas para los clasificadores con tasa de error más alta. Los mejores resultados se obtienen mediante la aplicación del mejor clasificador FST con posterior corrección de los errores por el TBL. Los resultados obtenido por otros clasificadores y corregidos por el TBL muestran mejoras entre 2-4 puntos porcentuales en la tasa de error de las palabras. La mejora que se obtiene mediante la aplicación del TBL para a los resultados del clasificador más simple, basado solamente en correspondencias letra-fonema presentes en el corpus de entrenamiento, ML, es enorme (77-83 puntos porcentuales dependiendo del léxico), lo que demuestra la eficacia del TBL por si solo. El éxito del algoritmo TBL demuestra la eficacia del aprendizaje basado en los errores, que es bastante similar al aprendizaje de lenguas por los humanos.

Otra técnica que los seres humanos utilizan de forma regular en el aprendizaje de idiomas es pronunciación por analogía. Esto es aún más cierto para lenguas con ortografía profunda, donde la correspondencia entre la forma escrita y hablada es bastante ambigua. Para mejorar aún más la capacidad de adaptación de nuestro módulo de pronunciación fonética, se ha estudiado un algoritmo de pronunciación por analogía. Este algoritmo encuentra arcos de letras a los que corresponde la misma pronunciación y calcula su frecuencia. La pronunciación de una nueva palabra se construye con los arcos más largos que constituyen el camino más corto a través del grafo de todas las pronunciaciones disponibles para esta palabra. Se basa en parámetros tales como la frecuencia de arco, posición en la palabra, etc., las pronunciaciones que contienen el menor número de arcos (si hay más de una) se dan un rango y las estrategias de puntuación escogen la mejor opción.

En esta tesis se han propuesto nuevas estrategias de puntuación, obteniéndose resultados prometedores. Una de las nuevas estrategias propuestas claramente supera a los demás. Además, las estrategias propuestas también aparecen seleccionadas al observar las mejores combinaciones de estrategias. Los mejores resultados para PbA son entre 63 y 88% palabras correctas según el léxico. Se obtienen resultados G2P no solamente para el inglés, sino también para otros idiomas europeos. También se ha considerado el caso del habla continua, adaptando la pronunciación para el habla continua del inglés, utilizando las llamadas formas débiles. Los resultados generales muestran que la capacidad de adaptación del módulo fonético ha sido mejorada.

Otra línea de investigación en esta tesis se encamina a aumentar la fiabilidad del módulo fonético. Aunque, los resultados experimentales para el G2P son bastante buenos, todavía existen errores que pueden impedir que la inteligibilidad de ciertas palabras y, por lo tanto, reducir la calidad del habla en general. Para lograr un mayor nivel de fiabilidad se propone utilizar la fusión de diccionarios. Las pronunciación de las palabras presentes en los distintos diccionarios depende de muchos factores, por ejemplo: opinión experta, especificaciones del acento local, alfabeto fonético elegido, nivel de asimilación (para nombres propios), etc. A menudo hay discrepancias entre la pronunciación de la misma palabra en diferentes

léxicos. Por lo general, estas discrepancias, aunque a veces significativas, no obstaculizan gravemente la pronunciación global de la palabra ya que todas las pronunciaciones léxico han sido previamente validadas por un lingüista experto. Estas discrepancias normalmente se encuentran en la pronunciación de vocales y diptongos. La sustitución de vocales por otras similares no se considera un error grave porque no afecta la inteligibilidad y por lo tanto la calidad de voz. El sistema de fusión estudiado es un sistema P2P que transforma las pronunciaciones del léxico de origen en pronunciaciones del léxico destino (el sistema está capacitado para aprender estas transformaciones). Para entrenar el clasificador, se seleccionan las entradas comunes entre el léxico fuente y destino. Se han realizado experimentos tanto para las palabras comunes como para los nombres propios, considerando los métodos de transformación basados en DT y FST. Los resultados experimentales muestran que la calidad del habla en general se puede mejorar significativamente dadas las bajas tasas de error de G2P y la amplia cobertura del diccionario del sistema. Un sistema TTS adaptable y fiable tiene que estar preparado para afrontar el reto del multilingüísmo, fenómeno que ya forma parte habitual de nuestras vidas cotidianas.

Esta tesis también ha considerado contextos que contienen la mezcla de lenguas, en los que la lengua puede cambiar de forma inesperada. Este tipo de contextos abundan en las redes sociales, foros, etc. Se propone un esquema de G2P multilngüe incluyendo la *nativización*. El primer componente de un TTS multilngüe es el módulo de identificación de idioma. Se ha desarrollado un identificador de idioma basado n -gramas (de letras) que proporciona buenos resultados. Los contextos en los que intervienen varias lenguas deben ser tratados con especial delicadeza. Por lo general, cada frase o párrafo tienen una lengua principal y las palabras extranjeras presentes en ella se consideran inclusiones. Al definir la estrategia sobre cómo pronunciar frases en varias lenguas puede partirse de dos escenarios: 1) aplicar a cada lengua un clasificador G2P distinto e independiente (que produciría cambios fonéticos bruscos que sonarían muy poco natural); 2) aplicar el clasificador G2P para el idioma principal de la frase suponiendo que esta pronunciación sería más aceptable que la que contiene fonemas extranjeros. Pero, ¿y si ninguno de los escenarios anteriores ofrece una calidad aceptable? Para países como España, donde el dominio de lenguas extranjeras por la población general es bastante limitado proponemos *nativizar* la pronunciación de palabras extranjeras en frases españolas. ¿Qué criterios se deben utilizar dadas las significativas diferencias en el inventario de fonemas? El objetivo ha sido obtener pronunciaciones que no son del todo desconocidas y que sean aceptadas tanto por hablantes nativos o con alto dominio del idioma extranjero como por hablantes de esa lengua con nivel medio o bajo. La nativización se lleva a cabo estudiando específicamente las inclusiones inglesas y catalanas en frases en castellano. Cuando hay diferencias significativas en los inventarios de fonemas entre las lenguas nativización presenta retos adicionales. Con el fin de validar rápidamente la idea de nativización se crearon tablas de mapeo de fonemas extranjeros a los nativizados y se llevó a cabo una evaluación perceptual. La nativización

basada en tablas muestra un mayor nivel de aceptación por parte del público que la síntesis sin nativización.

A fin de mejorar aún más los resultados de nativización de forma eficaz se propone aplicar un método basado en datos. Como gran parte de pronunciaciones extranjeras se aprenden por analogía, la aplicación del PbA a esta tarea es idóneo, sobre todo porque ya ha demostrado excelentes resultados para la tarea de transcripción fonética. Para ello se explora la analogía tanto en el dominio ortográfico como fonético. Todos los métodos basados en datos requieren un corpus de entrenamiento y PbA, por supuesto, no es una excepción. Ya que ningún corpus de nativización adecuado para la tarea estaba disponible se tomó la decisión de crear un corpus de entrenamiento y test para entrenar y validar nuestro clasificador para inclusiones inglesas en castellano y otro similar para las catalanas. Ambos corpus de entrenamiento contienen 1.000 palabras y son ortográficamente equilibrados. Se aplica la nativización por analogía basada en la forma ortográfica de la palabra G2Pnat y también basada en la forma fonética P2Pnat con el fin de nativizar palabras comunes y nombres propios en inglés y palabras comunes en catalán en frases en castellano. Los resultados obtenidos muestran que la analogía fonética da un mejor rendimiento que la analogía en el dominio ortográfico para los nombres propios y palabras comunes. Sin embargo, los resultados obtenidos para los nombres propios ingleses se encuentran unos 12 puntos porcentuales por debajo de los obtenidos para las palabras comunes en inglés. Esto es debido al hecho de que la pronunciación nombres propios está influenciada por factores más complejos e incluso para los seres humanos presenta importantes retos. El algoritmo TBL también se ha aplicado para mejorar los resultados de nativización para inclusiones inglesas. Se han obtenido mejoras tanto para los resultados obtenidos por P2Pnat, como para los resultados obtenidos por las tablas de nativización. Los buenos resultados obtenidos por el algoritmo TBL aplicado a la predicción del método ML demuestra la eficacia del método de aprendizaje a partir de errores también para esta tarea. En la evaluación perceptual llevada a cabo para ilusiones inglesas en castellano, se pidió a los oyentes que votaran el mejor de los tres métodos disponibles: G2P (para castellano), NatTAB y P2Pnat. P2Pnat es elegido como el mejor en el 50 % de los casos mientras que el G2P para español obtiene la mayoría de votos negativos (45 % de los casos). Estos resultados perceptuales así como los alentadores resultados objetivos demuestran la idoneidad de nativización para sistemas TTS multilngües.

Agradecimientos

Me gustaría agradecer a cada uno en persona pero por si no tengo esta oportunidad quería decir unas palabras de agradecimiento a todos que me han ayudado y acompañado en el viaje.

Quería agradecer al director de esta tesis, Toni Bonafonte, en primer lugar, por darme esta gran oportunidad al invitarme hacer la tesis doctoral en la UPC y por supuesto por su paciencia, comprensión, dedicación y continuo apoyo durante todo el proceso de mi maduración el la UPC.

De forma muy especial quiero agradecer a mis padres por el apoyo brindado, por la motivación y los buenos consejos, por compartir alegrías y preocupaciones conmigo y sobre todo por seguir creyendo en mí. Dedico esta tesis a vosotros con muchísimo cariño.

Esta viagem não seria a mesma sem ter o Eduardo ao meu lado. Nem saberia por onde começar por agradecer-lhe o seu apoio e atenção prestada durante tudo este tempo, por ter me ensinado a ser perseverante e paciente, qualidades imprescindíveis para a investigação cientifica e também para outros aspectos da vida. Poderia continuar e continuar. Um sincero obrigado de todo o coração.

Muchas gracias a mi amigo y compañero de despacho Dani, con quién empezamos el doctorado juntos y compartimos muchos cafés, comidas, cenas, viajes y además muchos debates lingüísticos interesantes. Muchísimas gracias por toda tu ayuda y tu amistad.

A todos los demás compañeros del despacho 215 Max, Yésika, José María, Jean-François, Coralí por hacer más agradable el día a día y hacer de nuestro lugar de trabajo un espacio más agradable y entretenido para investigar.

A todos los fantásticos compañeros del grupo veu: Asunción, Pablo, Javi, Jordi, Helenca, Ignasi, ha sido un placer compartir reuniones, cafés, viajes, congresos y más con todos vosotros.

Muchas gracias a mis amigos Andrey, Ania, Giovanna, Gregory, Israel, Katia, Mariella, Marina, Patrik y Victoria por todos los ánimos que me habéis dado a lo largos de los años y por ser excelentes compañeros y testigos de esta metamorfosis científica y personal.

Por supuesto no quería acabar sin agradecer a todos mis profesores de materias de carácter científico y lingüístico en la UPC.

Y por último quiero dar las gracias por el apoyo económico al Ministerio de Educación por la beca FPU concedida y a proyectos TC-STAR, ALIADO y AVIVAVOZ.

Contents

List of Tables

List of Figures

Acronyms

ASR	Automatic Speech Recognition 1
BDLEX	BDLEX dictionary
BEEP	BEEP dictionary of English 31
CART	Classification and Regression Trees 21
Celex	Celex dictionary 31
CMU	Carnegie Mellon University 21
DT	Decision Trees v
EM	Expectation Maximization 16
F_LANG	F_LANG label
FSA	Finite State Automata 47
FST	Finite State Transducers v
G2P	Grapheme-to-phoneme v
G2Pnat	Grapheme-to-phoneme Nativization vii
HMM	Hidden Markov Models v
LANG	LANG label
LTS	letter-to-sound 4
ME	Maximum Entropy 27
ML	Most-likely phoneme assignation vi
MLR	Machine learning 63
ML+TBL	Most-likely phoneme assignation and TBL
NETtalk	NETtalk xxv
NatTAB	Nativization Tables xxi
NatTAB+TBL	Nativization Tables (NatTAB) and TBL

OALD Oxford Advanced Learner's Dictionary 22

P2P Phoneme-to-phoneme vi

PbA Pronunciation by Analogy vi

POS Part of speech 22

P2Pnat Phoneme-to-phoneme Nativization vii

P2Pnat+TBL Phoneme-to-phoneme Nativization (P2Pnat) and TBL

Pronlex Pronlex dictionary 31

S2ST Speech-to-speech translation 2

TBL Transformation-based error-driven learning v

TTS Text-to-speech v

TWB Teacher's word book 31

Unisyn Unisyn 31

Chapter 1

Introduction

For many years scientists have dreamed of building machines able to converse with their creator by providing them with a measure of "intelligence" together with speech recognition and synthesis capabilities (Damper, 2001). Even if building a machine that would have human-like understanding and talking abilities is impossible, applications producing artificial speech are highly demanded on the market, especially now when the quality of the synthesized speech is much better than a decade ago. Speech technologies have as their main objective to ease the interaction between the humans and the computers. Speech technologies focus on the development of 4 major types of systems: Text-to-speech (TTS) systems that allow the computer to produce intelligible speech signal by imitating a human voice; Automatic Speech Recognition (ASR) systems that transform voice signal into a computer readable text therefore allowing to establish a bi-directional conversation; and automatic dialog systems that are equipped with special human-computer interface that allows to collect or provide information or to carry out different human-computer transactions; the latter system includes the first two as submodules. Automatic speech-to-speech translation systems are very popular and are probably the most perspective application of the speech technologies mentioned above. This kind of systems has the capacity to transform a speech signal in the source language into a speech signal in the target language, it includes, speech recognition, statistical machine translation, and speech synthesis. A voice conversion system maybe also necessary depending on the source-target speaker voice preferences. Voice conversion systems transform the speech waveform changing the voice of a source speaker to be perceived as if it was target speaker's voice. Llisterri (2003). The number of potential applications of TTS systems in the modern technological world is growing on a daily basis. Some examples are: reading aloud applications for people with vision impairments, voice production devices for mute people and in general for people with speech disorders, applications providing language learning assistance (pronunciation training), distance learning assistants, spoken dialog systems,

GPS navigation systems, spoken dialogue systems, video games, audio books, and a large number of applications in PDA systems, Smartphones and Androids, etc., and finally in Speech-to-speech translation (S2ST). S2ST systems have been very popular in the last years; the globalization phenomenon makes the knowledge of foreign language essential for successful professional and private life. That is why the S2ST systems have gained a world-wide popularity of the last decade. Learning languages is time-consuming, while high-quality S2ST application can solve communication problems and help overome language barrier in a number of limited fields. Would it not be nice if you said something in your own language and the device would say it for you in the language of the country you are visiting at the moment? Next imagine that the response would be spoken to you in our own language. The TTS makes it possible.

1.1 Introduction to text-to speech synthesis

Early electronic speech synthesizers had robotic voice and were often barely intelligible. The quality of synthesized speech has been steadily improving, but any output from contemporary speech synthesis systems is still clearly distinguishable from actual human speech. The first computer-based speech synthesis systems were created in the late 1950s, and the first complete text-to-speech system was implemented in 1968. Despite the success of purely electronic speech synthesis, research is still being conducted into mechanical speech synthesizers.

In human communication, voice has a very important role. However, not all of human speech production and control mechanisms are clearly understood. Human speech mechanism produces complex movements of the vocal organs and the vocal tract; the former are the lungs and the vocal cords in larynx, the latter consists of the tongue, the palate, the nasal cavity and the lips.

In the human speech mechanism, human vocal cords, which are thin folds on the larynx, generate sound as they are vibrated by airflow from the lungs. This source sound enters the vocal tract, which is a resonance tube constructed by the tongue, palate, nasal cavity, and lips. The vocal tract articulates the source sounds into vowels and consonants. This basic mechanism has been clarified; however, the control of the complex movement of speech organs and detailed phonemes in speech production are still unclear.

The two rapidly developing and most successful branches of speech synthesis are concatenative synthesis and synthesis using Hidden Markov Models.

Synthesis using Hidden Markov Models (HMM)s models the spectrum, excitation and duration simultaneously using context-dependent HMMs. The speech waveforms are

generated from the HMMs themselves based on the maximum likelihood criterion Zen et al. (2007).

Formant synthesis and articulatory synthesis do not use pre-recorded human speech databases and do not reach any level of naturalness that could be mistaken for a human speech, however, they are more flexible than concatenative methods in some aspects, such as intonation and emotions control and are clearly preferable for those application where acoustic glitches at concatenation boundaries present a bigger issue than the lack of naturalness.

A TTS system, is a computer based speech system capable of transforming the input text into intelligible natural speech signal. The TTS consists of a *front-end* part that converts texts to linguistic specification and a waveform generator that uses this linguistic specification, a sequence of phonemes annotated with contextual information, to generate a speech waveform.

The *front-end* is responsible for three main tasks: text analysis, phonetic transcription and prosody specification. Phonetic transcription is also necessary for automatic speech recognition and computer assisted language learning, however this thesis has been carried out entirely in TTS framework.

Text analysis consists in sentence and word separation followed by normalization and disambiguation of non-standard words. The raw input text can contain a number of elements that prior to further processing need to be disambiguated. The normalization of non-standard words usually is a very complex task and includes several language dependent problems (Sproat, 1996). Numbers, abbreviations, dates in different formats, special symbols designating monetary units, such as € and $, etc., generally need to be expanded to full words. In case of acronyms, if there is no commonly used expansion, of if the acronym is more usual than the expansion, they can spelled out or pronounced "as word", for example, North Atlantic Treaty Organization is better known as NATO (Sproat et al., 2001). When dealing with abbreviations, finding the correct expansion is not always as easy. In English, the abbreviation "Dr." can be expanded into doctor or drive while "ft." can be transformed into fort, foot or feet. For highly inflected languages such as Swedish, Czech and other Slavic languages the task of correctly expanding non-standard tokens into full words is even more complex and requires additional linguistic knowledge.

Grapheme-to-phoneme (G2P) conversion system is responsible for converting words from a normalized text into their corresponding phoneme forms. The difficulty of such conversion is highly language-dependent. English, the language with huge research interest is a language with deep orthography, in other words, with no obvious grapheme-to-phoneme correspondence. Spanish, however, is a language with very shallow (opposite to deep) orthography where the correct pronunciation of common names can be derived using a

rather simple rule based G2P converter. The situation with proper names is different, especially those of foreign origin. Here, simple letter-to-sound (LTS) rules were found to be insufficient.

It is important to emphasize, that pronunciation of foreign proper names is a very difficult task even for human beings because they differ from other words morphologically, orthographically and of course phonetically. Proper names comprise a large portion of unknown words in the grapheme-to-phoneme module. Sometimes the correct pronunciation depends on many variables and unpredictable factors such as personal wish of the bearer, the level of assimilation, etc.

Grapheme-to-phoneme correspondence is a crucial issue when it comes to synthesizing speech. For the languages with non-transparent writing system the transcription task is usually assisted by an available pronunciation lexicon. A lexicon or pronunciation dictionary is usually a two-column document featuring words and their pronunciations. Sometimes alternative forms of pronunciation and/or part-of-speech tags are included.

Furthermore, knowledge of syllabic structure of the words is helpful in the composition process of concatenative synthesis the duration of a phone is affected by its location within a syllable and correct syllabification influences the correct pronunciation of a word just as much as the correct identification of its phonemes (Marchand and Damper, 2007). Syllabification can be inferred from orthographic input simultaneously with the pronunciation or as a separate process. Sometimes better syllabification is achieved when it is derived from the phonetic form (Weisler and Milekic, 2000). It is believed that the stressed syllables have longer duration and are considerably louder than the unstressed ones. Lexical stress should be distinguished from phrasal stress, the first one only can be spoken of in case of isolated words. When words form sentences the lexical stress of each words is not usually kept, moreover the stress can be moved to another syllable, previously unstressed. Most languages do not have a fixed lexical stress position, therefore, it is very important to be able to predict both lexical and phrase stress correctly in order to strive for a more natural synthesized speech (Dutoit, 1997). The correct stress placement influences not only the pronunciation but also prosody. Connected speech also presents some pronunciation variation in comparison with pronunciation of isolated words, articulation phenomena at word boundaries, liaisons, fast speech consonant and vowel deletions are some of the factors influencing the pronunciation of words in sentences in this case some phonetic post-processing is needed.

After the the words have been converted to phonemes the prosody generation module uses previously trained models to assign the correct rhythm, intonation, duration and other related attributes to the phoneme form obtained in the previous modules. The above listed values are extracted from the corpus used for training the system. The last module generates the final output of the TTS system, the waveform.

The branch of concatenative synthesis that allows a good ratio of naturalness versus flexibility is the unit selection synthesis which uses large databases of pre-recorded human speech. The pre-recorded units can vary from large units such as sentences or phrases to much smaller units such as diphones or even half-phones. Speech database is aligned with the waveform and the segmentation of the database is achieved using a speech recognizer and a text transcript of the recordings. The units are classified by different criteria such as fundamental frequency, duration, surrounding context and position in the sentence or the syllable.

Using all the information obtained from the previous two modules, the incorporated unit selection system (used in high quality TTS systems) is applied in order to find the most appropriate segments from the available speech database, for their further concatenation into synthetic utterances. The best units for the target utterance are usually selected using a weighted decision tree or a similar classifier. The architecture of a standard TTS system is shown in Figure 1.1.

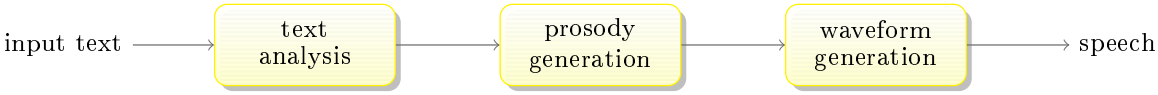


Figure 1.1: Architecture of a TTS system.

Sometimes words can be uttered in several ways and it is important to have methods capable of providing different alternatives of pronunciation of the same word. This is of a particular importance for speech recognition and for concatenative speech synthesis.

1.2 Phonetics and letter-to-sound rules

There are languages where the orthographic transcription (letters) and the phonetic one (phonemes) are quite different. A word can have more phonemes than letters, but usually there are more letters than phonemes. The languages where this phenomenon occurs are said to have deep orthography, and their pronunciation is "non-transparent".

In English, generally, there are more letters than phonemes as in the following example: the word each is pronounced as `/i: tS/`, in some case although more rare ones there are more phonemes than letters as in fox `/f A k s/`, therefore no one-to-one correspondence can be inferred. This phenomenon occurs due to the natural delay in the evolution between the spoken language and the written one (Dutoit, 1997). Another difficulty is represented by homophones, words that have different orthographic representation but the same pronunciation as in *made* and *maid* in this case both are pronounced as `/m e I d/`. The final letter *e* in *made* is not pronounced however it indicates that the previous syllable

is open and it should be read as `/e I/` opposed to `/ae/` in *mad*. English alphabet has only 26 letters but they represent around 45-60 phonemes, depending on the chosen *phonetic alphabet*.

Each abstract unit is called phoneme, and even though a language has a fixed set of phonemes, their particular pronunciation depends on contextual effects, speaker's individual voice characteristics, his mood and his intentions. The acoustic representation of a phoneme is called *phone*. In continuous speech the pronunciation of each phone is influenced by the articulation of the neighboring phones. The pronunciation variations caused by the phone context are represented by allophones. Allophones represent the same phoneme pronounced with different vocal tract configuration.

Each language has a fixed set of phonemes (not considering allophones) which usually varies from 20 to 60 units. The mean size of phonetic alphabet for the Western and Southwestern Eurasia is 36.50 phonemes according to Donohue and Nichols (2011). The phoneme inventory or the phoneme alphabet is the graphic representation of phonemes. Each alphabet is usually divided into vowels and consonants. Vowels are the voiced sounds produced by the vibration of vocal cords. Meanwhile consonants may not contain the voiced part. This is one of the reasons why consonants are more difficult to synthesize, they have low amplitude, short duration and significant articulatory changes at the boundaries. One of the most widely used phonetic alphabets is speech technologies is SAMPA (Speech Assessment Methods Phonetic Alphabet)(Wells, 1997) and IPA (International Phonetic Alphabet) (Handbook, 1999). However, speech experts believe that SAMPA is much more convenient to use for automatic computer processing than IPA (Dutoit, 1997).

IPA is a system of phonetic notation based on the Latin alphabet, devised by the International Phonetic Association as a standardized representation of the sounds of spoken language(Handbook, 1999). The IPA is used by foreign language students and teachers, linguists, speech pathologists and therapists, singers, actors, lexicographers, and translators. IPA is designed to represent only those qualities of speech that are distinctive for spoken language: phonemes, intonation, and the separation of words and syllables. To represent additional qualities of speech such as tooth gnashing, lisping, and sounds made with a cleft palate, an extended set of symbols called the Extensions to the IPA is used.

SAMPA is a machine-readable phonetic alphabet. It was originally developed under the ESPRIT project 1541, SAM (Speech Assessment Methods) in 1987-1989 by an international group of phoneticians, and was applied in the first instance to the European Communities languages Danish, Dutch, English, French, German, and Italian (by 1989); later to Norwegian and Swedish (by 1992); and subsequently to Greek, Portuguese, and Spanish (1993). Later Hebrew, Russian, Thai, Croatian, Cantonese, Arabic and Turkish were added. Unlike other proposals for mapping the IPA onto ASCII, SAMPA is not one single author's scheme, but represents the outcome of collaboration and consultation among speech

researchers in many different countries. The SAMPA transcription symbols were developed by or in consultation with native speakers of every language to which they were applied, and are standardized internationally.

The challenge of phonetic transcription was approached in many different ways. Languages are expanding continuously, and most of the dictionaries contain entries restricted only to morphemes, thus making it impossible to have a lexicon with all the words in it. The lack of coverage of the pronunciation lexicon and the need for accurate pronunciation raises a demand for a grapheme-to-phoneme converter.

There are several ways to approach pronunciation derivation for out-of-vocabulary words, some of them combine morpheme-lexica and a set of phonemization rules, others are based on a set of letter-to-sound rules and and exception lexicon that contains all the words that do not obey the defined set of rewrite rules.

However, for languages with deep orthography, manual crafting of the pronunciations or letter-to-sound rules has elevated costs and is not quite suitable for real-time applications. This is the main reason why over the past decade there has been a growing tendency to use data-driven methods to obtain the pronunciation of out-of dictionary words. Moreover, the traditional context-dependent computer-applicable rewrite rules were found to be very inefficient. Usually, more than one rule applies at each transcription stage. The order of the rules application is crucial for the final result. In general, the rules are ordered from more specific to less specific. These rules and their order should be provided by an expert linguist. These rules are unadjustable to the vocabulary rapidly expanding with neologisms and are certainly absolutely language specific. Eventually, the traditional rules were left behind by many speech researchers and substituted by the automatic methods for grapheme-to-phoneme conversion. The possibility of using such numerical measures as grapheme-to-phoneme correspondence probabilities, frequency counts, etc. makes it possible to improve the robustness and the flexibility of the G2P methods without a need to elaborate and maintain complex rules and rule-ordering schemes. Data-driven statistical algorithms are much less time-consuming and can function as fully language-independent, easily adaptable to the changes within any language. Nowadays, many automatic G2P systems are used to infer pronunciations of the unknown words, the results reported are good but still there is room for further improvement. There is a need for a comparative evaluation in order to choose the best technique in each specific situation. Automatic G2P methods find their applications in text-to-speech synthesis, automatic speech recognition and language learning.

1.3 G2P in Speech-to-Speech translation

To speak your native language is the most natural way of communication and interaction between humans and between humans and machines. The variety of languages spoken in Europe is a result of its cultural and historical diversity. Nevertheless, this diversity can create development barriers for the speech-centered applications. The speech translation, after getting rid of the linguistic barriers between users of different origins, would favor the economic development of many areas such as tourism, client service and others. Voice generation in the framework of the speech-to-speech translation sets its own goals. It is necessary to emphasize that we are referring to the statistical translation systems, which are able to translate the text represented in the language of origin into its superficial representation in the target language, without using the intermediate conceptual representation. The statistical translation is able to offer better results in wide semantic domains as well for text input as for the cases when the input is speech. The first special feature is that in the speech translation the generated language can be closer to the spoken language than to the written one. All the research work that has been carried out in speech synthesis was focused on reading of texts. Each text corresponds to a correct grammatical structure and has to be read in a certain pre-established way. Nevertheless, there are many other interesting applications of speech synthesis. Usually the people do not "read" but "speak" and the spoken language is much more expressive than reading. The synthesis of the informal speech finds its application in many areas, such as entertainment, education and many others. Lately, such synthesis systems form part of speech-to-speech translation systems, where the study of prosodic and acoustic models is necessary. Usually when translating the news from one language to another it could happen that some concepts used in the source language could be unknown for the listeners of the target language. Dealing with this phenomenon requires elevated cognitive levels but still it is possible to take it into consideration when expanding abbreviations or when reading names of persons or entities that are not so common in the target language. The proper names present a particular interest. However, the speech translation has an advantage with respect to voice generation from text: the source language is known. This information can be used in order to adapt several parameters of the synthetic speech, as for example proper name pronunciation. For human listeners a foreign accented voice presents a barrier of intelligibility. Common proper names are understood better than the rest of the text since they are heard more often by the listeners, but the rare ones present an important comprehension problem (Tomokiyo et al., 2005). As proposed in Chapter 3 one of the solutions might be opting for a nativized pronunciation, but it could happen that drastically changing the pronunciation according to the target languages G2P rules may make it unrecognizable that is why the original waveform of the proper name in the source language may be helpful to find the optimum pronunciation. To decide whether it is necessary to use this information about the name

frequency and phonetic distance measures between the source and proposed target language pronunciation the may be used. Since there is no suitable corpus available for the task described above the first necessary step will be the creation of a bilingual acoustic and a bilingual phonetic corpus.

1.4 Framework of this thesis

In Spain, text-to-speech synthesis and other speech technologies have been given a lot of attention in the framework of national and international projects. The research carried out at the UPC was mainly focused on synthesizing speech in English and in multilingual scope. An important international project TC-STAR that partially involved the development of this dissertation was carried out in cooperation with leading speech technology developers in Western Europe both in the commercial and academic spheres. It was aimed at the speech-to-speech translation of European Parliament transcriptions from Spanish to English and vice versa. Speech synthesis and speech recognition components were enhanced in improved to suit the framework (Bonafonte et al., 2005). Other projects that involved speech synthesis or speech recognition carried out at the UPC were: ALIADO - Speech technologies for a personal assistant. ALIADO undertakes the developing of spoken and written language technologies for the design of personal assistants in a multilingual environment (Mariño and Rodríguez, 2003). TECNOPARLA - this project focused on the development of speech recognition and speech synthesis systems for Catalan specially adapted for the phone applications. AVIVAVOZ -supported the research in the field of speech technologies that are comprised inside speech-to-speech translation technology. The goal of the project was to develop state of the art speech technologies for the 4 official languages of Spain (Marino, 2006). BUCEADOR- Is the continuation of the AVIVAVOZ project, it is focused on advanced research in all core Spoken Language Technologies (SLT), (diarization, speech recognition, speech machine translation, and text-to-speech conversion). The goal of the project is to achieve improvements in all the SLT components to improve human-machine and human-to-human communication among all the official languages spoken in Spain as well as between these languages and English. The active project *SpeechTech4All* includes multilingual speech synthesis and integrates translation results. The technology produced in this thesis is being applied to this project.

The raised demand for high quality and flexibility in the field of speech technologies motivates the researchers to seek further improvements is such areas as continuous speech recognition, speaker identification and verification, voice conversion and translation, which, however, being a rather new research field already presents a special research interest due to increased the mobility and globalization phenomenon. Inside the S2ST framework there

is a need for a good automatic grapheme-to-phoneme module as the presence of grapheme-to-phoneme errors for languages with complex orthography does not allow good quality of the TTS system therefore it affects negatively the overall performance of the translation system.

1.5 The objectives of the thesis

Modern TTS systems use the dictionary look-up as the primary method to derive the pronunciation of the words but since it is not possible to list exhaustively all the words present in one language and also due to the entrance of new foreign words a backup strategy for deriving the pronunciation is needed. Since in terms of time and effort automatic data acquisition methods are the most economic ones, the application of machine learning methods to the G2P is the most appropriate choice.

The first objective of this thesis is to analyze and compare different state of the art language-independent G2P conversion methods, propose improvements for the existing techniques finding an automatic and efficient way to improve the pronunciation of unknown words, making an important step towards producing high quality synthesized speech, where naturalness and intelligibility are considered to be very important desirable aspects of synthesizer performance. Intelligibility is the obvious requirement that any synthetic speech should meet, but a high degree of naturalness has been widely accepted as easing listening strain. It is important to have different grapheme-to-phoneme methods evaluated using the same training and test data. This will allow a better understanding of the advantages as well as of the weak points of each method. This brings us to the next objective of this researh work - to provide consistent evaluation of the G2P methods on the same datasets.

In order to further improve the quality of the synthesized speech, the size of the system dictionary used as a back-up pronunciation strategy in TTS can be increased. However, the lexica created by different experts are incompatible and they usually use different phoneme inventories. Another objective of this thesis is to find a reliable way to improve the compatibility of the lexica with the goal of creation of a larger look-up system ditionary. This objective is based on the hypothesis that for some words, especially for proper names, pronunciations from less realiable lexica are less damaging to the speech quality that the automatic pronunciations.

The G2P errors for languages like Spanish and Catalan are mostly derived from the mispronunciation of foreign words and foreign and loaned proper names. Although, for languages with complex writing systems like English the pronunciation of novel words is problematic, proper names represent the main source of the most severe pronunciation errors. Texts written in several languages are a rapidly spreading phenomenon that

should not be ignored when talking about high quality speech applications. Worldwide globalization is responsible for an entirely new form of multilingualism present in all types of communication resources. This problem reaches a very important level in such countries where there are several official languages, e.g. Spain or Switzerland. Types of mixed-language inclusions vary from word parts to entire sentences. Therefore a language identification approach is needed prior to phonemization. This thesis sets the goal of language identification of the inclusions and application of the language-independent G2P conversion methods to the multilingual scope.

The problem of pronunciation of foreign words and phrases in a text almost entirely written in another language (target language) needs special attention since for some languages it is unusual to have foreign phonemes in a native language sentence. In Spanish, the usage of correct foreign pronunciations, for instance, it is very unnatural and even is considered pretentious due to considerable differences between Spanish phonetic inventory and phonetic inventories of other languages. However, the other extreme, would be correct either, e.g. pronouncing the foreign name *Jackson* `/dZ 'ae k s @ n/` according to the Spanish letter-to-sound rules would result in `/x 'a k s o n/` an unintelligible and unnatural pronunciation. Next objective deals with finding the pronunciation of the inclusions, which language has been identified. A balance between the two extremes of foreign words pronunciation should be found in a way that it would be acceptable by both native and foreign audiences.

1.6 Thesis overview

The rest of the document is organized as follows. Chapter 2 reviews the state of the art techniques of automatic phonemization of out-of-vocabulary words. Among these there are Decision Trees (DT), Finite State Transducers (FST), Hidden Markov Models (HMM), Pronunciation by Analogy (PbA), and latent analogy. The main advantages and disadvantages of each method are pointed out. The most important G2P results found in the literature for different conversion methods and lexica, are summarized. Since this work is intended for multilingual framework and its final goal is to be able to synthesize speech from multilingual texts, the language identification task is described as well as proper names pronunciation, which adds even more difficulty to the task. Previous approaches to pronunciation of foreign words are described towards the end of the chapter.

In Chapter 3 different automatic and language-independent methods for G2P conversions are studied and compared. In the first place the experimental setup is defined in such a way that the results obtained would be comparable with those found in the literature and also among them. The methods are studied and compared, and several improvements are proposed with the goal to obtain better performance. The errors obtained

by the best-performing methods (Transformation-based error-driven learning (TBL) and PbA are analyzed and conclusions drawn. Main differences in the pronunciation generation for connected speech and isolated words are explained in detail.

Usually, the more words you have in your look-up dictionary in the phonetic module of the TTS system, the less is the probability of getting erroneous transcriptions, however, it is not possible to merge several lexica without any adaptation. The main reason for this lies in lexicon incompatibility. Different pronunciation databases are transcribed using different expert criteria and different phonesets without a trivial one-to-one phoneme symbol correspondence between them. The goal set in Chapter 4 was to investigate these differences and propose a way to automatically homogenize the pronunciations in the lexica.

The number of multilingual texts is increasing on a daily basis because of the globalization phenomenon that affects all spheres of human activity. The information circles the globe at the speed of light and multilingual sources are becoming more and more common to people from all corners of the planet. For speech researchers it raises an urgent need to be able to offer not only language independent tools, but also those that would be able to process mixed information. In TTS processing of multilingual texts requires several additional components. First of all, a language identification tool is needed. Once the language is determined the following step is to derive the pronunciation for the words or sentences in the language identified. At this point, depending on the language, it is necessary to decide whether or not the pronunciation sought should be adapted to the main language of the text in any way. Spanish requires this type of adaptation, which is called "nativization".

Proper names present an additional challenge since their pronunciation is highly dependant on their language of origin and the level of assimilation they had undergone. Chapter 5 closely studies these issues from the point of view of English and Catalan inclusions (words that come from other languages) in texts where the main language is Spanish. A method for pronunciation of inclusions in Spanish language is proposed. Subjective and perceptual evaluation results are given and discussed.

Chapter 6 concludes the dissertation work by giving a full summary of the achievements attained in this thesis and possible research directions for future work.

The research work presented in this dissertation addresses several important questions regarding the evaluation and efficiency of the phonemization methods in the framework of multilingual text-to-speech synthesis in the modern world.

Chapter 2

State of the art

This chapter provides an overview of the state of the art G2P methods, and also and gives a review of the issues and factors influencing the accuracy of G2P conversion in mono and multilingual scopes, language identification as well as corpora suitability. It also summarizes the most significant works from the rapidly growing field of TTS that are relevant to this Ph.D. research.

2.1 Grapheme-to-phoneme conversion methods

Text-to-speech synthesis systems use several tools for automatic phonemization of input words. Usually, for languages with a rather deep orthography, such as English or French, a system lexicon forms an essential part of the phonetic module of the TTS system. For each input word, a dictionary-look up procedure is performed and the corresponding pronunciation assigned. However, the pronunciation dictionaries cannot cover the continuously expanding language and, therefore, are not able to list all the words exhaustively. Neologisms, foreign proper names, unlisted morphological forms, etc., are the most common sources of our-of-vocabulary words. Thus, in order to have an unrestrained and transferable to new domains TTS system, it is necessary to have a tool or a method capable of deriving pronunciations for the words absent from the system lexicon. In modern TTS system, different backup phonemization approaches are used, these methods are either knowledge-based or data-driven machine-learning methods. While knowledge-based systems are costly and highly time-consuming other than non-transferable to other languages, machine-learning methods are flexible, language independent and rather inexpensive. In this section, both knowledge-based and data-driven approaches to phonemization are reviewed as well as different techniques for lexicon alignment. In Section 2.1.1 knowledge-based approaches to automatic phonemization are reviewed briefly, later Sections 2.1.2 through 2.1.8 are focused on data-driven approaches to G2P conversion.

2.1.1 Knowledge-based approaches

For languages with shallow orthography the challenge of G2P conversion is usually approached by series of ordered rewrite rules. These rules are proposed by expert linguists and are context dependent. Usually they are represented in the following form:

A [B] C $\rightarrow$ D (Chomsky and Halle, 1968),

where B is the central letter, D the corresponding phoneme, A and C the surrounding left and right context. Rule-based approaches to phonemization are often used in TTS systems as an alternative for dictionary look-up, since they were extensively studied long before computers had gain a center place in the development of the mankind (Dutoit, 1997). Rules may involve different linguistic characteristics such as: syllable boundaries, part-of-speech tags, stress patterns or etymological origin of a word.

Pronunciations of words vary according to different conventions adopted by expert linguists, and for the same language, there are usually several expert systems. These systems use different phonemic alphabets, rule formalisms and disambiguation criteria. Typical letter-to-sound rule sets are described in (Ainsworth, 1973; Elovitz et al., 1976; Hunnicut, 1976).

To derive the pronunciation for the input word, the rules are applied in the order that they appear in the rule list. This order is established by the experts and usually goes from the most specific rule to the least specific one. Whenever several rules exist for the same letter in different contexts the rule that appears at the top of the list is applied in the first place. The words are usually scanned form left to right and the rule triggers are searched. Every time a rule match is found a phoneme is output and the search window is shifted to the right N characters, N being the number of characters that were necessary to trigger the rule. If no match is found, the size of the sliding windows is decremented and the rules are scanned again until a match triggers the rule. The default rules are based on single characters, therefore a match is always found. The larger character clusters are given priority when scanning, therefore, every time the window is shifted after having emitted a phoneme, its size is reset to the maximum value. In English, consonant clusters are usually converted first, as for vowels, the letter-to-phoneme correspondences are rather ambiguous and they account for the main part of the errors. Previously converted consonants can be used as a part of the context for converting vowels.

Damper et al. (1998) evaluated *Elovitz's et al.* rules on a *Teacher's Word Book* dictionary of 16280 words (Thomdike and Lorge, 1944). The word accuracy as low as only 25.7% was achieved. This result is very different from the 80-90% accuracy reported by Elovitz et al. (1976). This can be explained by the fact that (Damper et al., 1998) used a stricter evaluating technique that did not classify pronunciations not containing any severe errors as "good" pronunciations. Also, this later evaluation was performed on TWB dictionary that

uses a phoneset of 52 phonemes, while the rewrite rules include only 41 phoneme symbols. Such a discrepancy in phoneme inventories may be one of the main causes of errors.

Rule-based system require hiring an expert linguist and therefore have a high production and maintenance cost, they clearly lack in flexibility and are language-dependent. Moreover they do not take into consideration any kind of statistical measures such as rule probability, frequency counts, etc., that could be helpful in order to improve robustness (Damper, 2001). In the last two decades data-driven approaches have been widely used to solve the problem of automatic phonemization. They are flexible and mostly language-independent, which makes them a perfect alternative to rule-based approaches. Older TTS systems used automatic phonemization methods because memory limitations made the storage of large lexica impractical (Taylor, 2005). However, some data-driven techniques are in many ways similar to hand-written rules. For example, decision trees and the questions formulated at each iteration can be represented as a set of context-sensitive rules. The key difference between data-driven techniques and hand-written rules lies in the rule development. In the first case, the rules are learned directly from data and in the latter case, the rules are crafted by expert linguists.

In the rest of the sections data-driven language-independent approaches to G2P conversion are reviewed in detail. The main results found in literature are given in Table 2.3. However, different alignment tecniques are summarized in the first place since the alignment is required by the majority of the automatic G2P converters.

2.1.2 Alignment

Almost all automatic G2P methods require the training data to be previously aligned. The alignment is the correspondence between the orthographic and the phonetic forms of the word. For earlier experiments such as those described in (Sejnowski and Rosenberg, 1987) and (McCulloch et al., 1987) manual alignments were used, however manual elaboration of alignments is very costly and language-dependent. The use of automatic alignments is preferable because it is the best solution in terms of time and cost. For languages with deep orthography (e.g. English and French) automatic alignment is a difficult problem mainly due to the lack of transparency in the writing system of these languages (Damper et al., 2004). The lack of clarity in the English orthography adds complexity to the alignment task since any phoneme can potentially align to a maximum of 4 letters (Taylor, 2005). The cases of 4-to-one correspondences are not so common but 2-to-one are numerous, for example, who [h u]. The cases where one letter aligns to more than one phoneme are less frequent but also deserve special attention. An example is the word *six* [s ɪ k s].

Letters	M	E	A	D	O	W	S
Alignment 1	m	E	_	d	o	_	z
Alignment 2	_	_	m	E	d	o	z

Table 2.1: Possible alignment candidates.

One-to-one alignments

Automatic epsilon-scattering method can be used to produce one-to-one alignment (Black et al., 1998b). For the cases where the number of letters is greater than that of phonemes a "null" phoneme symbol is introduced into phonetic representations to match the length of grapheme and phoneme strings. Firstly, the necessary "null" symbols are added into all possible positions in phonetic representations. This process is repeated for every word in the training lexicon. In Table 2.1 we give an example of two of the possible alignment candidates for the word *meadows*.

Such probabilistic initialization allows obtaining all possible imperfect alignment candidates. The goal of epsilon-scattering algorithm is to maximize the probability that letter g matches phoneme ϕ and, therefore, to choose the best alignment from possible candidates. It is done by applying the Expectation Maximization (EM) (Dempster et al., 1977). The EM is associated with joint grapheme-phoneme probabilities. Under certain circumstances, the EM guarantees an increase of the likelihood function at each iteration until convergence to a local maximum. The obtained alignments are not always logical, e.g., the word *through* may be in some cases aligned to [θ r _ _ _ u]. This alignment imposes the correspondence between grapheme h and phoneme [u], which introduces additional ambiguity to the training data. One way to overcome this obstacles is to build a list of allowables as in Black et al. (1998b). It is a simple table, that does not require any expert knowledge of the language. The allowables table defines for each grapheme a set of phonemes to which they can be aligned. All other alignments are prohibited. Some words with very opaque relationship between letters and phonemes would require adjustments made in the allowables table in order to produce alignments.

Another way to find a relationship between letter and phonemes is to use dynamic programming(DP).

DP based alignment uses a letter-phoneme association matrix A, of the dimension $L*P$, where L is the size of the letter set and P is the size of the phone set. At the first step the matrix A is initialized in a naive way with the elements ${a_{\mathrm{l,p}}}^0$ which are incremented each time the letter l and the phoneme p are found in the same word. At the next iteration ${a_{\mathrm{l,p}}}^1$ are incremented if the letter l and the phoneme p are found in the same alignment position.

Letters	S	P	EA	K	ING
Phonemes	s	p	i	k	IN

Table 2.2: An example of an alignment based on graphones.

At this first iteration the nulls are introduced into the dictionary as a consequence of the DP matching where both phonemes and graphemes can be associated with nulls. At the EM step the matrix A is updated in a way that the word alignment score is maximized. Nulls are not entered as a part of the updated matrix A in order to avoid the tendency to generate unnatural alignments. The role of nulls is restricted to the DP matching phase. The DP matching phase can be considered a path-finding problem. Dynamic programming is more efficient than epsilon-scattering method and it allows nulls in both letter and phoneme strings. The alignment x to [k s] is done automatically while the epsilon scattering method requires an a priori introduction of double phonemes [k] [s] to [ks].

Many-to-many alignments

Finite state transducers (Section 2.1.6) and multigram models can use many-to-many alignments. In (Bisani and Ney, 2002; Chen, 2003; Deligne et al., 1995; Galescu and Allen, 2001) the authors used G2P alignment as the first step to infer the pronunciations of unknown words. Bisani and Ney (2002) baptized the alignment element as "graphone", or a grapheme-phoneme joint multigram, which is a pair $q = (g, \varphi) \in Q \subseteq G* \times \Phi*)$. Letter sequence and phoneme sequences can be of different length (G and Φ are the grapheme and phoneme sets respectively). An example is shown in Table 2.2.

Those graphones that map one phoneme to one letter are called singular graphones. Graphone alignments can be inferred by using hand-crafted rules, dynamic programming search with predefined alignment constraints or costs, or by an iterative estimation of alignment probabilities.

The best sequence of graphones is induced from the dictionary data by searching for the most probable sequence of graphones, first assigning uniform distributions to all possible graphones (within the manually set length constraints) and then applying the EM algorithm. After graphones are aligned joint multigram sequence model is applied to automatically derive pronunciations (Bisani and Ney, 2008; Chen, 2003; Galescu and Allen, 2001).

Some of the automatic phonemization methods, however, do not require alignments since the letter-phoneme correspondences are calculated during the training, e.g., Hidden Markov Models (HMM) use Baum-Welsh training (Jelinek, 1997).

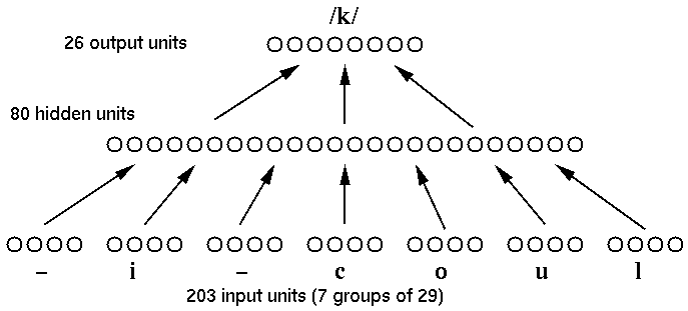


Figure 2.1: Schematic drawing of the NETtalk network architecture. A window of 7 letters in an English text is fed to an array of 203 input units. Information from these units is transformed by an intermediate layer of 80 "hidden" units to produce patterns of activity in 26 output units. The connections in the network are specified by a total of 18629 weight parameters (including a variable threshold for each unit) (Sejnowski and Rosenberg, 1987).

2.1.3 Neural Nets

One of the first and well-known approaches to automated G2P conversion is the NETtalk system created by Sejnowski and Rosenberg (1987). The authors were pioneers in applying back-propagation neural networks (Rumelhart et al., 1986) to a learning problem with such an important practical application. NETtalk system was designed as a feed-forward multi-layer perceptron with three layers of units and two layers of weighted connections. This architecture features an input layer of letter units, a hidden layer used for optimal feature detection, and an output layer of phoneme units. The input layer received a 7 letter window, where the central letter represented the source output letter being the other 3 letters to each side a way to define a context of surrounding graphemes. Each input letter was represented by a code of 29 bits, 1 for each of the 26 letters of the English alphabet and 3 additional bits for the punctuation marks and word boundaries. Therefore, the total count of units in the input layer was equal to 7x29 = 203 units. Phonemes were represented using a set of 21 articulatory features, using 5 additional bits to encode, stress level and syllable boundaries. The articulatory features that represented the phonemes in the output layer were voicing, point of articulation, vowel height, etc. Altogether the number of output features used to represent 51 output phonemes was equal to 26.

The number of units varied along the experiments, although the authors report that 80 hidden units were found to be a good match point between good performance and rather low computational complexity. However the best results were obtained using 120 hidden units.

The network architecture is shown in Figure 2.1.

The goal of the learning procedure is to minimize the average square error between the truth $s*_{\mathrm{i}}$ and the output values (1): $Error = \sum_{i=1}^{J}(\dot{s}_i - s_i^{(N)})^2$ (1) where J is the number of the output units, and N designates the output layer. The updated weights are calculated in (2): $w_{ij}^{(n)}(t+1) = w_{ij}^{(n)}(t) + \varepsilon\Delta w_{ij}^{(n)}$ (2), where t is the number of weight updates and l is the learning rate (typically 1.0).

The output phoneme was calculated from the set of vectors that were used to encode each phoneme, taking the inner product of the output phoneme vector code and each one of the vectors from the desired phoneme inventory, the desired phoneme that formed the smallest angle with the output phoneme was chosen as the "best guess", in cases when the difference was as small as 0 or 1, it was said that the "perfect match" was found.

As most of the automatic G2P systems, the NETtalk requires the data to be aligned in a one-to-one manner. Sejnowski and Rosenberg manually aligned a 20,012 word corpus created from Merriam Webster's Pocket Dictionary (Sejnowski and Rosenberg, 1993). When there number of letters exceeded the number of phonemes in a word, the so-called *silent phonemes* were introduced, in the opposite case new double phonemes were invented , e.g., the phonemes [k] and [s] in the word *axes* were joined in one [æ k_s].

The system was both trained and tested on continuous speech and isolated words from the dictionary. The continuous speech corpus of 1024 words featured alternative pronunciations for the same words. The best results achieved in terms of phoneme accuracy when tested on the continuation of the corpus (439 words) were 78% best guesses and 35% perfect matches. The system was also trained on a 1000 word subset from a 20k corpus of most common English words. The number of hidden units varied across the experiments. The best results on the training corpus were obtained using 120 hidden units. The same number of units was used to test the network on randomized version of 20,012 word dictionary, and the best performance was 77% of best guesses and 28% perfect matches.

One of the advantages of the method is that it is language independent. Only the input and truth vectors need to be adapted. The phoneme error rate allows concluding that the network has well captured the complex structure of English language. This system has strong similarities to human learning and memorizing processes, however, it does not come close to modeling human reading acquisition skills.

However, there are some points of downfall to the evaulation method used (Damper, 2001). The problem of automatic G2P conversion was simplified since the letter-to-phoneme alignment was carried out manually. The biggest issue with the results presented by the authors is that the generalization ability of the system was never tested on a totally unseen set of words,. Furthermore, phoneme error rate is not a good enough measure to compare the methods since the quality of synthesized speech can decrease quite quickly even if

there is only one erroneous phoneme per word. Nevertheless, this system gave acceptable performance. Due to the distributed input representation no single unit or link was essential, the system was reported to be rather tolerant to weight damage and it recovered much faster than trained.

Very soon after the publication Sejnowski and Rosenberg's paper an extension of NETtalk was presented by McCulloch et al. (1987). NETSpeak had a few changes in comparison with NETtalk. First of all, the authors claimed that a more concise representation of the input data would help achieving better performance. The number of input units was reduced to 11. The letters were grouped into 5 different mixed phonological sets according to the proximity of their manner of articulation with the exception of vowels which were all placed in one set. The remaining 6 bits were to indicate the position of the letter in the group. The output coding uses less phonological and more stress and punctuation features. The number of hidden units throughout all experiments was equal to 77. Another distinctive feature of this approach is that it was tested on a completely unseen set of words, however the authors used a different lexicon which makes the results difficult to compare. The results obtained on 1,200 unseen words by a network trained on 15,080 words from "Teachers Word Book" were equal to 86% of best guesses. The impact of word frequencies on the results was also studied. The words from the dictionary were replicated in appropriate proportions to make a distinction between common and uncommon words. The authors' hypothesis that the system would perform worse on common words due to their rather irregular G2P correspondences was not proved. A hybrid network that combined two separate networks trained on common an uncommon words was also trained and tested(McCulloch et al., 1987).

2.1.4 Decision trees

CART (Breiman, 1984; Mitchell, 1997), a classical machine learning technique widely used in speech recognition and language modeling is frequently applied to obtain G2P transcriptions. Decision trees represent a symbolic, instance-based, learning approach. They constitute a greedy, top-down classification method, which has great flexibility and generalization capability. The idea of the method is to split the data in such a way that the generalization of training data is maximized allowing a more accurate prediction of out-of-vocabulary words. The basic component includes a set of yes/no questions and a procedure to select the best question at each node to grow the tree from the root. For G2P conversion, the decision tree normally has as the input grapheme sliding window with 3 to 5 letters to the left and to the right in addition to the central letter. The basic yes/no questions are asked about the surrounding context of the letter to be transcribed, e.g. *Is the second left letter 'h'* and *Is the third right letter 'g'*?

Daelemans and Van Den Bosch (1996) propose to build decision trees using the information gain criterion (IG-tree). The starting node represents the target letter and the consecutive nodes refer to consecutive contexts. The order of the questions is defined by calculating the importance of the contexts in the disambiguation process. Right positioned contexts were found to have more importance than left-positioned ones. No pruning is involved and the information gain is computed only once for each attribute. The deeper the path the more ambiguous is the phonemic correspondence for a grapheme. If the tree fails to produce a phoneme, the *best guess* strategy is activated. This strategy finds the most probable phoneme based on occurrences. In an earlier work by the same authors (Van Den Bosch and Daelemans, 1993) k-nearest neighbor metrics was used as a back up strategy to find the phonemic output if the IG-tree failed to do so. The performance was evaluated on a Dutch corpus and compared to a Dutch G2P system analogous to NETtalk. The decision trees gave better accuracy.

In (Andersen et al., 1996) decision trees are grown using the gini splitting criterion that aims at the decrease of impurity measure. Five letter context to left and right is considered, and 10 graphemic classes serve as possible additional characteristics. The questions about the graphemic classes are about whether the grapheme is a vowel or consonant, its manner of articulation or whether it belongs to a diphthong. The leafs of the tree assign the probabilities to possible phonemic outputs. The best results obtained on Carnegie Mellon University (CMU) corpus are 91.1% phonemes and 57.9% words correct. The experiments on NETtalk data are 53.0% word and 89.9% phoneme accuracies correspondingly.

Jiang et al. (1997) build a G2P converter based on Classification and Regression Trees (CART) trees. Both graphemic (5 letters to the left or to the right) and phonemic (3 letters to the left) contexts were involved in the tree creation. In addition to simple questions about the context, complex questions were allowed. The complex questions combined several simple questions in one and allowed a significant reduction of the tree depth. Moreover, they did a good job in capturing transcription for morphemes, e.g. *-tion*. Questions that allowed maximum entropy reduction were prioritized. The system was tested on CMU and NETtalk databases. In order to improve baseline results, questions about categories (vowels, consonants, fricatives, etc.), context distance measures, phonetic 3-grams, phonetic rescoring and multiple tree combination(several trees build from partitioned data) were used to enhance the performance of the system, the best results on NETtalk database were 65.8% words and 91.9% phonemes correct. The results for the CMU data were 73.1% in word and 91.8% in phoneme accuracy. The results reported for NETtalk are among the best results found in literature. In the same year Bakiri and Dietterich (1997) published another work dedicated to G2P conversion using DT. Their tree is build recursively from the root, using the information gain criterion to further expand the tree. It is a binary tree that uses no pruning for overfitting. The input and output formulation was very similar to

that reported in (Sejnowski and Rosenberg, 1987) for training a back-propagation neural network. A subset of 2000 words from NETtalk was chosen and divided in equal proportion into training and test data. The best results obtained in terms of word accuracy were 64.8%, however such a high accuracy can be influenced by the fact that the test data was also used for development and therefore some overfitting may be present.

Black et al. (1998b) also use the information gain criterion. In their work, information gain is recomputed at each node at the time of the split unlike in (Daelemans and Van Den Bosch, 1996) where the information gain is computed only once for all attributes. The resulting structure, called $ID3$, is a decision tree containing questions and return values on terminal nodes. Each leaf of the tree corresponds to a specific pronunciation rather than to a distribution of pronunciations. Pruning is used to prevent overtraining. The experiments for English were carried out on Oxford Advanced Learner's Dictionary (OALD) (Mitten, 1992) and CMU (Weide, 1998) lexica, the best results for OALD were 74.56% words and 95.80% phonemes correct. CMU lexicon is built using sources of variable word difficulty and the results obtained are lower than those for OALD (57.80% words and 91.99% phonemes correct. Stress was predicted separately and together with the phonemes, the latter method seemed more effective.

Pagel et al. (1998) grow decision trees that are very similar to those reported in (Black et al., 1998b), however, they include the following phonemic context in the node-splitting questions. This phonemic context covers up to 3 previously predicted phonemes for the current unknown word. This is similar to hand-written rules defined for other languages. For English, the information about preceding vowels can be crucial to determine if the next vowel is a full vowel or a schwa, also this may be helpful in terms of stress prediction (Pagel et al., 1998). This requires the word to be processed in reverse order from right to left, since the phonemes are considered to be the result of decisions taken previously. They also report improvements achieved by adding questions about the Part of speech (POS) of the word considered. For OALD lexicon, the phonemic contexts allows obtaining 76.66% words and 93.60% phonemes correct. For CMU lexicon the accuracies were 62.79% for words and 87.84% for phonemes. Similarly, the information gain criteria was also used to build G2P decision trees in (Häkkinen et al., 2003). Up to 4 letters before and after the central letter were involved in the questions. The method uses the information about phonemes previously predicted for the out-of-vocabulary word in question. Phonemic classes are used as additional characteristics.

Decision trees are appropriate for discrete characteristics and produce rather compact models, whose size is defined by the total number of questions and leaf nodes in the output tree. The fact that the tree uses only grapheme context on both left and the right sides by DT has a disadvantage: it assumes that the decisions are independent one from another so is that it cannot use the prediction of the previous phone as the reference to predict the

next one. Another limitation introduced by the binary decision trees is that every time a question is asked the training corpus is divided into two parts and further questions are asked only over the remaining parts of the corpus.

2.1.5 PbA

For the first time, Pronunciation by Analogy (PbA) was proposed for reading studies by Glushko (1981), and later, Dedina and Nusbaum (1991) introduced this method to TTS applications. The idea of deriving pronunciations of unknown words from large portions of existing transcription was tested by Glushko (1981) on two sets of mono-syllabic English pseudo-words. The first set contained regular pseudo-words while the other one consisted of those whose pronunciation was considered irregular o an "exception". Several groups of native English speakers were asked to pronounce regular pseudo-words and "pseudo-exceptions" mixed with existing regular words and exeptions, the experiment was carried out and the results for different experimental set ups were analyzed. The results obtained from the experiment refuted the claim that the reading aloud process involved a retrieval of full word pronunciations from memory. Neither they sustained the idea of pronouncing pseudo-words by abstract pronunciation rules. The analogy is the ability to retrieve and use specific multiletter rules to pronounce unfamiliar letter strings. The activation is the implicit use of knowledge of letter-to-sound system, which can be also described as the automatic availability of pronunciations. Analogical pressure from the irregular letter-to-sound patterns resulted in some unexpected pronunciations of pseudo-words. For example, a regular pseudo-word *tave* was often pronounced as [t æ v], because the conflict between the irregular *have* and regular *gave*.

Dedina and Nusbaum (1991) adapted Glushko's hypothesis that pronunciation of novel words can be derived on the basis of lexical knowledge without any mechanism needing rules. Large chunks of existing words were retrieved and recombined into new lexical items. Their analogy-based system called PRONOUNCE (Dedina and Nusbaum, 1991) consists of four major components.

- Aligned lexicon (in a one-to-one manner)
- Word matcher (searches for the existing chunks in the system lexicon)
- Pronunciation lattice (a graph that represents all possible pronunciations recombined from existing chunks)
- Decision maker (chooses the best pronunciation among all present in the lattice)

The only condition imposed by the system in order to recombine two chunks is that two adjacent parts must share exactly one common phoneme. Each possible path through the lattice was considered to be a possible pronunciation, if there was a unique shortest path it was given the preference. Frequency count of chunks was used to break the tie between pronunciations consisting of the same number of chunks. More frequent chunks were considered more reliable. The silence problem was reported when no complete path through the lattice was found. The authors tested their system on a 70 word subset of Glushko's pseudo-words (Glushko, 1981). The test was not representative of a real TTS framework, since those words were too short and did not belong to general English. Moreover, during the further development of the algorithm their results were reported impossible to replicate (Bagshaw, 1998; Damper and Eastmond, 1996, 1997; Yvon, 1996a). The partial pattern matching used to detect existing chunks in the lexicon was found to be inefficient since it did not cover the range of all possible overlaps between the lexicon entries and unknown words.

Yvon (1996b), proposed an extension of the algorithm for the pronunciation module of a French TTS system. The first hypothesis was to find pronunciations of novel words combining heads and tails of words with known pronunciations. The size of the overlap, or common portion in the head and the tail parts, influenced directly on the reliability of the resulting pronunciations. The larger was the overlap the smaller was the chance of obtaining erroneous output. The overlap was required both in the orthographic and phonemic strings and its size was not limited.

In a later work Yvon (1996a), uses the idea of large overlapping chunks was maintained and extended to any number of chunks. The paths in the pronunciation lattice were weighted according to the overlap size. The scoring function searched for a compromise between the number of chunks constituting the pronunciation and the size of the overlap. The information about the chunk frequency was also used to break the ties in the same wasy as in (Dedina and Nusbaum, 1991), In some cases, when no overlapping chunks were found, no pronunciations were produced. The system, SMPA, was evaluated on several public-domain lexica. For the NETtalk lexicon, SMPA showed an improvement of about 7% in comparison with the results obtained using a reimplementation of PRONOUNCE (65.96% word accuracy).

Another reimplementation of PRONOUNCE, called PbA was provided by (Damper and Eastmond, 1996, 1997). The purpose of the authors' work was to improve the scoring function responsible for choosing the best candidate among possible alternatives. The authors also performed a more realistic evaluation, which was also more relevant to TTS systems, on a large corpus of real words, namely the NETtalk lexicon. Two new scoring methods were proposed. The maximum sum heuristics was replaced by the maximum arc frequency product. The total product that summed all the frequency products for the same

pronunciation was introduced as well. The best results obtained were with 60.7% words correct and 91.2% phonemes correct.

Later and most successful implementation of the PbA algorithm was published in (Marchand and Damper, 2000). The main contribution of this work was the introduction of totally novel scoring strategies that allowed making more reliable choices among all available alternative pronunciations.

Brown and Besner (1987) mention that lexical analogy in reading is influenced by such factors as the size of the common part between the unknown word and known lexical entry, its position in the two letter strings, frequency of occurrence in the language, and, in particular, frequency of occurrence of words containing that part. The strategies proposed to enhance the PbA algorithm in (Marchand and Damper, 2000) are:

1. Maximum arc frequency product (PF)
2. Minimum standard deviation of arc lengths ($SDPS$)
3. Highest same pronunciation frequency (FSP)
4. Minimum number of different symbols (NDS)
5. Weakest arc frequency (WL)

These scoring strategies take into account some of the factors mentioned by Brown and Besner (1987). Also, two methods of strategy combination were introduced. Each strategy assigns each candidate a score and based on this score each candidate is assigned a rank. According to the rank, each candidate is awarded points. If a strategy gives the same score for several candidates, they are given the same rank and the same number of points. There are two manners of determining the winner candidate; the sum rule, that chooses the candidate with the largest value of the sum of points for all of the included strategies and the product rule, that chooses the candidate with the largest value of product of the points awarded by each of the included strategies.

Before evaluating the multi-strategy scoring, the authors prepared a baseline system. A single scoring strategy approach and full pattern matching vs. Dedina and Nusbaum's partial matching is used to obtain preliminary results. The silence problem reported in (Dedina and Nusbaum, 1991; Yvon, 1996a), was overcome by adding a null-labeled juncture between adjacent chunks that did not share an overlapping phoneme. The results without silence problem were slightly better. The baseline system, besides G2P evaluation was applied to other problems of similar importance for the field of speech technologies such as phoneme-to-grapheme mapping and grapheme-to-stress conversion. Combined prediction of stress and phonemes gave a rather low word accuracy of 41.8%. The testing strategy

consisted in leaving out each words at a time from the pronunciation lexicon and deriving a pronunciation of an unknown word by analogy with the remaining words. This is called leave-one-out or n-fold cross validation (Daelemans et al., 1997) where n is the size of the lexicon.

In the framework of a multi-strategy approach to pronunciation by analogy all possible strategy combinations were evaluated and the results analyzed. The new strategies allowed obtaining statistically significant improvements in comparison to the results obtained in (Damper and Eastmond, 1996, 1997; Dedina and Nusbaum, 1991; Yvon, 1996a) and to the preliminary results obtained in the same work.

For the NETtalk dictionary (Sejnowski and Rosenberg, 1993), the best accuracy obtained was equal to 65.5% for words and 92.4% for phonemes using all 5 strategies (Marchand and Damper, 2000), which is better than using any one of the strategies alone. The sum and the product rules of strategy combination gave similar results. The best single strategy was the one that prioritized same pronunciations, giving a word accuracy of 63.0%.

2.1.6 Joint multigram models and finite state transducers

Joint multigram approach model is a statistical model that allows to learn variable length grapheme and phoneme from the training corpus and later to decode a string of orthographic symbols into a phonetic sequence.

For the first time many-to-many alignments for G2P were used by Deligne et al. (1995). Joint sequences of graphemes and phonemes of variable length were extracted from the training lexicon using the maximum likelihood criterion. The maximum sizes of corresponding sequences are defined before the training. The algorithm was initialized by computing the relative sequences of all possible many-to-many alignments available from the training lexicon. Then the authors trained 2 different models, one based on EM training and another one based on Viterbi (maximum approximation) training.

The decoding was carried out sequence by sequence and not grapheme by grapheme as in the majority of G2P classifiers. Different sequence sizes and thresholds (setting a minimum the number of times a consequence had to appear in the training corpus in order not to be discarded) were tested.

The evaluation on a French lexicon BDLEX (De Calmès and Pérennou, 1998) containing 23,000 words and compounds showed that 64.52% words and 95.0% phoneme accuracy was achieved by the best model. Thresholding was found very effective in order to improve the performance of the model on unseen words. When phonotactic bi-gram model estimation was used for decoding, smaller values of sequence size parameters were needed to achieve the same results than without it.

Bisani and Ney (2002) applied a similar joint-multigram approach to align joint sequences of graphemes and phonemes. They introduced the term "graphones" to refer to the corresponding graphemic and phonemic chunks of variable length. The pronunciation of the unknown words was inferred using the standard maximum likelihood training (EM algorithm) as well as Viterbi training. The minimum length of graphones was set to 1 and the maximum to 6 for both graphemic and phonemic domains. The best results for Celex lexicon (CELEX) containing 66,278 words were obtained using a 3-gram model. Longer graphones were more difficult to estimate, however the alignments restricted to 1-to-1 graphones seemed to perform worse than when longer chunks were involved. Thresholding and marginal trimming were used to enhance the models. The best results achieved were 95.02% phonemes correct. Galescu and Allen (2001) built a similar $4-gram$ model although they used a different alignment procedure. Each letter-to-phoneme correspondence was restricted to having at least one grapheme and one phoneme, these correspondences were inferred using the EM algorithm. The performance was evaluated on two English lexica, NETtalk and CMU. The experiments included stress prediction, however only for latter lexicon. A back-off n-gram model with Witten-Bell discounting was used to train the model. 1-to-1 manually proofed alignment available for the NETtalk data was also evaluated in the experiments, showing that chunk-based alignments perform slightly better. The results obtained on NETtalk data were 63.93% words and 91.74% phonemes correct. For CMU including the stress markers 62.6% word and 91.0% phoneme accuracies were obtained. When phonemes were predicted disregarding the stress the corresponding accuracies were 71.5% words and 93.6% phonemes correct. Furthermore, the authors also carried out the reverse task of predicting letters from phonemes using the same models. Chen (2003) aligns letters and phonemes using a conditional Maximum Entropy (ME) model with Gaussian priors. Nulls are allowed both in grapheme and phoneme strings and the letters and phonemes are continuously realigned during training unlike other fixed chunk models (Bisani and Ney, 2002; Galescu and Allen, 2001). To train a joint maximum entropy 8-gram both conventional and Viterbi versions of the EM algorithm are used. The results are evaluated on Pronlex lexicon (Kingsbury et al., 1997) for English. The stress markers were not included. Pronlex contains 91,216 words. Syllable boundaries used as an attempt to enhance the model by preventing the syllable splitting, were found rather ineffective. The results were obtained for three datasets: regular words, proper names and foreign words. For regular words the transcription accuracies obtained were 72.7% for words and 92.85% for phonemes. Bisani and Ney (2008) use a similar model as in their previous work (Bisani and Ney, 2002) and test the performance of their system over a variety of English datasets in order to make their results comparable to those reported in literature. Moreover they study different model initialization and training schemes, the influence of the held out set and the effect of different smoothing techniques and the size of graphones on the overall results. The results obtained showed that the joint multigram models proposed performed

better or on par with best performing G2P methods. The results obtained for OALD lexicon were 82.51% words and 96.46% phonemes correct. For NETtalk dictionary (size variable form 15K to 19K) the results ranged between 66.33% to 69.00% for word accuracies and from 91.74% to 92.34% for phoneme accuracy. For CMU dictionary the 75.47% words and 94.22% phonemes correct were obtained. For Pronlex the corresponding accuracies were 72.67% and 93.22% words and phonemes correct. For BEEP dataset(Robinson, 1997) 79.92% words correct and 96.46% phonemes were obtained. Joint models are believed to be beneficial because they handle the alignment problem intrinsically. Caseiro et al. (2002) build a joint model for European Portuguese, however, they only used singular graphone (or 1-to-1 letter-phoneme correspondences). Such a model can be transformed into a finite state transducer, each pair of symbols is converted into a pair of input/output symbols.

2.1.7 Hidden Markov Models

Taylor (2005) proposed to use hidden Markov models to confront the difficult problem of phoneme prediction. HMM is a statistical method that does not require letters and phonemes to be aligned before training. The alignment is generated during the model training stage by Baum-Welch training (Jelinek, 1997) in which the Hidden Markov Models uses the probabilities of the G2P correspondences found in the previous step of the algorithm. Each phoneme is represented by one HMM and letters are the emitted observations. The probability of transitions between models is equal to the probability of the phoneme given the history (previous phoneme). The objective of this method is to find the most probable sequence of hidden models (phonemes) given the observations (letters), using the probability distributions found during the model training.

$$\hat{\varphi} = \underset{\varphi}{\operatorname{argmax}}\{p(\varphi|g)\} = \underset{\varphi}{\operatorname{argmax}}\{p(g|\varphi)\, p(\varphi)\}$$

where $p(\varphi)$ is the probability of phoneme sequence, and $p(g|\varphi)$ is the grapheme-phoneme joint sequence probability. One model is trained for each phoneme; the maximum number of letters that a phoneme is able to generate is set to 4, since it is uncommon that more than four letters represent a single sound, at least in English. No looping states are allowed unlike in the general model configuration that serves for speech recognition. In the phoneme domain, certain constraints and patterns determining the sequences of possible phonemes were imposed. This is similar to phonotactic grammar. Phonotactically illegal sequences could cause a severe problem for TTS because the synthesis system will not be able to generate a corresponding waveform. The automatic speech recognition toolkit can be used to train the HMM models and to decode graphemes into phonemes. However, to achieve better results, some pre-processing was needed. Some letters were swapped and

words rewritten. This measure was necessary because HMMs cannot model dependencies between observations. However, one of the advantages of the HMM is that they allow to model context-sensitivity in the phoneme domain. This was achieved by cloning the context independent models and applying further runs of Baum-Welch for those tokens of the training data that appeared more than 20 times. The experiments were carried out on Unisyn dictionary (Fitt, 2000) of approximately 110K words, most of which are regular English words. There are 42 phonemes in the Unisyn lexicon. The results obtained for a 4-gram model without preprocessing were 39.13% words and 85.12% phonemes correct, preprocessing allowed raising the bar to 49.64% and 87.02% words and phonemes correct correspondingly. Context-sensitive modeling brought the results up to 57.31% words and 90.98% phonemes correct. Stress prediction was included in the experiments. The large portion of errors consisted in schwa-full vowel confusions and stress misplacement.

In (Ogbureke et al., 2010) HMMs are also used for Grapheme-to-phoneme conversion. The authors propose an extention to HMM described above. In previous approaches, only phoneme context, which for first-order HMMs includes only the preceding phoneme, was used. Here, both grapheme and phoneme contexts are modeled. In order to include grapheme context, each observation sequence was transformed increasing at the same timethe number of possible observation symbols. No rewrites were necessary. Stress prediction was not considered. The approach combining context-sensitive grapheme, context-dependent phonemes and a 4-gram language model allowed obtaining 57.85% words correct for CMU dictionary and 79.19% for Unisyn lexicon for British English which is significantly better in comparison to the results obtained in (Taylor, 2005). Increasing the number of observations allows obtaining higher accuracy.

2.1.8 Latent analogy

Common inductive learning techniques used in automatic G2P conversion methods (e.g., decision trees) do not always generalize well, as in the case of proper names of foreign origin for example. Bellegarda (2005) proposed a new method which avoids the traditional top-down in favor of a bottom-up strategy. This approach was designed to exploit all potentially relevant contexts, regardless of how sparsely seen they may have been in the training data. It works by constructing neighborhoods of locally relevant pronunciations, using a latent semantic analysis framework operating on n-letter graphemic forms. The latent semantic analysis determines the most characteristic grapheme strings for all in-vocabulary words. There is one such string for every in-vocabulary word and they are called orthographic anchors. An out-of-vocabulary word is compared to each such anchor and the degree of closeness is determined. All words that score above a preset threshold of closeness are added to the orthographic neighborhood of the out-of-vocabulary word for which the pronunciation is sought. An orthographic neighborhood or a subset of similar lexicon entries

allows to gather the corresponding set of pronunciations or a pronunciation neighborhood. All entries from the latter are aligned and the maximum likelihood is applied to choose the most frequent phoneme for each position. This method was observed to be effective on a large test corpus of proper names (Bellegarda, 2005). Some preliminary experiments were carried out on a training lexicon of 56,514 proper names, mostly of Eastern European origin. The test data was a completely different set of 84,193 proper names of diverse origins. The results obtained on a 500 word test subset from this database were among the best obtained for automatic G2P for proper names. The phoneme accuracy achieved by latent analogy was 86.4% and the word accuracy was as high as 62.0%. Decision trees (Black et al., 1998b) on the same corpus scored 76.7% phonemes and 19.8% words correct.

The summary of the G2P results for English databases found in the literature are given in Table 2.3.

Some recent work on G2P conversion has been published. Jiampojamarn and Kondrak (2009) propose to use a probabilistic approach to choose several best candidate pronunciation generated by analogy. Many-to-many alignments are used and the reported results are quite promising. In (Illina et al., 2011) a probabilistic approach called Conditional Random Fields with HMM-based one-to-one alignment is proposed. Different features such as: POS-tag, context size, unigram versus bigram, etc.; has been studied. The proposed system has been validated on two pronunciation dictionaries and the results compared favorably with the results of the Joint-Multigram approach. Moreover, the method has been proven more suitable for generation of different pronunciation alternatives.

2.2 Pronunciation of proper names

Many TTS systems require an extensive coverage of different proper names such as person's names, place names, addresses, etc. Some speech applications use mostly proper names. Voice controlled GPS navigation systems or directory assistants are only a few examples of such applications. The proper names have been studied by the phoneticians for a long time and it is a known fact that they have much more irregular pronunciations that ordinary words. Pronunciation of proper names is a hard task even for the human brain, however, proper names constitute a great percentage of out-of-vocabulary words and present a serious problems for any kind of dialog-based assistance systems as well as for speech synthesis and recognition systems. Moreover, the error analysis (Tomokiyo, 2000) shows that the foreign words are the major cause of errors in the task of G2P transcription. Data labeling is a difficult task because of the removed accent mark and the fact that some names can belong to more than one language, having a different pronunciation in each one of them. Since the people adapt the pronunciation of their names according to its origin, including this knowledge as an additional feature for classifier should improve the pronunciation accuracy.

Dataset	Author	G2P method	Phoneme acc.,%	Word acc.,%
NETtalk	Sejnowski and Rosenberg (1987)	neural networks	78.0	35.0
=	Torkkola (1993)	DT	-	90.8
=	Yvon (1996a)	PbA	65.96	-
=	Andersen et al. (1996)	DT	53.0	89.9
=	Jiang et al. (1997)	DT	65.8	91.9
=	Bakiri and Dietterich (1997)	DT	64.8	-
=	Galescu and Allen (2001)	joint multigram	63.93	91.74
=	Damper and Eastmond (1997)	PbA	60.7	91.2
=	Marchand and Damper (2000)	PbA	65.5	92.4
NETtalk 15k-19k	Bisani and Ney (2008)	joint multigram	66.33-69.00	91.74-92.34
CMU	Andersen et al. (1996)	DT	57.9	91.1
=	Jiang et al. (1997)	DT	73.1	91.8
=	Black et al. (1998b)	DT	57.80	91.95
=	Pagel et al. (1998)	DT	62.79	87.84
=	Galescu and Allen (2001)	joint multigram	71.5	93.62
=	Bisani and Ney (2008)	joint multigram	75.47	94.22
=	Ogbureke et al. (2010)	HMM		57.85
Unisyn (Unisyn)	Taylor (2005)	HMM	57.31	90.98
=	Ogbureke et al. (2010)	HMM		79.19
Pronlex dictionary (Pronlex)	Chen (2003)	joint multigram	72.70	92.85
=	Bisani and Ney (2008)	joint multigram	72.67	93.22
Celex dictionary (Celex)	Bisani and Ney (2002)	joint multigram	-	95.02
=	Bisani and Ney (2008)	joint multigram	88.68	97.50
OALD				
=	Black et al. (1998b)	DT	74.56	95.80
=	Pagel et al. (1998)	DT	76.66	93.60
=	Bisani and Ney (2008)	joint multigram	82.61	96.46
BEEP dictionary of English (BEEP)	Bisani and Ney (2008)	joint multigram	79.92	96.46
Teacher's word book (TWB)	McCulloch et al. (1987)	neural networks	-	86.0
=	Damper et al. (1998)	Elovitz rules	25.7	-

Table 2.3: Summary of G2P results found in literature for different English datasets (Bisani and Ney, 2008).

Same name can be pronounced in a different manner depending on the language where it comes from, e.g. *David* can be pronounced [d ˈ e v ɪ d] in English manner o [d a β ˈi ð] if the name refers to a Spaniard. Already assimilated pronunciations of common nouns of foreign origin are normally strongly influenced by the letter-to-sound rules of the foreign language, although sometimes the language receptor imposes its own pronunciation due to absence of source (foreign) language phonemes in it. This usually happens to very frequently used words of English origin in languages such as Spanish, Italian and French, which have a tendency to "nativize" foreign words for the sake of linguistic purity, and lack of English knowledge by the population. For example *Microsoft* in Spanish can be pronounced as [m i k r o s ˈ o f], loosing the diphthong and the final /t/, or sometimes the diphthong is conserved resulting in [m a j k r o s ˈ o f]. The stress shift observed in this particular example is not as frequent in Spanish as it is in French, where the stress always falls on the last syllable. However, in Spanish, normally the stress position is reminiscent of word original stress in the source language. The pronunciation of such words often present a cultural issue, especially when it comes to person's names, in this case the grade of assimilation becomes an issue of individual character. Other languages such as English also "nativize" the pronunciation of unfamiliar phonemes by replacing them by the nearest neighbor in the local phonological system (Dutoit, 1997; Liberman, 1979). In general, the globalization phenomenon contributes to better acceptance of the foreign pronunciations by the receptor languages, however this rate is individual for each linguistic society. In G2P conversion the influence of foreign name origin was widely studied. The most complete overview of this approach, is given in (Font Llitjós, 2001).

In (Font Llitjós, 2001) it was suggested that the information about the language origin can significantly reduce the errors obtained during G2P conversion of proper names. The language identifier described in (Font Llitjós, 2001) consists of trained language models for all languages to be classified, however only trigrams are considered since they were found to be sufficient for language identification (Vitale, 1991), the word beginning and end markers are included. For every input word, the trigrams are found and the probability of belonging of each one of the trigrams to every one of the languages in question is estimated. Then the probability of belonging of the word to each language is calculated. The language with the highest probability is chosen to be the name's language of origin. Bi-gram and 4-gram information is also considered important as, for example it can contain valuable information about language-specific 4-letter suffixes. These features are used by implementing special filter (Vitale, 1991) or by using n-grams and a back-off model. The language origin information can be used by the G2P converter directly or indirectly. The direct way of incorporating this information lies in building specific letter-to-sound rules for each language in the database, after having identified the language origin of the input word. Direct usage of the language information is quite risky due to data sparseness problem and possible lack of accuracy of the language identifier. Another disadvantage of this approach

is the necessity to store letter-to-sound rules for all languages considered. The indirect usage as an additional feature of CART classifier (Section 2.1.4) allows more flexibility as it was done in (Font Llitjós, 2001). This information was entered at the word level, and was used in the training only when relevant. Since for a large number of possible language origins, this kind of information could introduce additional confusion for the classifier, language information was replaced by language family information. A native speaker of one language can only identify another language that is close to his own. For the rest of languages he can recognize which language group it belongs to. The language family and language information can be combined. First the language family is identified and then the language is chosen only from the set of language in that language family.

The results obtained in (Font Llitjós, 2001) using direct language origin information show that for those languages that had reliable labels (like English or German) the proper names accuracy pronunciation using language-specific G2P converter is about 7% of word accuracy higher than the baseline. For languages for which the training data was scarce, the results are below the baseline. Using the language origin indirectly as a CART feature, allowed obtaining results higher than baseline in all cases. Exclusion of the names belonging to the languages whose labels were not quite reliable gave about 4% of improvement with respect to the baseline, and when the languages with insufficient training data were also disregarded, resulting in a total of two languages, the results surpassed the baseline by almost 9%. Using language family information as an additional CART feature gave a small improvement of 0.37% in word accuracy compared to the case where the information about all 25 languages was included. Combining the language family information with language models for all languages of that family to obtained the additional CART feature allowed another 0.16% of improvement. Furthermore, the language identification is more accurate than when using 25 language models. In a similar work (Font Llitjos and Black, 2001) the pronunciations generated by the language based and baseline systems were subjectively compared by users. Overall results show that the language-based system was preferred by 17% of the users. These users were exposed to synthesized speech before and had some foreign language knowledge.

Language identification has been also proposed for improvement of proper name pronunciations in (Chen et al., 2006). The authors used n-grams of syllable-based letter clusters to train the language models. The problem of identifying language of proper names in English text is more complicated since all non-English alphabets are normally converted to English. Syllables carry more language origin information than letters and are stable natural units. However, the syllable information is available only from the phoneme string. To obtain the syllable boundaries for the unknown words, first, the lexicon was aligned by Viterbi algorithm in a one-to-one manner. The null phonemes were inserted where necessary. Then, the syllable boundaries were copied directly from the aligned phoneme

string to the letter string. If nulls were found at syllable boundaries, they were moved to the previous syllable. For every language only the core syllables with frequencies above the preset threshold were selected. Subsequently, for n-gram training, words were decomposed into syllables and those in their turn were decomposed into core syllables and letters in the following way: 1) if the syllable contained only one core syllable it was decomposed into that core unit and the surrounding letters; 2) if the syllable contained more than one core syllable it was decomposed into the longest core syllable and surrounding letters; 3) if there were no core syllables the word was split into letters. The likelihood of a word w belonging to language l was calculated using a *tri*-gram. Any new word was segmented in all possible ways using core syllables for each language. The path that gives the highest n-gram score was selected as the final path, and that score gave the final likelihood of w belonging to l. The names appearing in more than one language were removed for the training data to avoid ambiguity. The accuracy of the language identification varied with the number of core syllables above the threshold. The best results were obtained using a list of about 500 to 1,000 core syllables for segmentation. The accuracy of language identification by letter n-gram and core syllable n-gram were compared. For English and German the core syllable gave better results, while for French and Portuguese, the letter n-gram performed better. Despite such a small improvement, the error analysis showed that the error distributions between the two methods and different core syllable sets were quite divergent. Therefore, (Schapire, 1999) algorithm was used for combining n-grams trained on different sets of core syllables. In the experiment with 4 languages, AdaBoost gives about 18% of improvement over language identification accuracy.

In a comparative study the pronunciation accuracy of the proper names for French was assessed by 4 G2P systems (De Mareüil et al., 2005). In French, the pronunciation of foreign names highly depends on the origin and usage. However, there is a conflict between the original spelling and its approximation to the original French pronunciation. The most frequently used proper names tend to be pronounced conserving the similarity to the original pronunciation. Since it is difficult to define what a proper name is, the experiments were restricted to only person's names. A list of name-surname pairs were extracted from the newspaper *Le Monde*. Only the pairs of capitalized words that appeared in the corpus between 100 and 200 times were considered because those names were believed to be of the average difficulty. Then capitalized common words, brands and company names were filtered out. The material was manually transcribed and some pronunciation alternatives were added. For more accurate transcription, the transcribers had access to the context surrounding the selected names-surname pairs (about 100 words to the right and to the left). Their situation was close to that of radio journalist confronted with proper names that he/she has to pronounce. A set of linguistic labels, taking into consideration the naive linguistic knowledge of native French speakers was defined. Spanish and Italian names were placed in a separate group, while Slavic and Germanic were put in a group together.

The language identification was also carried out manually using the context information and Google queries. The authors compare 3 rule-based and a machine-learning system. The overall results show that the data driven approach performs slightly worse than the best rule-based system, both for first names and surnames. The first names were generally transcribed better than surnames. The linguistic labels were not used by the G2P converters but they allowed filtering the results by language origin. French names were transcribed in a more accurate way, while the error rate for English and Germanic names was the highest.

Another issue that arises when dealing with proper names is studied in (Lewis et al., 2004), where the authors described a method for improvement of automatic transcription of transliterated Arabic and Russian words in English text. The proposed system consists of an n-gram base language identifier and a set of language specific G2P rules. Proper names that originate from non-English alphabets present an additional transliteration problem which makes the G2P conversion even less trivial. Applying English G2P rules to Arabic transliteration could result in severe pronunciation errors. The authors used a classifier to identify non-English words in the corpus. Another classifier applied specific G2P rules trained for each language in question. Since the unseen foreign words are generally proper nouns, the "unknown" English words were believed to be morphological variations of lexicon words or misspellings. An n-gram for language identification was trained on the words segmented into individual letters. Word boundaries were marked and each n-gram (4-gram in this case) was assigned a probability based on its frequency. The set of words from the CNN news transcription was compared to the CMU dictionary, and 1001 unknown words were found. Each unknown word was labeled as Arabic, Russian or "other" manually. Prior probabilities were assigned according to the language distribution in the news transcription corpus. The add-one smoothing method was used for previously unseen n-grams. The training corpus consisted of the transcribed English conversations found on the web. The lexicon used for training is the CMU pronunciation dictionary. The transliterated words collected from several resources were hand-transcribed. Two corpora were built, one contained 844 Russian and the other one 582 Arabic words. It is rather small but reliable since it was hand transcribed. The language identifier gave pretty good results, the results ranged from 80 to 90% . In G2P task, the CMU dictionary was taken as the baseline. Both language specific G2P systems (Arabic and Russian) showed a significant improvement over the baseline in this experiment the languages were known beforehand. The overall results of the language identifier and the language specific G2P also beat the baseline by about 8%.

2.3 Multilingual TTS

The globalization phenomenon takes place all over the word, however, it does not influence in the same way on different societies Spain is a country of a remarkable linguistic patrimony, which is a cultural treasure but also represents an additional challenge in terms of speech technologies. The the influence from its own linguistic societies together with the growing international mobility and other effects of the globalization phenomenon make it a very interesting case of study since in the framework of the rapidly expanding field of applications the speech tools must be adapted to the multilingual scope allowing a higher level of flexibility and answering the needs of modern users. Currently in Spain, hearing proper names from all over the world has become commonplace. Text-to-speech synthesis finds many important applications in the emerging market of Spanish speech technologies. Voices that can embrace more than one language are highly demanded in the era of mass media globalization.

2.3.1 Issues with pronunciation of foreign words in a language

Every language receives a constant incoming flow of new words. In addition to the obvious acquisition of neologisms during morphological and semantic word formation, many new words enter the current language from foreign languages (Real Academia Española, 1992). There are several ways that words of foreign origin can become incorporated into a receptor language. On many occasions words are translated through semantic borrowing or calque, e.g., *computer mouse* to *ratón*, or *weekend* to *fin de semana*. Another source of foreign-derived neologisms is lexical borrowing where the lexical form and the semantic meaning are adopted directly from the donor language. This form of borrowing implies adaptation of the pronunciation of the new word to the receptor language and almost always that of the orthographic form as well. For example: *football* to *fútbol*; *whiskey* to *güisqui*; *scanner* to *escáner*. This adaptation has two steps. In the first step, the pronunciation is altered to imitate the pronunciation of the language of origin in regards to the limitations of the phonological system of the receptor language. Then, after the word has been used frequently in everyday life, it loses its original foreign form and its orthography is transformed according to the pronunciation of the receptor language, which is Spanish in our case (see Real Academia Española, 1992). Usually, this involves deletion of the unpronounced consonants, one of the double consonants, the unpronounced final *e*, or other changes. The lexical stress presents an even bigger challenge than the orthographic representation. In English, the stress position is quite irregular. As a matter of fact, many words have primary and secondary stresses that sometimes makes the auditive recognition of where they should be placed an issue, especially for non-native speakers. Due to these factors and also to the receptor language accentuation patterns the stress in the assimilated word does not always

match its original position. For example, the pronunciation of the French word *élite* [e 'l i t] in Spanish varies from ['e l i t e] to [e 'l i t e]. In the first case, the stress is shifted as the consequence of the Spanish interpretation of the French graphic accent (which is used to designate if the vowel is open or closed); the second pronunciation, however, is also accepted by Real Academia Española (1992). Every language has its own accentuation patterns and specific characters that cannot be copied to the new language. This is one of the reasons why it is such a delicate matter to decide the best graphic representation and pronunciation for the new word. In this work, we focus on pronouncing English words in Spanish, before they undergo any graphic assimilation, in the scope of multilingual texts.

2.3.2 Mixed-language texts

Texts written in several languages present a rapidly spreading phenomenon that should not be ignored when talking about high quality speech applications. Worldwide globalization is responsible for an entirely new form of multilingualism in all types of communications resources. Types of mixed-language inclusions vary from word parts to entire sentences. Pfister and Romsdorfer (2003) classify foreign language inclusions into three classes:

1. Words containing a foreign stem but following receptor language morphology
2. Full words following foreign morphology that do not always agree syntactically
3. Syntactically correct sentence chunks

Single foreign words or phrases such as movie titles and proper names already present language identification and pronunciation issues, while more complex language mix-ups that can be found in chats, forums, and other sources make the disambiguation even more problematic. Multilingual texts vary in their nature and their degree of multilingualism depending on the document source. For example, text extracted from a newspaper obeys the strict style determined by the editor, which dictates whether and how the foreign words should be used and when they should be translated to the official language of the issue. Peoples' names and geographical names would be the only inevitable foreign words in this case. Some popular free of charge newspapers, such as international *Metro* and European *20 minutes*, are usually not very restrictive in their use of foreign terms; there are numerous foreign words and phrases in articles on culture and entertainment events. However, texts originating from sources such as blogs, online forums, emails, and short messages reach the highest degree of multilingualism. In such texts, orthographic errors are abundant and unusual abbreviations are frequent, making the search for the correct pronunciation rather challenging.

2.3.3 Information sources

The information about correct foreign word pronunciation can be obtained from different sources. For instance, the *book of styles* used by the television channels and radio stations provide a general idea of how different foreign words should be adapted to the official language (see for instance Llorente and Díaz Salgado, 2004). This book for Spanish is consistent, and although it does not give enough detail on the nativization of foreign words in all cases, it sets the main guidelines to follow. The tendencies for the pronunciation of frequently used words are rather clearly defined, yet the degree of multilingualism for spoken programs is considerably inferior to that of written texts. Usually, the only foreign words that appear during a news flash are the well-known proper names and orthographically assimilated foreign words. Nonetheless, to synthesize high quality intelligible speech from multilingual texts, it is necessary to be able to pronounce any new word that one may encounter. The criteria to be applied for nativization should depend on the frequency of use of the word in the language and the target audience. In the case of Spain, unfortunately, only a small percentage of TV viewers are fluent in English (EF EPI).

2.3.4 Phoneset extension

A phoneset or phoneme inventory is a set of symbols that defines the sounds of a language. Extension of the phoneset phenomenon occurs more often in bilingual communities or in cases when a speaker is at least bilingual; however, it is impossible to study foreign word pronunciation on the level of the individual. In bilingual societies, it is much easier to observe general tendencies. Spain has five officially bilingual autonomic regions: Catalonia, Valencian Community, Balearic islands, Basque country, and Galicia. English phonemes /ʃ/, /z/, /ʒ/, /ʤ/, /ə/, and /ŋ/ (as a phoneme but not as an allophone) are absent from Spanish but exist in Catalan; others exist in Galician. English dental fricative /ð/, for example, finds its analog only in Basque. Better coverage of the English phoneset allows speakers from these autonomous regions to use all the sounds from their phonemic inventory in addition to Spanish sounds, bringing their pronunciations of English words closer to the actual English pronunciations.

In the particular case of Catalan, the phenomenon of nativization of foreign words also takes place; the Catalan phoneset as mentioned previously is much closer to English compared to that for Spanish. Therefore, nativization has to cope mostly with the adaptation of vowel pronunciations. It is curious to note that Spanish words in Catalan are pronounced using the regular Spanish phoneset, due to the fact that the majority of Catalan speakers are perfectly fluent in Spanish. An example of the latter is the pronunciation of Spanish name *Jorge* in Catalan being [ˈx o ɾ x e] and not [ˈʤ o ɾ ʤ ə] as Catalan phonetics would stipulate. The phoneme /x/ is absent from Catalan, but is used for Spanish words.

To maintain an up-to-date synthesizer, we need an ultimate automatic method for the derivation of the nativized pronunciation. The term *nativization* is usually used to designate the pronunciation adaptation process Trancoso et al. (1999).

2.3.5 Previous approaches to nativization

As described in Font Llitjos and Black (2001), knowing the language of origin of foreign words allows a significant improvement in automatic G2P conversion. In the same work dedicated to the pronunciation of proper names the main goal was to find their correct from the viewpoint of American English pronunciation rules, or in other words, to Americanize them.

Another example of a problem similar to nativization is the development of a cross-language synthesizer described in Black and Lenzo (2004). A Basque synthesizer was developed using an existing diphone Spanish synthesizer. The resulting voice was Spanish accented and sounded like one of the many speakers of Basque whose native language is Spanish. The phonemes in the Basque were mapped to the phonemes in the available language (Spanish). Even if the mapping was imperfect, it maintained the vowel-consonant relationships across the languages. This type of mapping can only make sense if there is a significant percentage of phoneme overlap between the source and the target language. Spanish and Chinese, for example, do not share enough phonemes for this type of mapping. In Pfister and Romsdorfer (2003), the language identification and foreign words pronunciation issue was approached jointly with a text analyzer. The text analysis was decomposed into two steps: a set of monolingual text analyzers was elaborated with their own lexica and grammar; and then for each pair of languages $\{L_i, L_j\}$ an inclusion grammar defined which elements of the language L_j were allowed as foreign inclusions in the language L_i. This work solved the problem regarding the use of German in Switzerland where there is a tendency to pronounce foreign words or even word parts according to the source language phonetic rules. Moreover, the text analyzer provided precise word and morpheme language identification for this narrow problem. The pronunciation of foreign proper names has also been addressed for the case of English and German names in Swedish (Lindström, 2004), but once again, the authors determined that Swedish speakers extended both their phoneme inventories and their phonotactics when pronouncing foreign names. Of course, the intelligibility of such names does not only depend on the speakers but also on the listeners and their linguistic and cultural backgrounds. On the other hand, for both English and European Spanish languages there is a clear tendency to adapt foreign proper name pronunciation to the phonetic rules of the receptor language. Indeed, in the two languages, and especially in Spanish, due to the smaller coverage of its phoneset, it would sound very unnatural to have foreign inclusions pronounced according to foreign pronunciation rules within utterances in either of these languages. The nativization issue was mentioned

and the factors influencing nativized pronunciations were analyzed in the framework of the Onomastica project dedicated to the creation of a multilingual lexicon (Trancoso, 1995). Later, a rule-based approach was applied to the derivation of alternative pronunciations with different degrees of nativization; both full and null knowledge of foreign language were considered for this purpose. These alternatives were used in voice-controlled navigation system queries for German and French (Trancoso et al., 1999). In French, as in Spanish, foreign words and proper names are nativized to French pronunciation and the phonemes are restricted to the French inventory. However, the lexical accent is placed on the last syllable in 99% of the cases as French pronunciation rules would suggest. The nativization phenomenon is very common for monolingual regions of European countries. In bilingual regions as well as in countries with significant English-speaking influence such as Sweden or Switzerland, tendencies for phoneset extension and closer proximity to foreign pronunciations can be observed (Lindström, 2004; Pfister and Romsdorfer, 2003; Trancoso et al., 1999).

Nevertheless, the problem of foreign word nativization is relatively new to speech synthesis researchers. The multilingualism problem in general has been given much more attention in the framework of automatic speech recognition (Trancoso et al., 1999; Van den Heuvel et al., 2009), where non-native and dialect variations are reported to be the cause of a great number of recognition errors. In synthesis, when dealing with the problem of non-standard pronunciations, we can divide it into two components: foreign pronunciation of native words or non-native speech, and native pronunciation of foreign words. The first component of the problem is highly variable, given the large number of different accents and corresponding phonesets. Moreover, foreign pronunciations hardly obey any regular pattern because they comprise phoneme pronunciation, intonation, and semantic and word-morphing errors, whereas, native or *nativized* pronunciation of foreign words follows traceable patterns. Prosody, intonation, speech rate, and other components, are defined by the phonetic rules of the target language, and only the pronunciation is influenced by the source language. Several social linguistic conventions based on the frequency of use of a particular word and its degree of phonetic assimilation in the target language help to define pronunciation adaptation criteria.

2.4 Other factors influencing pronunciation accuracy

2.4.1 Compatibility and consistency of the lexica

Different syllabification and alignment methods can add inconsistency to the lexica used for training automatic G2P systems; they also render automatic evaluation of these methods even more problematic. The lexica created by experienced linguists are usually more reliable than those automatically generated from parallel text-voice databases. Automatically

created lexica usually contain many identical word entries with different pronunciations, some of which are frequently inconsistent. Besides, any alternative pronunciations that do not depend on grammatical information, such as part-of-speech tags, introduce unwanted ambiguity to the automatic pronunciation generation problem. However, they can be quite useful for more accurate database segmentation. Furthermore, expert-proofed pronunciation dictionaries are not always compatible or even comparable. Phonetic alphabets and transcription criteria used are the two main points of dictionary incompatibilities.

2.4.2 Evaluation standards for G2P techniques

The lack of evaluation standarts can lead to obtaining different results by the same tecniques, examples of this can be found in literature. There is a lot of work done on the G2P conversion for English and for other languages (Black et al., 1998b; Damper and Eastmond, 1996; Galescu and Allen, 2001; Sejnowski and Rosenberg, 1987; Yvon, 1996a). But usually the results are presented in such a way that the performance is difficult to compare. Every author uses a different lexica, different amount of data in the training and test sets, different output features or data alignment methods. A variety of systems predicts different word and phoneme characteristics as, for example, primary stress, secondary stress, or syllable boundaries. Given that the error rate strictly depends on every one of these factors it would be very helpful to set up a unique evaluation framework and compare different G2P conversion methods in a more precise way. It is very important not only to evaluate the phoneme error rate, but also word error rate because it is a more exact measure that allows having a better estimate of the intelligibility expected for the speech synthesized by the method in question. A higher word error rate brings downthe speech quality and perceptual acceptability considerably. Strongest and weakest points of a performance of a G2P system can be exposed through performing the evaluation separately on proper names, geographical names, and common nouns and varying the size of the training data. In the 1^{st} and the 2^{nd} evaluation runs in the framework of the TC-STAR project `www.tc-star.org` the accuracy was evaluated separately for proper names and common nouns. To compare various G2P techniques, the systems should be tested on the same dictionary and a standardized output, including the same phonetic alphabet, and same evaluation techniques should be used. A comparison of most representative G2P techniques for English on the same data was performed by (Damper et al., 1998). The Evasy evaluation allowed comparing the G2P systems for French (De Mareüil et al., 2005). (Damper et al., 1998) proposed testing competitor G2P techniques on the same large datasets and employing strict scoring techniques without taking into account frequency-weighted characteristics. Last but not least they proposed to use the same phoneme set for the output.

2.5 Conclusions

This chapter summarizes the state of the art methods for automatic grapheme-to-phoneme conversion discussing the results achieved and the databases used along the past decade. Several partially unresolved issues needing improvement are pointed out. The alignment used for training the system plays an important role on the results, however no gold standard for alignment exists. If a lexicon with reliable pronunciations is available, the resulting pronunciations will less likely be erroneous. It is inappropriate to compare different conversion methods using different databases. The size, reliability and the coverage of the training data influence the results. Performance of methods tested on data gathered from different sources can neither be compared objectively nor subjectively. Grapheme-to-phoneme conversion is a rather trivial task for languages with a transparent correspondence between letters and phonemes and, on the contrary, it is a quite challenging task for languages with ambiguous letter to phoneme correspondences and abundance of irregular pronunciation patterns. Therefore, pronunciation for languages with shallow orthography can be derived rather successfully using decision trees or rewrite rules. Proper names pronunciation does not follow the standard pronunciation patterns and its derivation can be quite ambiguous for any language. Among the top of the line data-driven G2P conversion methods compared using different lexica of English, the best performance was obtained by joint multigram models (Bisani and Ney, 2008) and pronunciation by analogy (Marchand and Damper, 2000).

In the next chapter we compare different G2P state of the art conversion methods using different lexica for different languages and we propose ways to improve their performance. The pronunciation variation in connected speech is approached using weak forms and phonotactic rules. An error analysis is also provided.

Although the research carried out in this thesis is intended for the multilingual scope, a great part of it is dedicated to English since it is a crucially important language that presents several important issues when it comes to pronunciation derivation. The main difficulty lies is the pronunciation of vowels as there are only 5 vowel letters in English and at least 11 pure vowel sounds(this number varies across different dialects) disregarding allophones. Some vowel or consonant elisions, assimilations, etc, also take place therefore adding even more complexity to the pronunciation system.

Chapter 3

Data-driven approaches to G2P conversion

In this chapter, different G2P techniques are compared experimentally and some improvements are proposed. Errors obtained by the best performing method are thoroughly analyzed and some conclusions drawn. G2P conversion results are obtained for English and several other languages.

Experiments performed for isolated words with and without stress and connected speech are described. The pronunciation was predicted for different stressed and unstressed lexica using different data-driven methods and the results were compared. Different parameters influencing the error rate in grapheme-to-phoneme conversion were analyzed for a number of data-driven classifiers. Word and phoneme error rates were obtained, moreover, error rate as a function of the word length was studied. Some of the improvements were integrated into Ogmios TTS system, which was build in the framework of TC-STAR project. Sections 3.1 through 3.4 give the description of these methods, some improvements are already proposed in 3.4, and further improvements can be found in 3.5. Section 3.6 introduces G2P conversion results obtained for other languages, while the error analysis follows up in Section 3.8. And, finally, Section 3.7 describes the particularities that should be taken into account when dealing with connected speech and database segmentation for TTS.

3.1 Decision trees

Deriving the pronunciation automatically by using decision trees, such as Classification and Regression Trees, or CART, is a commonly used technique in grapheme-to-phoneme conversion. Pagel et al. (1998) introduced CART decision-trees into pronunciation prediction. A graphemic sliding window, containing three letters on the left and three

letters on the right of each center letter was used to build the input vectors. The advantage of using CART method is that it produces compact models. The model size is defined by the total number of questions and leaf nodes of the generated tree. Tree leaves correspond to the best prediction for a sequence of questions. At each step of the algorithm, all possible questions about all possible attributes at that step are asked. The entropy is measured after each partitioning of the data and the question that generates the least entropy is selected. The entropy increases as a function of the distance between the samples in a set. Each attribute is questioned individually and in this way only the relevant attribute can influence the outcome, however, a clear disadvantage of binary decision trees is that the learning data is split in two at each node. For example, if the central letter is *c*, one of the questions to generate a simple general rule could "Is the first letter on the right an *o*?" and so on. For building decision trees, the software include in Wagon as part of Festival TTS (Black et al., 1998a) software kindly provided by the University of Edinburgh was used. The decision tree architecture is represented schematically in Figure 3.1

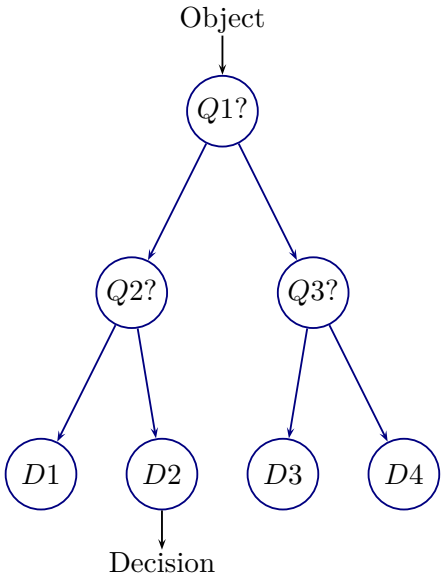


Figure 3.1: A binary decision tree architecture.

Experimental results

For the experiments described in Chapter 3 only common words from the LC-STAR dictionary were used (about 50K). No homonyms were considered for the experiments. The LC-STAR dictionary for American English is transcribed in General American dialect. The Unisyn lexicon, taken from a publicly available source (Fitt, 2000), comprises a rich

variety of different pronunciations of English including standard American English; it has about 110,000 word entries. The partition of the lexicons into training and test sets was carried out leaving 90% of each lexicon for training and 10% for testing on a random basis. In case of DT, in order to optimize different parameters adjustable for these methods, 10% of the training data was reserved for development.

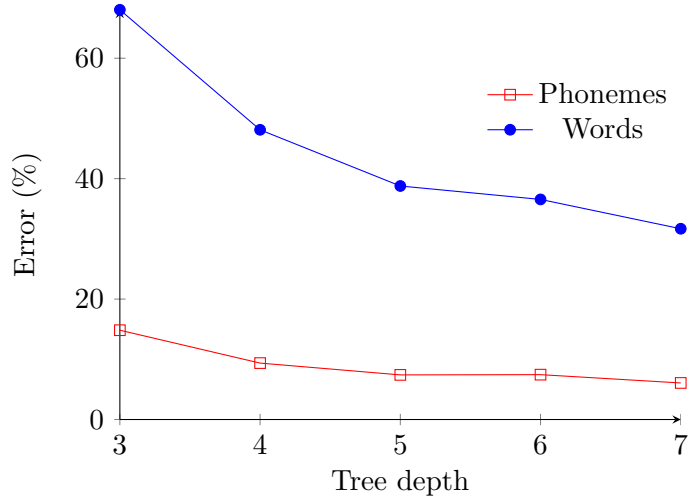
The system was trained and tested using the CART tree-based phoneme prediction. As this algorithm requires the training data to be pre-aligned, the alignment was carried out using the epsilon scattering method (see 2.1.2). Different CART parameters such as maximum tree depth and entropy gain were tested on the development set and the best parameters were applied to obtain the prediction on the test data set. Phoneme and word error rates for different decision tree parameters are plotted in Figure 3.2.

Figure 3.2a shows the error for different values of maximum tree depth. If we set this parameter to be very small, the word error rate tends to be very high. Figure 3.2b shows this error rate as a function of the value of minimum amount of the entropy gain, necessary to justify further tree expansion. The parameters that give us the best results were found to be 0.001 for the entropy gain and 7 for the tree depth. These values coincide with those obtained by (Black et al., 1998b). The results obtained for the lexicon excluding stressed marks, setting the tree parameters to the ones that showed best performance on the development set, are given in Table 3.1. Long non-standard words and abbreviations were removed from the corpus. The first row gives phoneme and word accuracies without taking into consideration the stress marks. The second row of the same table shows the results obtained for combined prediction of stress and phonemes. The lexical stress brings an additional difficulty to G2P conversion task, that is why the accuracies given for the stressed lexicon are significantly lower. During the evaluation of the pronunciations predicted with stress marks, a misplaced stress mark was counted as an error. However, in speech synthesis stress prediction is rather important, while for speech recognition it may be sufficient to predict only phonetic transcription.

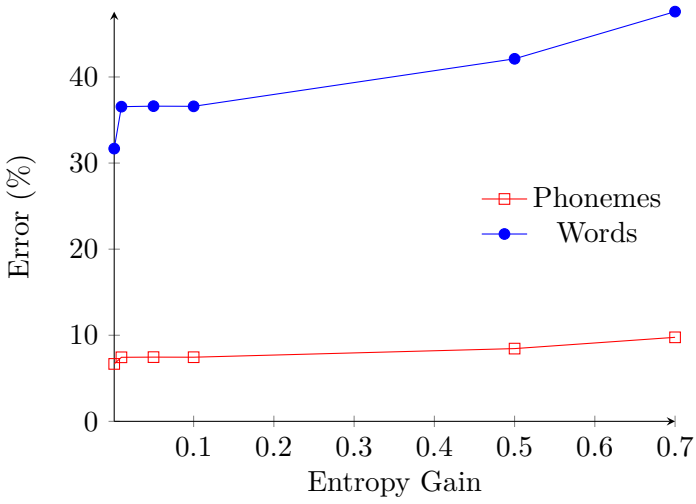
	Phon.accuracy	Word accuracy
w/o stress	93.93	68.32
with stress	91.29	57.8

Table 3.1: Percentage of correct phonemes, and words for stressed and unstressed lexicon using CART.

As it can be seen from Table 3.1 the system performs significantly better on the unstressed lexicon, especially if we take into the consideration the word accuracy results obtained.



(a) Phon. and word error rates (%) as a function of the max. tree depth.



(b) Phon. and word error rates (%) as a function of min. entropy gain.

Figure 3.2: Ajusting parameters of the decision tree.

3.2 Finite-state transducers

In this next approach, the pronunciation is inferred by using Finite State Transducers (FST). Letters are used at the input and phonemes at the output. For this task a stochastic FST was used.

Finding the pronunciation of an unknown words is the same as to chose the pronunciation that maximizes

$$\underset{\varphi}{\operatorname{argmax}}\{p(\varphi|g)\} \tag{3.1}$$

where $\varphi = \varphi_1 \dots \varphi_N$ is the sequence of phonemes, including the "empty" phoneme, and $g = g_1 \dots g_N$ is the letter sequence.

To solve the maximization problem we use a finite state transducer, which is similar to the one described in Galescu and Allen (2001). Equation 3.1 can be expressed as

$$\underset{\varphi}{\operatorname{argmax}}\, p(\varphi/g) = \underset{\varphi}{\operatorname{argmax}}\{p(\varphi, g)\} \tag{3.2}$$

This can be estimated, using standard n-gram methods. As for decision trees, letter-to-phoneme alignment was inferred before the training phase. Then, singular graphones or letter-phoneme pairs found in the aligned dictionary $\gamma_k = (g_k, \varphi_k)$ were defined. Singular graphones only allowed 1-to-1 letter-to-phoneme correspondence.

Applying Bayes rule to Equation 3.2 we obtain:

$$p(g, \varphi) = \prod_{i=1}^{N} p(g_i, \varphi_i / g_1^{i-1}, \varphi_1^{i-1}) \equiv \prod_{i=1}^{N} p(q_i / q_1^{i-1}) \tag{3.3}$$

where N is the number of letters in the word. This probability can be estimated using graphone n-grams.

n-grams can be represented using a finite-state automaton in which a transition (an edge) with label (g_k, φ_k) is added for each new graphone and a state γ_k (g_k, φ_k) is created for each history element h. The weight of each edge corresponds to the probability of the transition, $p(\gamma_k/h)$

For example, for the word *ALIGNED* we create the following singular graphone sequence:

$$< start > (A, [\text{a}])(L, [\text{l}])(I, [\text{aɪ}])(G, [\_])(N, [\text{n}])(E, [\_])(D, [\text{d}]) < end >$$

Figure 3.3 shows some states and edges for the bigram-based FSA. All training data is used to estimate the Finite State Automata (FSA), including the probabilities.

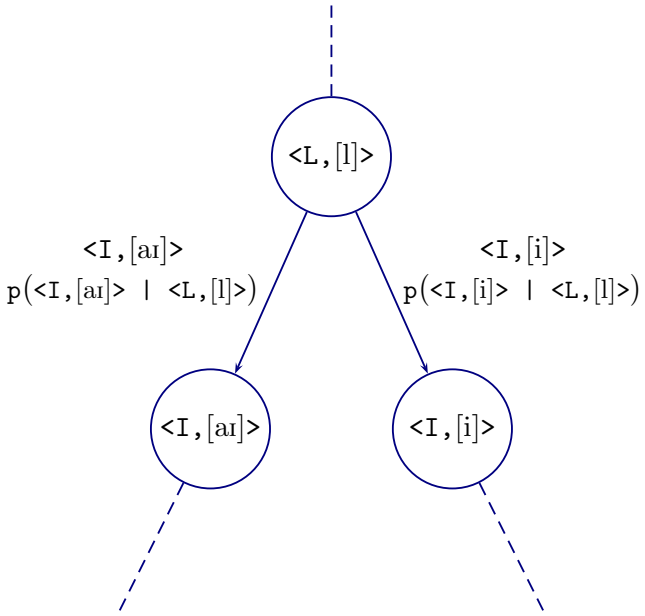


Figure 3.3: States of a finite-state automaton used to represent graphone probabilities.

The FSA allows to compute the probability of the graphone sequence $p(\gamma)$. In order to solve equation (3.2), we derive a finite state transducer in a straightforward way: the labels attached to the FSA are split: the letters of each letter-phoneme pair become input and the phonemes become output elements. Figure 3.4 shows the finite-state transducer derived for the finite-state automaton in Figure 3.3.

Note that the FST is non-deterministic. For instance, from the state labeled as $< L, [l] >$ there are two possible edges for the same input letter I.

To find the pronunciation we have to solve equation (3.3). This is equivalent to finding the path through FST maximizing the transition probabilities. The input letters limit the number of edges which can be followed. The best path is found using dynamic programming.

In this thesis, in order to estimate n-gram probabilities an x-gram model is used (Bonafonte and Marino, 1996). The x-gram model assumes that the number of conditioning grapheme-phoneme pairs (history length) depend on each particular case. The main idea of the x-gram model lies in applying a state-merging algorithm to reduce the number of states. Two criteria for state merging were used. First of all, merging takes place if the number of times a given history $< q_{i-m}, \ldots, q_i >$, where q_i is a grapheme-phoneme pair, has appeared in the training data is below the threshold, k_{min}. The probability of such rare graphones is calculated using the distribution associated to the more frequent state $< q_{i-m+1}, \ldots, q_i >$.

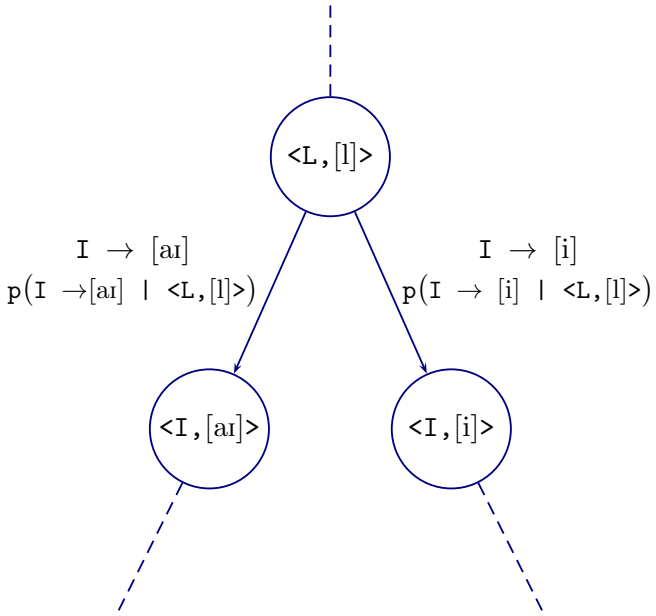


Figure 3.4: States of a finite-state transducer used to transuce letters to phonemes.

The states are also merged if their distributions $p(q_{i+1}/q_{i-m},\ldots,q_i)$ and $\hat{p}(q_{i+1}/q_{i-m+1},\ldots,q_i)$ differ less than a threshold D_{th}. This difference is measured as the divergence D defined as

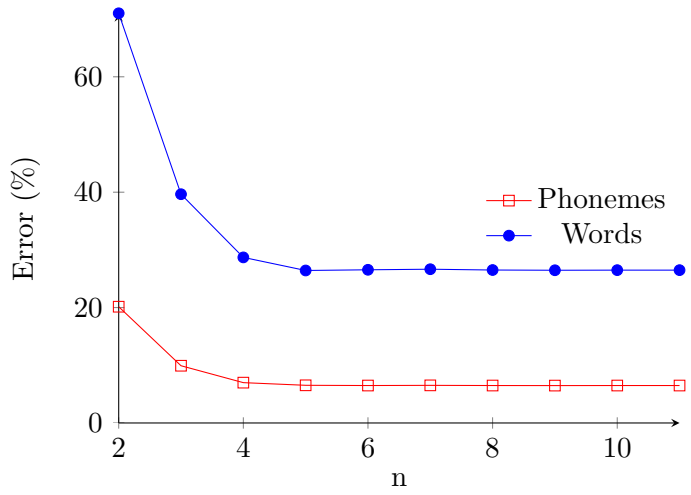
$$D(p||\hat{p}) = \sum_{j=1}^{J} p(i)\log\frac{p(j)}{\hat{p}(j)} \tag{3.4}$$

where J is the number of different graphones in the lexicon.

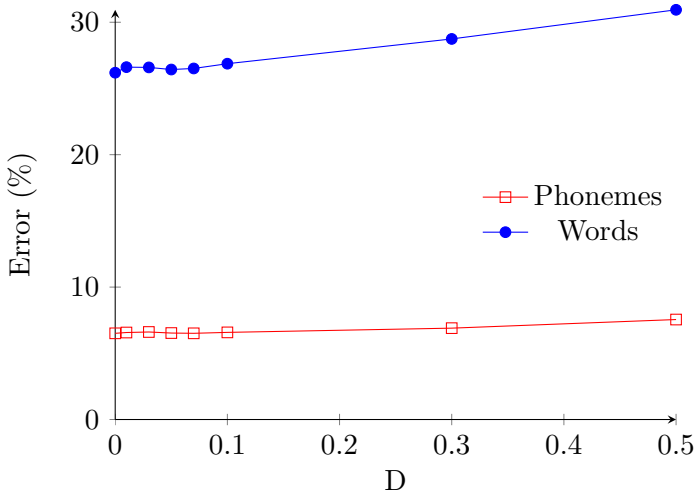
Choosing the proper values of k_{min} and D_{th}, one can significantly reduce the number of states without decreasing model performance. In this work, the threshold k_{min} was set to 1.

3.2.1 Experimental results

For FST it was important to analyze how the parameters of the x-gram models could affect the error rates. In order to do this, experiments for different values of the x-gram model parameters n and D_{th} were carried out, being n the maximum possible length conditional probability history. Figure 3.5a shows error rate as a function of n.



(a) Phon. and word error rates (%) as a function of n.



(b) Phon. and word error rates (%) as a function of the divergence threshold D_{th}.

Figure 3.5: Adjusting parameters of the x-gram.

Maximum history length $n = 5$ was found to be the optimal parameter of the n-gram model for the task in question. In Figure 3.5b error rate is plotted as a function of divergence threshold of the x-gram model.

The results plotted in Figure 3.5 show a significant reduction in number of states of the x-gram model gives good results even for rather high values of the divergence threshold. For $n >= 5$ error rates practically do not change if D_{th} falls within the range $0 < D_{th} < 0.1$. Thus, for the current task the error rate does not seem to decrease any further by increasing n or decreasing D_{th}. The best parameters for this lexicon were chosen to be $n = 5$ and $D = 0.05$. The results obtained by FST are given in Table 3.2. The accuracies obtained for the unstressed lexicon are significantly higher than those obtained for the stressed lexicon, once again proving the difficulty of stress prediction problem for English. FST scores almost 8 percentage points lower in word accuracy if the stress is predicted simultaneously with the pronunciation.

	Phon.	Word
w/o stress	93.63	75.66
with stress	91.07	67.91

Table 3.2: Word and phoneme accuracies for stressed and unstressed lexicon using FST.

3.3 Hidden Markov Models

As it was already explained in 2.1.7 HMMs can be applied to model the pronunciation in the G2P task. The phonemes in this case are represented by HMMs while graphemes are the observations. Each HMM is described by a number of parameters such as number of states and the initialization methods of initial state probabilities and the probabilities of transition between states. Here we use the same parameters as those proposed in Taylor (2005). No previous alignment was necessary since Baum-Welch algorithm as it usually happens in continuous speech recognition, took care of it during the HMM training. A model was created for each phoneme in the phoneset of the corresponding lexicon, and the initialization was carried out assigning uniform observation and transition probabilities to all models. First-order HMM were used. The maximum number of states for each model was set to 4 as it is quite probable that one phoneme could correspond to more than one letter in English. The topology is showed in Figure 3.6. The transitions from any state to the final state were allowed without looping. If, in the opposite case, one letter produced more than one sound, double phonemes were used. The procedure to decode the phoneme sequences from the grapheme sequences is the same that the one used in continuous speech recognition to find words given the spectral sequence. Here, the language model estimates

the probability of a phoneme given previous phonemes. *4-gram* models estimated from the phonetic transcriptions available from the training lexicon were used. The decoding was carried out by Viterbi algorithm. The combination of the letter probabilities associated to each phoneme (HMM) and probabilities of each phoneme sequence (phoneme *n-gram*) allow disambiguation between several possible outcomes.

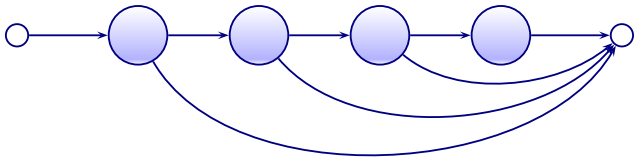


Figure 3.6: Topology of a HMM in G2P conversion.

The performance of the HMM was tested on both LC-STAR and Unisyn lexica for American English. Word and phoneme accuracies obtained for LC-STAR_TEST_SET equaled 47.54% and 84.16% correspondingly. For UNI_TEST_SET the word accuracy of 54.87% is comparable to the results reported in Taylor (2005) for the same lexicon, however,in the cited article a British version of the lexicon was used. The phoneme accuracy obtained for UNI_TEST_SET was 86.16%.

3.4 Pronunciation by analogy

Finding the pronunciation for English words given only their orthographic form is a challenging task for non-native speakers and it is even more so for automatic systems that are usually based on statistics.

The human brain handles statistics in a different way; humans use analogy to memorize how to pronounce words or word fragments in English and other languages with deep orthography. When trying to read something, it takes time and extra effort to apply the pronunciation rules of the language, while the analogy matching that our brain performs in thunder fast. Whether we say it correctly or not depends on the number of words with similar pronunciation rules that we have already learned before. This is where the computer has a great advantage in comparison to English learners. A computer is able to find similarity patterns in throughout all examples from the dictionary and apply statistics-based analogy to derive pronunciations for the new words in question of milliseconds. Pronunciation by analogy PbA is an interesting technique similar to language learning that was successfully applied to deriving pronunciation of out-of-vocabulary words (Dedina and Nusbaum, 1991; Marchand and Damper, 2000; Yvon, 1996a).

PbA system reported in Marchand and Damper (2000) was compared to other G2P classifiers and possibilities of further improvement of the system's performance were explored from different statistical and linguistic perspectives, new scoring strategies were proposed.

3.4.1 Algorithm description

For the first time, PbA was proposed for reading studies by Glushko (1981), and later, Dedina and Nusbaum (1991) introduced this method to TTS applications. The latest and most successful implementation of the algorithm was published by Marchand and Damper (2000), which we have reimplemented for our experiments. This system as well as the initial one, called PRONOUNCE (Dedina and Nusbaum, 1991) consists of four major components.

- Aligned lexicon (in one-to-one manner)
- Word matcher
- Pronunciation lattice (a graph that represents all possible pronunciations)
- Decision maker (chooses the best candidate among all present in the lattice)

Below we review the entire algorithm because it is necessary in order to understand the new strategies and introduce new terminology.

Alignment

The alignment required by data-driven approaches to G2P conversion is explained in detail in 2.1.2.

Pronunciation by analogy algorithm also requires an alignment, in this particular case a one-to-one match between orthographic and phonetic strings is necessary. In other words, each letter has to be aligned to a corresponding phonetic representation. For words that contain more letters than sounds, a null phone /_/ is inserted into the phoneme string, e.g., *thing* /θ _ i ŋ _ /. Otherwise, if the number of phonemes is greater than the number of letters, the phonemes corresponding to the same letter are joined together in one, e.g., *fox* /f ɑ k_s/. The alignment used to carry out the experiments is based on the EM algorithm, and it is similar to that described in Damper et al. (2004). However, the alignment is not always perfect and it can influence negatively on the results. Performance of analogy-based methods is especially vulnerable to the alignment imperfections.

Description of the algorithm

After the training dictionary is aligned, the matcher starts to search for common substrings between the input word and the dictionary entries. Every input word is then compared to all the words in the lexicon to find common "arcs". We called the substrings in the grapheme context *letter arcs* and the corresponding substrings in the phoneme context *phoneme arcs*. All possible letter arcs with a minimum length of two letters and a maximum length equal to the input word length are generated and then searched in the dictionary. For every letter arc from the input word, that matched the same letter arc in a dictionary word, the corresponding pronunciation or the phoneme arc is extracted. The frequency of appearance of each phoneme arc corresponding to the same letter arc is stored along with the starting position for each arc and its length. As an example, assume that the word `#top#` is absent from training lexicon; the list of all possible searchable letter arcs for this word can be given as "`#t, #to, #top, to, top, top#, op, op#, p#`". Now, suppose that in the lexicon, we have the word "`#topping#`" with the pronunciation /# t ɑ p ˍ ɪ ˍ ŋ #/. Here the matcher finds the letter arcs `#t, #to, #top, to, op` and `top` with their corresponding phoneme arcs /# t/, /# t ɑ/, /# t ɑ p/, /t ɑ/, /ɑ p/ and /t ɑ p/. Let us assume that the next word in the lexicon is `#cop#`. It gives us three more letter arcs matching with the word `#top#`, which are `op, op#`, and `p#` with their corresponding phoneme arcs /ɑ p/, /ɑ p #/ and /p #/. Each time the same phoneme arc is found for the same letter arc, the frequency of the phoneme arc in question is incremented. After the word *cop* is processed the frequency count for the phoneme arc /ɑ p/ becomes equal to 2, see Figure 3.7. The matching phoneme arcs are introduced into the pronunciation lattice that can be represented by nodes and connecting arcs. If an arc starts at a position i and ends at a position j and if there is yet no arc starting or ending at position i, the nodes N_i and N_j are added to the graph and an arc is drawn between them. All nodes are labeled with the corresponding "juncture" phoneme and its position in the word. The arcs are labeled with the remaining phonemes and their frequencies of appearance. An example of lattice construction for the word *top* using the arcs found in the words *topping* and *cop* is illustrated in Figure 3.7. These arcs and their frequency counts are updated when the search continues through all the words of the dictionary. After the pronunciation lattice is completed the decision maker chooses the best pronunciation. Each complete path through the lattice is called "pronunciation candidate". Throughout this manuscript, we considered only the shortest paths through the lattice, i.e. candidates consisting of minumum number of arcs (Marchand and Damper, 2000). If there is a unique shortest path through the lattice, it is automatically chosen as the best pronunciation and the algorithm stops. Usually, there are several shortest paths through the lattice, and a decision function is necessary to choose the best pronunciation candidate among them, e.g. in Figure 3.7, there are two shortest paths with the same pronunciation /ˈ t ɑ pˈ/. Please note that no single letter matches were considered. To solve the silence

problem, when no complete path through the lattice was found, concatenation of phoneme arcs was allowed, i.e. in case there was no complete path found through the lattice no phoneme overlap was required in order to concatenate two ajacent arcs. In our example, if we only had two arcs available e.g. /# t ɑ/ and /p #/ we would still build a path through the lattice by concatenating these two neighboring but not overlapping phoneme arcs (the overlapping arcs share the same phoneme).

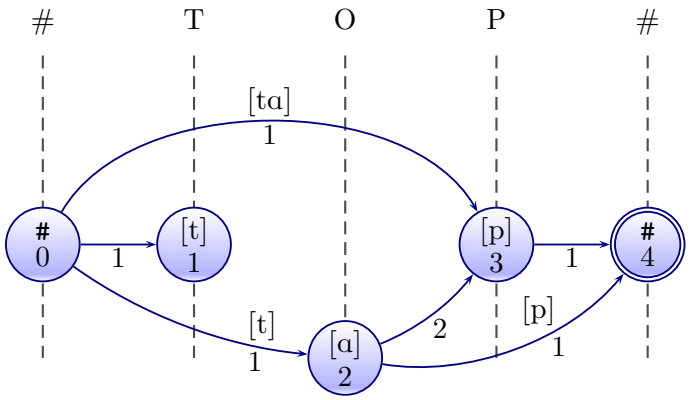


Figure 3.7: Pronunciation lattice for the word *top* using the arcs extracted from the words *topping* and *cop*.

Each candidate can be represented as $C_j = \{F_j, D_j, P_j\}$, where $F_j = \{f_1, \ldots, f_n\}$ are the phoneme arc frequencies along the j^{th} path, $D_j = \{d_1, \ldots, d_n\}$ are the arc lengths and $P_j = \{p_1, \ldots, p_k\}$ are the phonemes comprising the pronunciation candidate, with k being the pronunciation length. Marchand and Damper (2000) proposed to use five scoring strategies in order to choose the best pronunciation. Also, two methods of strategy combination were introduced. Each strategy gives us a score for each candidate and based on this score each candidate is assigned a rank. According to the rank, each candidate is awarded points. If a strategy gives the same score for several candidates, they are given the same rank and the same number of points. There are two manners of determining the winner candidate; the first one is the sum rule, which chooses the candidate that has the largest value of the sum of points for all of the included strategies. The product rule chooses the candidate with the largest value of product of the points awarded by each of the included strategies. For the NETtalk dictionary (Sejnowski and Rosenberg, 1993), the best accuracy obtained was equal to 65.5% for words and 92.4% for phonemes using all 5 strategies (Marchand and Damper, 2000), which is better than using any one of the mentioned strategies alone. The sum and the product rules of strategy combination gave similar results. The scoring strategies are based on the following parameters: frequency of appearance of a

given phoneme arc in the dictionary; its length; and the actual phonemes that constitute the candidate. Different strategies work with different aspects of analogy. High arc frequency is considered to be a major advantage over low arc frequency. Frequencies of suffixes and prefixes are prioritized by different strategies. In other strategies the final score for the candidate is directly proportional to the number of phonemes it has in common with the others. If two candidates share the same pronunciation, both of them are prioritized. These measures are used separately or combined across the strategies.

The strategies are described in detail below. The *original* 5 strategies were proposed by Marchand and Damper (2000). Later in Polyákova and Bonafonte (2008a, 2009) 6 new strategies were proposed with the goal to further improve the pronunciation by analogy algorithm.

Original strategies :

1. Maximum arc frequency product (PF) For each arc the corresponding arc frequencies are multiplied

   $PF(C_j) = \prod_{i=1}^{n} f_i,$

   n is the candidate length, or the number of arcs of which the candidate consists. Rank 1 is given to the candidates scoring the maximum PF.

2. Minimum standard deviation of arc lengths ($SDPS$)

   $SDPS(C_j) = \sqrt{\sum_{i=1}^{n}(d_i - \bar{d})^2/n},$

   where $\bar{d}$ is the mean arc length. Rank 1 is given to the candidate scoring the minimum $SDPS$.

3. Highest same pronunciation frequency (FSP) The privilege is given to the candidates that share the same pronunciation with the others

   $P(C_j) = cand\{P_j | P_i = P_k\}, j \neq k$ and $k\epsilon[1, N]$,

   rank 1 is given to the candidates scoring the maximum FSP.

4. Minimum number of different symbols (NDS) This strategy gives preference to the candidates whose phonemes appear in the majority of other candidates.

   $NDS(C_j) = \sum_{i=1}^{l}\sum_{k=1}^{N}\delta(P_{j,i}, P_{k,i}),$

   where l is the number of phonemes in a pronunciation P_j, δ is a Kronecker delta, defined as 1 if $P_{j,i} \neq P_{k,i}$ and 0 otherwise, and N is the number of candidates, rank 1 is given to the candidate scoring the minimum NDS.

5. Weakest arc frequency (WL) The candidate whose lowest arc frequency value is the highest

   $WL(C_j) = \min_i\{f_i\}$,

   rank 1 is given to the candidate scoring the maximum WL.

Proposed strategies (Polyákova and Bonafonte, 2008a, 2009):

6. Weighted arc product frequency (WPF) Similar to the 1^{st} strategy described in Marchand and Damper (2000), where for each arc, the corresponding arc frequencies are multiplied

   $$WPF(C_j) = \prod_{i=1}^{n} f_i,$$

   being n the candidate length, or the number of arcs that comprise the candidate. Rank 1 is given to the candidate scoring the maximum $WPF()$. The difference is that in this strategy for each phoneme arc, the frequency of its appearance is divided by k, the number of different phoneme arcs found in the dictionary for the corresponding letter arc, L_j. For example if our unknown word is `#infinity#`, and if in the pronunciation lattice we have a path that starts with a letter arc L_1= `#in`, and the corresponding phoneme arc with frequency equal to 12 is A_1=/ˈ ə ŋ/, in order to obtain weighted arc frequency, we have to divide 12 by the number of different phoneme arcs available in the dictionary for the letter arc `#in`.

7. Strongest first arc (SF) This strategy aims at finding analogy in prefixes. The candidate with the highest frequency score for the first arc is given rank 1.

8. Strongest last arc (SL) This strategy is analogous to the previous one but for the suffixes. The candidate with the highest frequency score for the last arc is given rank 1.

9. Strongest longest arc (SLN) The candidate who has at the same time the longest and the most frequent arc is given rank 1. First the longest arc is chosen and if there is a tie the next step is to choose the most frequent one. The candidate that have the longest arcs seem to be more reliable, and of course, the more frequent the arc is the stronger the analogy is.

10. Same symbols multiplied by arc frequency ($SSPF$) The 10^{th} strategy is similar to the fourth one (NDS). NDS gives preference to the candidates whose phonemes appear in the majority of other candidates.

    $$NDS(C_j) = \sum_{i=1}^{l}\sum_{k=1}^{N}\delta(P_{j,i}, P_{k,i}),$$

being l the number of phonemes in a pronunciation, δ the Kroneker delta, equal to 1 if $P_{j,i} \neq P_{k,i}$ and 0 otherwise, and N the number of candidates. In our strategy, when counting the common phonemes, we also take into consideration the phoneme arc frequencies. If a candidate has a common phoneme with other candidates, we assign it a higher score, depending also on the number of times the phoneme arc containing that phoneme appears in the dictionary

$$SSPF(C_j) = \sum_{i=1}^{l} \sum_{k=1}^{N} (1 - \delta(P_{j,i}, P_{k,i})) * f_i^j$$

11. Frequency product, same pronunciation ($PFSP$). This strategy is a combination of 1^{st} and 3^{rd} strategies in Marchand and Damper (2000). The 3^{rd} strategy gives the privilege to the candidates sharing the same pronunciation with the others, rank 1 is given to the candidate scoring the maximum FSP.

    $FSP(C_j) = \text{cand}\{P_j \mid P_j = P_k\}, j \neq k \ and \in [1, N]$

    In eleventh strategy all the candidates that share the same pronunciation obtain the same score equal to the combination of the scores assigned to each one of the candidates by the 1^{st} strategy

    $PFSP(C_j) = \sum_{\forall k, P_{k=P_j}} \sqrt[n]{PF(C_k)}.$

    Different strategies work with different aspects of analogy. They combine some hypotheses related to morphemic analogy as well as statistical analogy. For example, frequencies of suffixes and prefixes are prioritized by different strategies. The final score for the candidate is directly proportional to the number of phonemes it shares with the others. If two or more candidates share the same pronunciation, this pronunciation is believed to be more reliable and these candidates are prioritized. These measures are used separately or combined across the strategies.

3.4.2 Experimental results

The experiments described in this section were performed on three dictionaries, Unisyn and LC-STAR, used in our previous experiments and Nettalk previously used in literature to evaluate pronunciation by analogy (Marchand and Damper, 2000; Polyákova and A.Bonafonte, 2005; Polyákova and Bonafonte, 2008a).

The first thing to do was to find out how well each strategy performed. The strategy mask is a binary string, where 1 means the strategy is included in the final result and 0 otherwise. The results for eleven strategies for both dictionaries are given in Table 3.3.

Strategy name	Strategy mask	NetTalk		LC-STAR		Unisyn	
		Ph. acc.	W. acc.	Ph. acc.	W.acc.	Ph. acc.	W.acc.
PF	10000000000	89.70	57.48	94.76	73.59	96.28	79.94
SDPS	01000000000	88.00	50.59	92.68	65.31	94.82	73.49
FSP	00100000000	89.95	59.06	95.60	79.34	97.49	87.24
NDS	00010000000	90.27	57.43	95.53	76.73	97.10	83.46
WL	00001000000	88.56	53.75	94.07	71.44	95.66	77.09
WPF	00000100000	89.69	57.02	94.96	75.05	96.62	82.01
SF	00000010000	89.15	55.84	92.95	66.17	95.38	74.96
SL	00000001000	87.92	50.28	94.46	72.26	95.83	77.50
SLN	00000000100	88.68	54.01	92.82	65.23	95.09	73.62
SSPF	00000000010	89.99	58.30	94.95	74.61	96.48	81.31
PFSP	00000000001	91.14	62.94	96.01	80.32	97.72	88.09

Table 3.3: Word and phoneme accuracy for each strategy for NETtalk, LC-STAR and Unisyn dictionaries.

From the results above we can see that the strategies give different performance for different dictionaries. The best strategy is the proposed eleventh strategy and the second best is the original 3rd strategy for both dictionaries. For NETtalk dictionary, two proposed and three original strategies made it to the top 5 strategy list while for LC-STAR dictionary the top 5 strategies included three proposed and two original ones.

For our implementation of the 5 original strategies the best results were obtained for the combination of only the first and third strategies "10100". At the next step we evaluated all possible strategy combinations (2^{11}) of 11 strategy masks represented by binary numbers. The accuracy obtained for NETtalk lexicon was 63.04% words and 91.02% phonemes correct; and 80.94% words and 96.07% phonemes correct for LC-STAR lexicon. These results are slightly different from those reported in Marchand and Damper (2000) as well as the scores obtained for each original strategy with our system, but we believe that it is due to the implementation nuances. The top 5 combination results including the proposed strategies are given in Table 3.4 and Table 3.5.

Eleventh strategy is present throughout Table 3.4 and Table 3.5 and its contribution to the improvement of overall score is the greatest for both lexicons. The best strategy combination results obtained are higher than those obtained previously by combining only the original strategies. The word error rate decreased from 36.96% to 36.5% for NETtalk and for LC-STAR from 19.06% to 18.78%. That is between 1.5 and 2.5 percentage points of error decrease.

Finally, in order to be able to compare it to other methods PbA algorithm was also

S. combination	Ph. acc.	W. acc.
11110010011	91.28	63.50
01110110011	91.24	63.40
01100010001	91.30	63.40
01100010011	91.29	63.35
00100010001	91.31	63.35

Table 3.4: Top 5 strategy combination results for NETtalk lexicon.

S. combination	Ph. acc.	W.acc.
00101000001	96.13	81.22
01100001001	96.08	81.12
01111100001	96.11	81.04
01101001001	96.04	81.04
00101001001	96.09	81.04

Table 3.5: Top 5 strategy combination results for LC-STAR lexicon.

evaluated on Unisyn lexicon. The results are given in Table 3.6

S. combination	Ph. acc.	W.acc.
00101000001	97.80	88.54
00100000001	97.78	88.52
01100000011	97.75	88.47
11100000001	97.78	88.46
01100000001	97.77	88.46

Table 3.6: Top 5 strategy combination results for Unisyn lexicon.

As well as for the other two lexica, the eleventh strategy is present in all best strategy combinations. The word accuracy for this lexicon is higher because it is the biggest in size and it contains word derivations.

It is interesting to further analyze the specifics of the algorithm performance, for example to see how it varies depending on word length. Our hypothesis is that the performance of the PbA algorithm is different for short and for long words. For this purpose, the test dictionary was split into several dictionaries containing words of the same length. Word length ranged from 3 to 17 letters per word. The words that had only two letters were added to the 3-letter word lexicon subset. For the LC-STAR dictionary the distribution of words by length is a Gaussian with its mean situated approximately at length 8 3.8 (see figure 3.8). It is true for both train and test lexica.

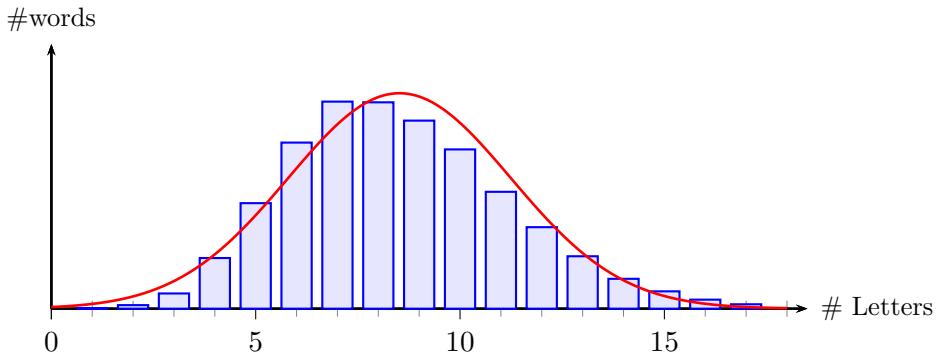


Figure 3.8: Distribution of words as a function of the number of letters

As we expected, the performance of each strategy depends on word length. This information could be used to select the most suitable strategy for each case. However, the new strategy 11 has happened to be the best in all cases. When looking at word accuracy, Figure 3.9, in the great majority of the cases the eleventh strategy is the best. For word length equal to 6 and 7 letters the word accuracy is the highest. The strategies strongly disagree on very short and very long words. Word accuracy is higher for shorter words, since there are less phonemes and the probability of having at least one phoneme wrong is lower.

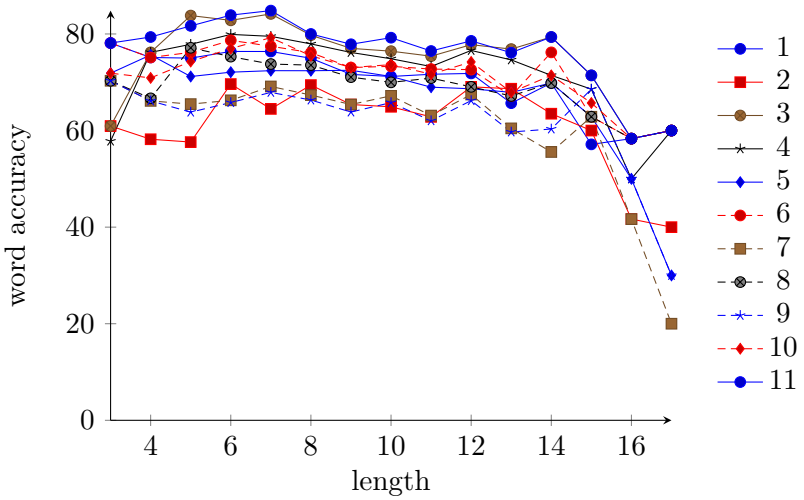


Figure 3.9: Word accuracy for each strategy as a function of word length.

For phoneme accuracy, Figure 3.10, the eleventh strategy gives the best results. The phoneme accuracy starts being rather high for 5-letter words and remains this way even for very long words, but like in for word accuracy case the strategies disagree for very short and very long words.

The best phoneme accuracy results very obtained for the words consisting of 14 letters using the eleventh strategy.

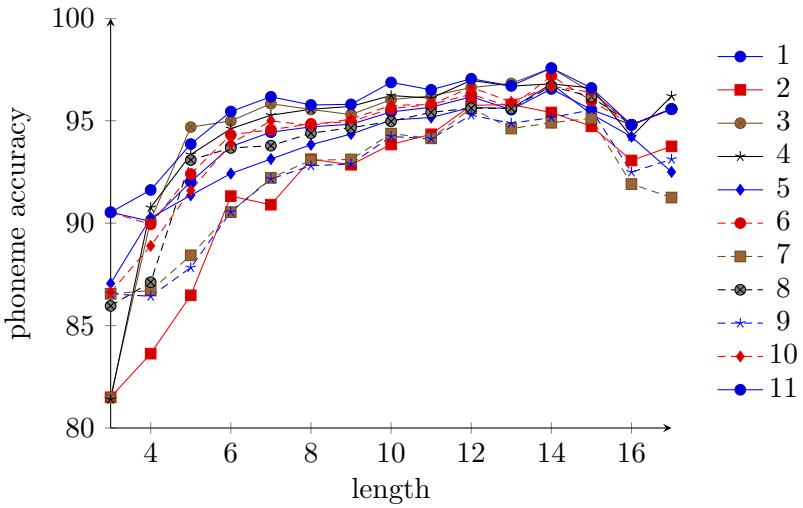


Figure 3.10: Phoneme accuracy for each strategy as a function of word length.

Finally, Table 3.7 compares the results obtained with the PbA algorithm to the ones previously obtained in Polyákova and Bonafonte (2006). These results are given for the LC-STAR lexicon as well. This comparison allows us to conclude that PbA is one of the best G2P methods up to now.

Classifiers	baseline
DT	67.47
FST	79.38
HMM	47.54
PbA	81.22

Table 3.7: Word accuracy for different G2P methods.

3.5 Learning from errors in G2P conversion

Another important aspect of language learning is learning from errors. When a new word is pronounced erroneously it is usually corrected by a native speaker or a teacher, this usually creates an embarrassing situation and our brain, emotionally stimulated, learns not to make the same mistake in a situation similar to this one. The more examples of similar errors, given a similar error occurrence situation we have, the better we learn not to repeat it.

The Machine learning (MLR) methods applied to the task of automatic pronunciation up to now leave room for some further improvements. In English, errors mostly appear when assigning pronunciation to vowels, whereas consonants are usually predicted better. To improve the performance of the G2P transcription system, in this chapter, we propose to use an approach based on learning from errors made by a G2P converter at the first step.

The transformation-based error-driven learning algorithm invented by Brill (1995), was successfully applied to such NLP tasks as part-of-speech tagging, word sense disambiguation, phrase chunking etc. The high level of accuracy achieved for these tasks proves the effectiveness of this data-driven method.

The experiments were performed on two lexica of American English. Four baseline conversion methods were used to obtain the necessary initial prediction and to train the error-driven system. These methods are: CART (Black et al., 1998b), FST (Galescu and Allen, 2001), HMM (Taylor, 2005) and such a naive prediction as the most-likely phone.

The error correction capacity highly depends on the size of the corpus and the number of errors available for training the transformation rules. The data was divided into 3 sets, 45% of which was designated for training, another 45% for development and finally the remaining 10% for test. The prediction obtained by G2P methods for the training set was rather low in accuracy, thus leaving a lot of errors for training of the rules. The improvements obtained for the test set by TBL in percentage points were higher than those obtained when a 90% 10% data split was used. However, the initial results were so low that it was less effective than when a standard partition was used.

The TBL algorithm, originally invented by Brill (1995), consists in learning transformation rules from the training corpus labeled with some initial classes. The TBL algorithm uses templates to generate rules that generalize the transcription errors obtained by the initial G2P method. The templates consist of several features that for this particular task can be the phoneme predicted at the previous step of the algorithm, letter context, etc. Some examples of rule templates are given below

```
                let_-1 let_0 let_1  →  ph
            let_1 let_0 let_1 ph_0  →  ph
let_-1 let_0 let_1 ph_-1 ph_0 ph_1  →  ph
```

Here `let_0` represents the letter corresponding to the current phoneme, while `let_-1` and `let_1` define the surrounding orthographic context. In this case, `ph_-1` and `ph_1` represent the surrounding predicted phoneme context and `ph_0` represents the predicted phoneme itself. `ph` is the correct phoneme to which `ph_0` should be transformed.

The erroneous tags in the training corpus serve as the basis for deriving error-correcting transformation rules. During the learning process TBL algorithm learns rules iteratively. Its goal is to correct as many errors as possible in the training corpus. Rules are generated and

applied to the current state of the training corpus at each iteration. The number of errors corrected is also called the number of good applications. The number of bad applications is defined by the number of times that application of a rule has introduced a new error. The score of each rule equals to the difference between the number of good and bad applications. The rule capable of correcting the largest number of errors at each iteration(the one with the highest score) is applied to the entire training corpus and appended to the final rule list. The scores of other rules affected by the application of the best rule at current iteration are also updated. The rule learning process continues until no rule that improves the accuracy of the training prediction could be found or a best rule with a score lower than the preset threshold is generated.

Using the TBL algorithm to correct the prediction previously obtained by another classifier allows capturing the imperfections of previous approaches into a set of context-dependent transformation rules, where the context serves as a conditioning feature. During the test phase, a rule from the list is applied whenever a match between the input set of features and those defined in the rule is found. In the evaluation phase the rules are applied to correct the errors in the initial prediction for the test data, in the same order that they were generated.

Figure 3.11 shows the scheme of combination of data-driven G2P methods with TBL. The transformation rules are derived from the errors in the initial prediction obtained by a

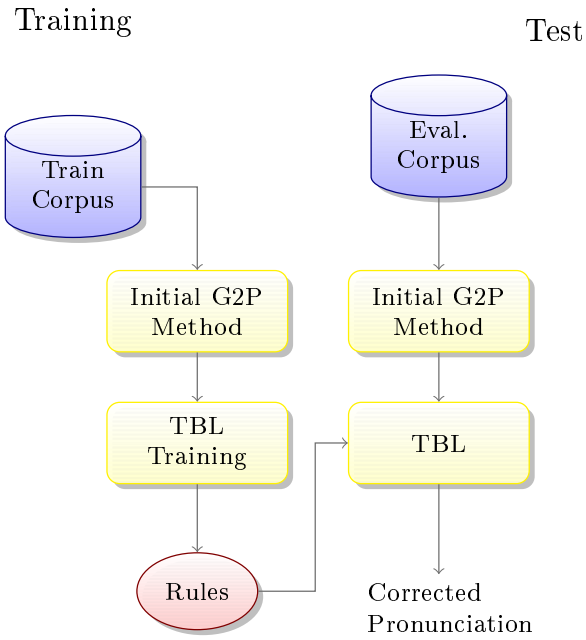


Figure 3.11: Scheme of combination of TBL with other G2P methods.

previous classifier for the training data. The TBL algorithm not only allows to correct errors

obtained in the previous predictions but also an easy combination of different conditioning features for error correction. The performance of TBL algorithm highly depends on the size of the training data and the number of prediction errors obtained by the initial classifier. A higher error ratio in the training phase and a larger size of training corpus lead to better correction results. This was analyzed for the case of Mandarin polyphones prediction in Zheng et al. (2005).

Advantages of the TBL

The main difference between a manually derived set of rules and the set of rules extracted by TBL is that the latter does not need to be elaborated by experts. The method is fully automatic apart from the rule template creation step. The order of rule application also does not require any knowledge about the language. It is established automatically during the system training. The rules that have the highest best score are placed at the top of the list and then the rules with lower scores are added. The rules are language-independent and can be applied to any supervised prediction task in combination with any machine learning technique, while the manually elaborated rules are non-transferable to other languages, which is an additional disadvantage to their already very high development cost. Like CART or FST, this method requires the data to be aligned in a one-to-one manner. As sometimes data alignment is not unique, it introduces a limitation as to what errors can be corrected by rules.

During the training process the main goal of the algorithm is to capture the regularity exiting between the errors in the first prediction and to choose best transformation rule according to the context where the error was made. The learning process is similar to when a human learner is trying to master a language. A human learner acquires knowledge from errors by memorizing the circumstances where the error occurred trying to avoid the same error in the future, given alike circumstances. If compared to foreign language learning, many examples of similar situations can be found. These could be incorrect word order, erroneously memorized noun gender (for languages that differentiate gender), stress misplacement, etc. This learning mechanism is activated every time the error occurs. It can not only be applied to language learning but to many other human activities also. When doing something for the first time, making mistakes is rather natural, but after having found out the right way to do it, it is unlikely to repeat the same or similar mistake given the same conditions. The TBL algorithm works in the same way: it generates rules that try in the best way to generalize the transcription errors obtained by the initial prediction method. Once the patterns *transformation condition* $\rightarrow$ *correct answer* are captured, the TBL applies these patterns to correct the errors. The error itself usually forms part of the transformation condition as well as the conditions of its occurrence.

Another advantage of the transformation-based learning algorithm is that it uses previously corrected predictions. These intermediate results generate more reliable corrections.

For instance, if the rules A and B are the following:

Rule A: if `let_-1 = _, let_0 = k`, and `let_1 = n`, change `fon_0` to `null` (e.g. as in the word *know*).

Rule B: if `fon_-2 = null, fon_-1 =` [n], `let_-2 = k, let_-1 = n, let_0 = i`, and `fon_0 =` [ɪ], change `fon_0` to [aɪ] (e.g. *knife*, where the letter *k* is silent)

The rules are applied to correct the prediction in the same order they appear in the list, first rule A, then rule B. If there are any errors in the transcription of the silent *k* in our prediction, rule A corrects them and, therefore, rule B also applies. Our initial prediction has significantly improved using rule A before rule B.

3.5.1 Experimental results

In this section, we present the outcome of combined application of various classifiers to the grapheme-to-phoneme task as well as the results obtained by each classifier alone.

The experiments were carried out using LC-STAR and Unisyn lexica for American English. To obtain the results, the fnTBL toolkit, kindly provided for public use by its authors was used (Ngai and Florian, 2001). The fnTBL differs from the original Brill's TBL in the way that the objective function is calculated and reaches a speed up, without reducing the system's performance.

First, the baseline results for both lexicons were obtained, then TBL algorithm was applied to the output of all machine-learning methods used at step 1, to correct their pronunciation as shown in Figure 3.11. Besides G2P methods described at the beginning of this chapter, a naive prediction as assigning the most-likely phoneme seen in the training to each letter, was considered. A one-to-one alignment was necessary in all cases.

In Table 3.8 and Table 3.9 grapheme-to-phoneme results for LC-STAR lexicon of American English are given.

Table 3.10 and Table 3.11 represent grapheme-to-phoneme results obtained for the Unisyn lexicon of US English.

It is seen that FST gives a much higher word accuracy than Hidden Markov Models (HMM) or Decision Trees (DT) for both lexica. The method that gives the worst results is HMM. To improve these results a preprocessing similar to the one proposed in Taylor (2005) might be necessary.

	baseline	cont=3	cont=4	cont=5
ML	59.70	95.17	95.59	95.76
DT	93.93	94.64	94.85	95.00
FST	93.63	95.73	95.87	95.92
HMM	84.16	92.66	93.15	93.29

Table 3.8: Baseline G2P results and those improved by combining 4 transcription methods with TBL for the LC-STAR lexicon(phoneme accuracy).

	baseline	cont=3	cont=4	cont=5
ML	1.07	75.67	77.46	78.26
DT	68.32	73.06	74.13	74.68
FST	75.66	78.79	79.33	79.63
HMM	47.54	67.01	68.70	69.08

Table 3.9: Baseline and improved G2P results for the LC-STAR lexicon (word accuracy)

	baseline	cont=3	cont=4	cont=5
ML	56.36	96.68	97.01	97.12
DT	95.26	96.44	96.60	96.67
FST	97.30	97.46	97.47	97.49
HMM	86.93	94.38	94.45	94.75

Table 3.10: Baseline and improved G2P results for the Unisyn lexicon (phoneme accuracy)

	baseline	cont=3	cont=4	cont=5
ML	1.42	82.39	84.00	84.61
DT	72.67	80.74	81.63	82.08
FST	86.65	87.25	87.28	87.36
HMM	54.87	74.19	74.32	74.95

Table 3.11: Baseline and improved G2P results for the Unisyn lexicon (word accuracy)

Applying a number of rewrite rules to the lexicon, would allow HMM to obtain better results with, as well as the use of context-sensitive models and stress patterns (in case of a stressed lexicon). Table 3.8 through Table 3.11 show the result of combination of TBL with four classifiers for different context lengths, the baseline results are also given. The goal was to learn rules that would be able to correct the prediction obtained by other classifiers. The rules were learned for 3 different sizes of letter context. The maximum letter context included in the rules varied from 3 to 5 letters to the left and to the right. Applying error-transformation rules to the output of four different algorithms shows significant improvements. The most significant improvement was achieved for the methods whose performance was the poorest at the first step; it can be explained by the abundance of errors that allowed their better generalization by the transformation rules.

The word accuracy results obtained by DT were improved by a measure of 8-10 percentage points, the HMM results were improved by about 20 percentage points, and the FST prediction improvement range was equal to 1-5 percentage points. The hugest improvement was achieved for the most likely phone prediction, where the preliminary prediction scored only about 50% phonemes and 1% words correct. Before TBL application the correctly predicted phonemes were mostly consonants. The improvement ranged from 75 to 80 percentage points in terms of word accuracy and the results are similar to those obtained by combining TBL with the best performing method, FST, which shows the effectiveness of the rules. The better was the baseline prediction, the less context-sensitive was the improvement.

The largest context gave the best results, although it was more expensive in computational terms. The time needed for computation also depended on the number of baseline errors. If there were few errors to correct in the training corpus the algorithm converged faster.

The TBL algorithm was also applied to improve pronunciations obtained by analogy. A way to combine strategies by using predictions obtained by different strategies as additional features for the TBL algorithm was attempted. However, no significant improvements were achieved. The improvements varied somewhere between 0.10% and 0.40% percentage points for the LC-STAR lexicon. An attempt to combine PbA strategies using TBL was also made. Outputs of different strategies were included as conditioning features, in order to derive rules to correct the prediction obtained by the best PbA strategy combination this didn't lead to any significant improvements either. After achieving such little improvement in the experiment, it was not repeated for other lexica.

Figure 3.12 shows the distribution of the number of errors per word before and after the application of the TBL algorithm to correct pronunciation. This analysis was carried out for the results obtained with HMM method using the LC-STAR lexicon. This method gave the poorest results.

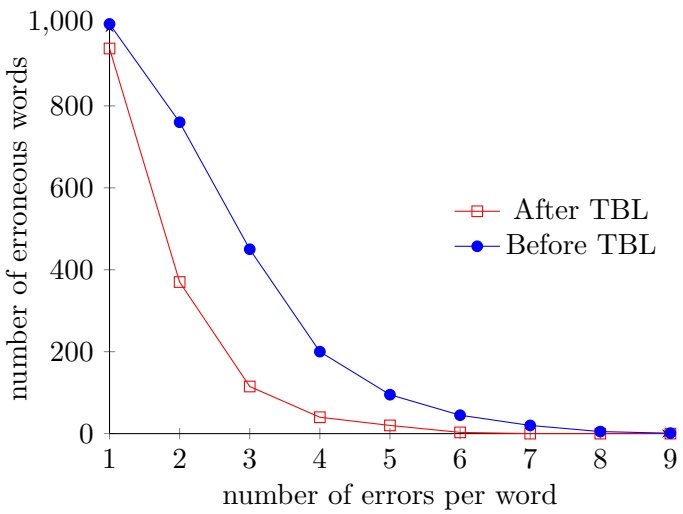


Figure 3.12: Number of words as a function of number of errors per word.

Figure 3.12 shows that the application of TBL reduces the number of words with severe pronunciation errors (more than 3 errors per word) by almost seven times and the rest of errors by 1.5 times. Words with only one error are still numerous and it could be due to the fact that these errors are very specific. They usually occur as a result of a confusion between different vowels represented by the same letter. These are the language irregularities that are very difficult to generalize as for the automatic algorithms that try to pronounce English text as for human English learners. The generalization capacity of the TBL is also limited by the inconsistency of some alignments and reference transcriptions.

The goal to obtain better G2P conversion results was set and achieved by means of applying a set of transformations learned from errors. The transformation rules were learned automatically from a training corpus previously labeled using four classifiers. The combination of all methods with transformation-based error-driven algorithms significantly improved the results obtained by these methods alone. The best G2P results were obtained by combining FST and PbA combined with TBL algorithm, although for FST these improvements were more significant.

3.6 G2P results for other languages

Besides English, the pronunciation was automatically inferred for Spanish, Catalan, Slovenian, Turkish and German using DT classifier and subsequently learning the transformation rules aimed to correct the errors of the first classifier. The results are given in Table 3.12 The maximum context used was limited to +-5 letters and +- 3 phones.

For Spanish and Turkish the results obtained with DT are very good due to its shallow orthography. For Catalan we obtained an improvement of 5 percentage points after applying the correction rules. The improvements obtained using the transformation-based rules for Slovenian and German varied between 9 and 2 percentage points.

	Baseline(DT)	TBL
Spanish	98.49	98.91
Catalan	83.72	88.79
Slovenian	66.39	88.87
Turkish	97.45	97.57
German	71.53	80.13

Table 3.12: Baseline (DT) and improved by TBL word accuracy for Spanish and Catalan and other languages for LC-STAR lexica of common words.

3.7 Specifics of G2P conversion for English

This section explains what specific linguistic aspects should be taken into consideration in order to obtain a higher quality synthesized speech. Phonetic weak forms, phonotactic rules and syllable boundaries may influence the output speech intelligibility and quality. The research in the framework of this thesis regarded these aspects for English, although phonotactic rules are also applicable in Spanish and Catalan, and had been already integrated into the system. This work was motivated by UPC's speech synthesis team participation in 2007 and 2008 Blizzard Challenge Initiative, an evaluation campaign whose objective was to compare TTS systems at an international level. The global goal set in the framework of Blizzard challenge was to improve the quality and intelligibility of the synthesized speech.

The system used for synthesis was Ogmios, the UPC Text-to-Speech system used for the evaluation. Initially, the system was designed to cover only Spanish and Catalan languages (Bonafonte et al., 2006b) and, for the Blizzard Challenge its features were extended to cope with American (Blizzard 2007) and British English (Blizzard 2008) as well.

3.7.1 Phonetic transcription in connected speech

The first task of the TTS system is to detect the structure of the document and to transform the input text into words. The system was extended to cover English language as well. The rules for tokenizing and classifying non-standard words are very similar to those used for Spanish and Catalan. The rules for expanding each token into words are language dependent

but are based in a few simple functions (spellings, natural numbers, dates, etc.). The second process is the POS tagger. Ogmios includes a basic statistical tagger. The n-gram statistics were estimated using 1 million of tokens from the WSJ Corpus using the Penn Tree bank POS system.

The disambiguated text was fed to the phonetic module, whose goal was to provide the pronunciation of the words. This is used not only for producing the test sentences but also for transcribing the training database which is used for building the voices.

When a word was absent from the system dictionary, the G2P conversion was necessary. A finite state transducer trained on the Unisyn dictionary was inferred for this purpose (Galescu and Allen, 2001; Polyákova and A.Bonafonte, 2005; Polyákova and Bonafonte, 2006).

Some hand-crafted rules were applied to model the pronunciation changes produced in continuous speech. For function words, a set of rules was produced based on factors such as word position in the sentence, part-of-speech and phrase accent 3.7.2. In continuous speech, function words usually lose their accented form and full vowels are reduced to shorter ones or schwa. Furthermore, a set of phonotactic hand-crafted rules was applied. These rules cover different phenomena from aspirated plosives, to consonant assimilation and elision 3.7.3. In the training phase, the rules provided several pronunciation hypotheses which were considered by the segmentation process.

3.7.2 Weak forms

Some words in connected speech may undergo phonotactic variations. Isolated words bear always at least one primary stress and, sometimes, a secondary and a tertiary stress as well. Lexical or function words loose they accented form depending on the pitch of the sentence and the surrounding phonemic context. These words become unstressed under the influence of the *sentential* stress that accounts for the prominent syllables on a sentence level (Dutoit, 1997). It is important to define correctly the rules in order to apply the phonetic weak forms for these words since an incorrect stress pattern can be very harmful for the overall naturalness of the TTS output. As opposed to strong forms the unaccented phonetic weak forms show vowel length reduction, vowel and consonant elision (Gimson and Cruttenden, 2001). Some function words such as *I, your, bus, my, nor, so* that usually are used in their strong forms may be reduced to weak forms when the rate of speech is very high.

The list of the weak forms and the conditions under which they were applied in the framework of Blizzard challenges 2007 and 2008 (Bonafonte et al., 2007, 2008) are given in Table 3.13.

Weak forms	Phonetic transcription	Comments
a	[ə]	
am	[m], [ə m]	
	['æ m]	end of the sentence or rhythmic group
an	[n], [ə n]	
and	[ə n d], [n d], [ə n], [n]	
are	[ə]	before a consonant
	[ə r], [r]	before a vocal
	['ɑː]	end of the sentence or rhythmic group
as	[ə z]	
at	[ə t]	
	['æ t]	end of the sentence or rhythmic group
be	[b ɪ] or [b i]	
	[b iː]	end of the sentence or rhythmic group
been	[b ɪ n]	
but	[b ə t]	
can (aux)	[k ə n], [k n]	
	['k æ n]	end of the sentence or rhythmic group
could	[k ə d], [k d]	
	['k ʊ d]	end of the sentence or rhythmic group
do (aux)	[d ə]	before a consonant
	[d]	after a vocal and before a consonant V + [d] +C
	[d u]	before a vocal
	[d uː]	end of the sentence or rhythmic group
does (aux)	[d ə z]	
	['d ʊ z]	end of the sentence or rhythmic group
for	[f ə]	before a consonant
	[f ər], [f r]	before a vocal
	['f ɔː]	end of the sentence or rhythmic group
from	[f r ə m], [f r m]	
	['f r ɒ m]	end of the sentence or rhythmic group
had (aux)	[h ə d], [ə d], [d]	only if auxiliar
	['h æ d]	end of the sentence or rhythmic group
has (aux)	[h ə z], [ə z]	only if auxiliar
	['h æ z]	end of the sentence or rhythmic group
have (aux)	[h ə v], [ə v]	only if auxiliar
	[h'æ v]	end of the sentence or rhythmic group
he	[h ɪ], [iː], [ɪ] or [h i]	
her	[h ə]	the forms with [h] are used after a pause
	[ɜː], [ə]	
him	[ɪ m]	
his	[ɪ z]	
is	[s]	before an unvoiced consonant
	[z]	before a voiced consonant
me	[m ɪ] or [m i]	
must	[m ə s t]	

Continued on next page

Table 3.13 – continued from previous page

Weak forms	Phonetic transcription	Comments
	['m ʊ s t]	end of the sentence or rhythmic group
not	[n t]	
of	[ə v], [v], [ə]	
Saint	[s ə n t], [s n t]	
St.	[s ə n], [s n]	
shall	[ʃ ə l], [ʃ l]	
	['ʃ æ l]	end of the sentence or rhythmic group
she	[ʃ ɪ] o [ʃ i]	
should	[ʃ ə d] o [ʃ d]	
Sir	[s ə]	before a consonant
	[s ə r]	before a vocal
some(adj)	[s ə m], [s m]	does not apply if is a pronoun
than	[ð ə n], [ð n]	
that(conj, rel. pron.)	[ð ə t]	does not apply to pronoun or demostrative adjective
the	[ð ɪ] or [ð i]	before a vocal
	[ð ə]	before a consonant
them	[ð ə m], [əm], [m]	
there(indef adv.)	[ð ə]	before a consonant
	[ð ə r]	before a vocal
to (into, onto, unto)	[t ə]	before a consonant
	[t ʊ] or [t u]	before a vocal
	['t uː]	end of the sentence or rhythmic group
us	[ə s], [s]	
was	[w ə z]	
	[' w ɒ z]	end of the sentence or rhythmic group
we	[w ɪ] or [w i]	
were	[w ə]	before a consonant
	[w ə r]	before a vocal
	['w ɜː]	end of the sentence or rhythmic group
who	[uː], [ʊ]	
	[h ʊ], [h u]	after a pause
will	[ə l], [l]	
	['w ɪ l]	end of the sentence or rhythmic group
would	[w ə d], [ə d], [d]	
	['w ʊ d]	end of the sentence or rhythmic group
you	[j ʊ] or [j u]	

Table 3.13: Lexical words in connected speech or phonetic weak forms for British English (Gimson and Cruttenden, 2001)

3.7.3 Phonotactic rules

Besides phonetic weak forms, another type of constraint rules, called phonotactic rules were applied to the phonetic transcription of the words obtained in isolation. Phonotactics defines what sound combinations are allowed to occur in a language (Whitley, 2002). The list of phonotactic rules used for Blizzard challenges 2007 and 2008 are given in Table 3.14

3.7.4 Syllabification

In speech synthesis the syllable can be used as a feature in several processes from deriving segmental duration to selection appropriate speech units in concatenative speech synthesis. It is important to have consistent criteria to split words into syllables. The most sonorous are the open vowels, followed by close vowels, laterals, nasals, approximants, trills, fricatives, affricates, plosives and flaps in the descending order of sonority.

In most languages syllables are considered to be stable and natural units and are even believed to carry more information about the language than letters (Chen et al., 2006). In English, there are no standard syllabification rules, whereas in languages like French or Japanese the division of words into syllables is a straightforward process. The sonority hierarchy shows the number of syllables in an utterance. The syllable can be divided into three parts: the onset (all the phonemes before the syllable peak), the peak (the most sonorous vowel of the syllable) and coda (all the phonemes that follow the syllable peak).

The sonority scale helps to define these parts. The number of syllables is equal to the number of sonority peaks that can be represented schematically as for the word *Manchester* in Figure 3.13.

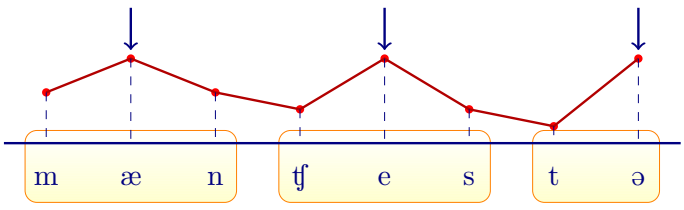


Figure 3.13: Sonority diagram for the word *Manchester*.

The consonants that do not constitute the sonority minimums and precede the peak are placed in the onset and the consonants that follow the peak are placed in the coda. It is difficult to decide whether to place the consonants situated at the local sonority minimums at the end of one syllable or in the beginning of the next one.

Gimson and Cruttenden (2001) mentions several types of placing the syllable boundaries: morphemic principle in a way that the syllable boundaries coincide with the division of

Rule type	Condition and Result	Example
palatalization	[d] before [j] → [ʤ]	*did you*
palatalization	[s] before [tʃ] → [ʃ]	*question*
vow. reduction	[unstressed_vow (except last syl.)] → [ə]	*spectrogram* [s p ’e k t r o g r æ m] →[s p ’e k t r ə g r æ m]
vow. reduction	[unstressed_vow., last syl.] → [ə]	*awful* [’oː f u l] → [oː f ə l]
vow. deletion	if [sylA][sylB][sylD] [a]	*chocolate* [’tʃ o ‘k ə l ɪ t] → [’tʃ o k l ɪ t]
vow. deletion	[ə] between 2 voiceless stops is deleted	*multiply* [’m ʊ l t ə p l aɪ] *to* [’m ʊ l t p l aɪ]
vow. deletion	[ə] after [consonant] before [sylA] is deleted	*police* [p ə l ’i s] → [’p l i s]

Table 3.14: Phonotactic rules for British English (Bonafonte et al., 2007, 2008)

.

[a] If a primary stressed syllable A is followed by a secondary stressed syllable B and then the unstressed syllable D, delete the vowel from B

word into morphemes; phonotactic principle that aligns syllable boundaries to parallel syllable codas and onsets at the word-initial and word-final positions; allophonic principle (syllable division is carried out in a way to best predict the allophonic variation); maximum onset principle sets preference for assigning the word-medial consonants to onsets. These principals are conflictive with each other. For our purposes, we place the syllable boundaries right after the sonority minima.

Different databases use different syllabification methods. Knowing the adequate syllable structure may be use useful in many speech technologies applications like representation of important information for word models building in ASR or taking the syllables or demi-syllables as the basic units in concatenative speech synthesis (Marchand and Damper, 2007). However, recent research carried out in the same work shows that pronunciation by analogy cannot be improved by introducing syllable boundaries. In the framework of this thesis, syllabic information is important for prosody generation, although the information about the number of the syllables in the word seems to be sufficient. This is achieved using the syllabication by sonority method described above (Bonafonte et al., 2007, 2008).

3.8 Error rate versus word length

It is interesting to know the error distribution versus the word length. For the LC-STAR lexicon excluding the stress marks we plotted the word error probability distribution, $f(l) = N_{er}(l)/N_{tot}$ as a function of the grapheme number per word, (where $N_{er}(l)$ is the number of errors in l-lettered words, and $N_{tot} = \sum_{l} N_{er}(l)$ is the total number of errors in the given experiment). The error distribution obtained for the LC-STAR lexicon as a function of word length is plotted in Figure 3.14 (open squares).

As it is seen from Figure 3.14 the erroneous pronunciations are most likely generated for 9-letter words. However, one should keep in mind that in English the word frequency is a non-monotonic function of word length. It was recently shown in Sigurd et al. (2004), that in English the word frequency obeys the distribution with the maximum at $l = 3$ (see Figure 3.14, filled circles).

$$f_w(l) = 11.74 l^3 (0.4)^l \tag{3.5}$$

Word distribution by length for the British English Example Pronunciation Dictionary was taken from Damper et al. (2004), for which average word length is 8.87 letters with a standard deviation of 2.58 letters (see Figure 3.14, filled triangles).

The difference between these distributions (filled circles and filled triangles in Figure 3.14) is that in Sigurd et al. (2004), the distribution is given for running words,

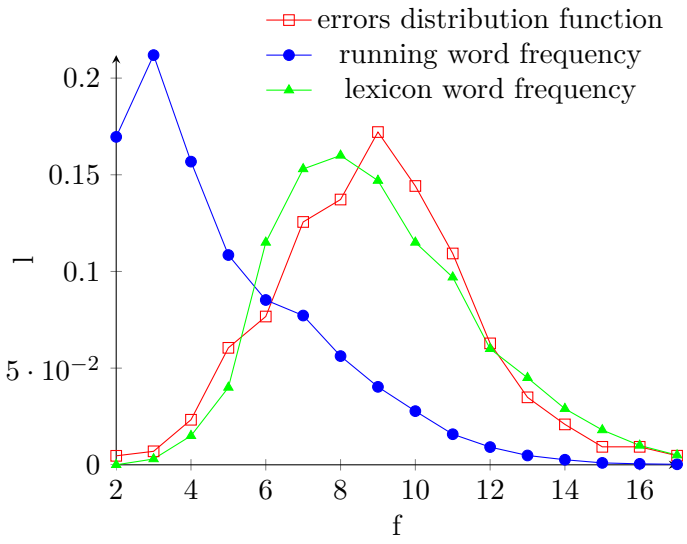


Figure 3.14: Probability distribution function of errors and word frequencies versus the number of letters per word.

while Damper et al. (2004) estimated the distribution for the BEEP dictionary (Robinson, 1997) after deleting words with length 3 or less, homonyms and words with incorrect pronunciation. In the light of these distributions, one can expect the probability of conversion errors to be lower in practice, as the probability of occurence of long words in spoken language is significantly inferior that for any lexicon.

3.8.1 Probability of errors in G2P conversion

In order to be able to optimize the phonemization process it is important to know the probability of error as a function of word length in a corpus. Let us define the probability of error in l-letter words to be $P_{er}(l) = N_{er}(l)/N_c(l)$, where $N_{er}(l)$ is the number of l-letter words containing errors, and $N_c(l)$ is the total number of l-letter words in the corpus. The total probability of error in the experiment is

$$P_{tot} = \frac{N_{tot}}{N_c} = \sum_l \frac{N_{er}(l)}{N_c} \tag{3.6}$$

Where N_{tot} is the number of erroneous words (of any length) and N_c is the number of words in the corpus. Knowing the frequency of l-letter words in the corpus $f_w(l) = N_c(l)/N_c$,

we can substitute N_c in Equation 3.6.

$$P_{tot} = \frac{N_{tot}}{N_c} = \sum_l \frac{N_{er}(l)}{N_c(l)} f_w(l) = \sum_l P_{er}(l) f_w(l) \tag{3.7}$$

Equation 3.7 allows us to estimate probability of error in our resulting grapheme-to-phoneme correspondences. Two limit cases of Equation 3.7 present special practical interest. The first case is when functions $P_{er}(l)$ and $f_w(l)$ have sharp and well separated peaks on l-axis, see exemplary plots in Figure 3.15a. As it is seen from Equation 3.7, in this case, the total probability of error can be very low. Indeed, in the sum over a large interval l in Equation 3.7, we have a lot of terms equal to the product of peak values of one function and close to zero values of the other. The second case occurs when maxima of functions $P_{er}(l)$ and $f_w(l)$ are reached at similar values of l, as it can be seen from Figure 3.15b In this case, the overall grapheme-to-phoneme error probability reaches its maximum, as in the sum in Equation 3.7, high values of one function are multiplied by high values of the other. This means that to reduce the error probability it is necessary to choose a different G2P method with an error distribution $P_{er}(l)$, with a maximum far from the maximum of $f_w(l)$, which depends strictly on the corpus and the language. Based on the analysis and Equation 3.7, we can state that if the functions $P_{er}(l)$ and $f_w(l)$ have sharp and well separated maxima on l-axis, the probability of error is minimal. How can we apply this rule in practice to improve the quality of the pronunciations derived? First of all, the text needs to be analyzed and the frequency of words depending on their length $f_w(l)$ needs to be found, this gives us the value of l_{max} which corresponds to the maximum of the function. Knowing l_{max} we can choose a strategy that with the maximum of the function $P_{er}(l)$ lying as far as possible from that of the function $f_w(l)$. According to the rule, the total probability of error will be minimal in this case. Note that if neither of available G2P methods has a suitable maximum of the function $P_{er}(l)$, in order to reduce the total error probability, different G2P methods may be used for different words lengths.

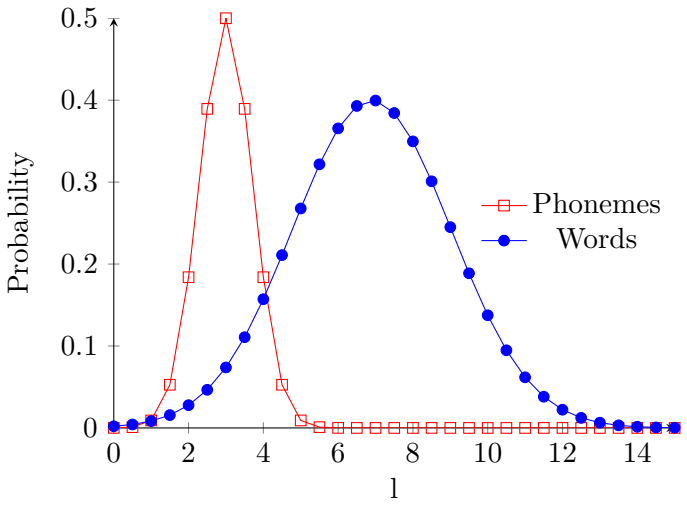
In practice the error distribution function seems to be rather independent from word length, exept for very short of very long words. This can bee seen from Figure 3.14 and from Figure 3.9.

3.8.2 Segmentation of the Speech Database

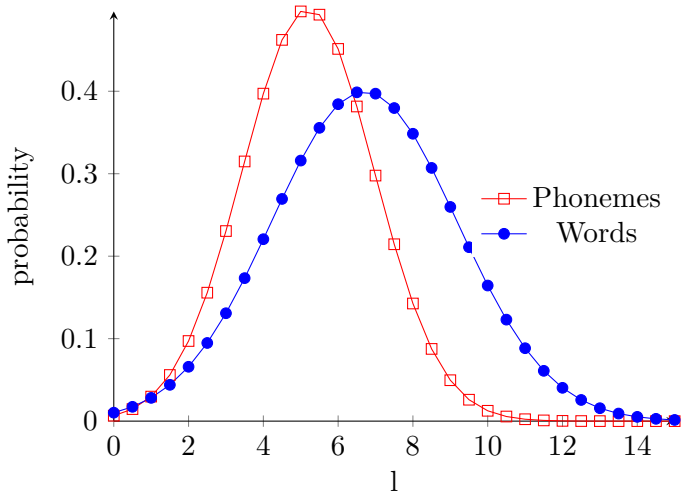
In order to create a voice we need to perform phonetic segmentation of the data.

Typically, a forced alignment between the speech signal the the HMM signal defined by the phonetic transcription is used for segmentation (Taylor, 2009).

At this point of the phonetic trancsription does not match the pronunciation, a segmentation problem exists. This can be caused either by artefacts in the signal (e.g.



(a) Functions $P_{er1}(l)$ and $f_{w1}(l)$ functions have well separated maxima.



(b) The maxima $P_{er2}(l)$ and $f_{w2}(l)$ are reached at similar word lengths.

Figure 3.15: Functions $P_{er}(l)$ and $f_w(l)$

noise) or by imperfect phonetic transcriptions. It is possible to eliminate many of these errors by discarding 10% of the training data with the lowest alignment score. Previous experiments have shown that when a correct phonetic transcription is given, HMM models can achieve similar speech synthesis quality than manual segmentation (Adell et al., 2006).

Anyway, the better is the match between the waveform and the phonetic transcription, the higher is the voice quality produced and the less data will need to be discarded.

Automatic phonetic transcription of a speech synthesis database has to cope with pronunciation variants and pronunciation errors. In order to overcome this problem, the alignment takes into account all possible transcriptions of a single word. The forced alignment process can select the transcription that better matches the signal.

The effectiveness of the phonetic weak forms for G2P is strictly speaker-dependent and sometimes can hamper the correct database segmentation. A better match between speaker's dialect and style and the phonetic transcription will positively influence the resulting synthesis quality. In our participation in Blizzard challenges 2007 and 2008 the use of phonetic weak forms and phonotactic rules led to low HMM probability scores for some pronunciations.

The solution we used in our system was interactive: for every weak form rule we checked if the probability of the segmentation obtained using this rule was high enough for the rule to make it to the list for the speaker in question.

3.9 Conclusions

This chapter compares different machine-learning techniques for G2P conversion module of a TTS synthesizer. The same evaluation techniques and datasets were used to allow a more precise comparison. DT,Finite State Transducers (FST) Hidden Markov Models (HMM), Pronunciation by Analogy (PbA) and Transformation-based error-driven learning (TBL) were applied and compared. These techniques are data-driven, language-independent and corpus-based. Their performance as well as their flexibility and portability match the ones required for high quality grapheme-to-phoneme conversion for mono- and multilingual synthesis.

The experiments were performed using 3 lexica (of American English): Unisyn, NETtalk and LC-STAR. The first series of the experiments performed compare two state of the art G2P conversion techniques, namely DT and FST. The results are obtained for phoneme prediction both with and without lexical stress. As expected, the number of errors obtained for the phonemes without stress marks was significantly inferior. Furthermore, FST gave better performance in both cases.

At the next step, the PbA was reimplemented and applied to the task of G2P conversion. The new scoring strategies were proposed and the improvements were obtained based on these strategies. The 1.5-2.5 percentage points of error reduction was obtained in comparison with the strategies used in Marchand and Damper (2000).

One of the proposed strategies (the eleventh strategy) was found to be the best one, as it can be seen from the results shown in Table 3.3 where it performs better than the other strategies on both dictionaries, as well as from Table 3.4 and Table 3.5 where it participates in all top 5 strategy combinations as well as some other proposed strategies. The performance of each strategy was analyzed for different word length; Figure 3.10 and Figure 3.9 show that for all word lengths the eleventh strategy performs best both for words and phonemes. Further improvements for the already mentioned state of the algorithms were achieved when these baseline results were enhanced by means of applying a set of transformation rules learned from errors. The transformation rules were learned automatically from a training corpus previously labeled using four different classifiers. The rule templates are language-independent and can be used to generate transformation rules for any language. The combination of any method with transformation-based error-driven algorithms significantly improved the results obtained by that method alone. The best G2P results were obtained for the combination of FST with TBL algorithm. The error-transformation rules were also trained and applied to a simple prediction obtained by assigning the most-likely phoneme to each letter based on letter-phoneme pairs seen in the lexicon.The results obtained proved the effectiveness of the transformations rules. In fact, the results were higher than those obtained by the widely used DT and HMM and closely comparable to those obtained by FST and PbA before the application of TBL. The transformation-based learning algorithm was also applied to improve the prediction the PbA and the results were analyzed. New strategy combination methods were considered and slight improvements attained. The fact that application of error-corretion rules did not give significant improvements allows concluding that the PbA method is quite capable of capturing the regularities in English orthography.

The pronunciation was also derived for other languages. For Spanish, the obtained results were high as expected, since for languages with shallow orthography the pronunciation of common names can be as easily inferred by a small set of simple rules as by MLR techniques. In the case of proper names and neologisms the simple rules might have difficulty to predict the pronunciation. The worst results were obtained for Slovenian and German.

The overall conclusion that can be drawn from the experimental results detailed in this chapter is that pronunciation by analogy algorithm including the new strategies gives the best results for G2P conversion for all English lexicons, although it is closely followed by the performance of the finite-state transducers enhanced by the transformation-based learning

algorithm. The size of the training corpus as well as the method used to align the training data have major influence on the system performance.

Furthermore, we analyzed different factors that could influence error rates in grapheme-to-phoneme conversion. The error distributions obtained for the LC-STAR corpus Figure 3.14 indicate that mainly 9-letter long words contribute to the total error rate, if the optimal model parameters are chosen for training of the system.

Moreover, the probability of grapheme-to-phoneme errors was analyzed analytically with regard to the word frequency in a given corpus. It was shown that knowledge of the word frequency for a given text and the distribution function of probability error versus the word length allows us to choose the best suitable G2P method and therefore reduce the error rate, although it seems to be independent from the word length, in practice. The number of unknown words in general cases is greatly inferior to that in the lexicon test set. The best word accuracies obtained for different lexicons seem to be high enough to obtain high quality synthetic speech, especially considering the fact that the system dictionary allows to reduce the number of unknown words that the system receives at the input.

The case of connected speech was also considered in the framework of Blizzard 2007 and 2008 challenges. Phonetic weak forms were introduced to our TTS synthesizer in order to find the best match between speaker's dialect and vocalization style. A more precise phonetic transcription positively influences the resulting synthesis quality.

Chapter 4

Dictionary fusion

4.1 Introduction

Although the error rate in G2P conversion is rather low, these errors still affect the quality leaving room for improvements.

As it was already mentioned in Chapter 2, phonetic transcription module of a TTS system is based on a system dictionary and automatically trained G2P converter (or converters, one for each language, in case of a multilingual system) used to derive of the pronunciation of the unknown words fed to the phonetic module of a TTS system. The average word accuracy of grapheme-to-phoneme conversion for out-of-dictionary words, achieved by language-independent G2P conversion methods for languages like English usually falls between 67 and 89 percentage points, depending on the method itself and the lexicon used for the evaluation. An error rate ranging from 33 to 11 percentage points could decrease significantly the intelligibility of the synthesized speech in some cases.

We can reduce the error rate by increasing the coverage of the system dictionary.

For English many lexica are available for the task of G2P conversion. For instance, in Chapter 3 we use NETtalk dictionary containing 20K common nouns; LC-STAR lexicon containing about 50K common and 50K proper nouns; CMU lexicon, containing both common and proper nouns, about 125K in the total, and finally, publicly available Unisyn lexicon containing around 110K of both common and proper nouns. All these lexica provide a version for American English, although a variety of different phonesets were used for their transcription.

If we set increasing lexicon coverage as our goal in order to improve the performance of our speech synthesizer we should consider merging the available lexica as the solution to this problem. Merging would reduce the percentage of never-before-seen words by our speech

synthesizer through expanding of dictionary coverage. For example, adding 40K words to 80K system dictionary would increase the lexicon size by 50%.

Although merging seems to be a quick and efficient solution to increase the quality of pronunciations obtained by the prosody module as the output of the phonetic module of our TTS, merging cannot be carried out directly. Lexica are often transcribed not only using different phonetic alphabets, but also using different pronunciation criteria (articulated vs. relaxed). The phoneset size varies across phonetic alphabets as well.

To solve the problem of phoneset compatibility at least one expert needs to define the mapping between phonesets before merging the data.

Such inconsistencies occur because the lexica are created by different expert organizations and for different purposes. The differences lie in use of particular transcription criteria as well as in and precision degrees (including or excluding allophones); all these result in a number of inconsistencies for a given word in a given dialect.

However, it would be desirable to standardize these criteria. Even if the perfect standardization might not be reached, these pronunciations could be considered valid. Discrepancies with other lexica, especially for very long words (see Figure 3.9) , seem to be less *severe* than errors found in pronunciations derived by data-driven methods. Usually alternative pronunciations found in other lexica do not contain errors in consonant clusters leading to severe intelligibility errors, but only some minor vowel confusions which do not produce highly unpleasant synthesis artifacts. For example, pronunciations derived using DT often have excessive consonant agglomerations. Having said all of the above, merging seems quite advantageous independently from the pronunciation criteria used to transcribe them, as the the pronunciations from other lexica are generally more reliable than any transcriptions not validated previously. The classification of errors will be explained towards the end of this chapter.

In order to address the reliability issue we would need an automatic method in order to bring uniformity to both phoneset and criteria used for transcription of those.

In Chen (2003) it was reported that direct merging of lexicons transcribed using different phonesets worked rather poorly. Grapheme-to-phoneme conversion techniques (joint ME n-gram model) were applied to decrease lexicon merging error rates. Some mappings between very different phonesets, however, were spelling-dependent and could not have been disambiguated without the spelling information. In order to deal with this case the authors extended their conditional ME model in order to include the letter information. A 3-way alignment was used to align the letter sequence to the corresponding phoneme sequences. The resulting mapping were evaluated against common examples from the target lexicon. The data-driven models significantly outperformed manual mapping rules and the letter

information allowed slight improvements. Merged lexicons were demonstrated effective for G2P model training.

The rest of the chapter is organized as follows: first in 4.2 the lexica available for merging are described, then the results and issues of direct dictionary merging is described in 4.3.1. Further along, an automatic fusion method based on data-driven algorithms is proposed in **??** and the objective results are given in 4.4. In the end of the chapter the overall conclusions are drawn.

4.2 Analysis of the available lexica

As it was already mentioned in Chapter 2 and Chapter 3 some of the most popular publicly available lexicons for the evaluation of the grapheme-to-phoneme methods are: NETtalk pronunciation dictionary (Sejnowski and Rosenberg, 1993), CMU dictionary (Weide, 1998) and Unisyn dictionary (Fitt, 2000). The CMU dictionary, provided by Carnegie Mellon University, includes about 125K North American words which were generated using independent sources of proper and common names among which were expert proofed transcriptions as well as some synthesizer-generated ones. For this work, the phonetic transcription had to be converted to Sampa (Wells, 1997). The Unisyn dictionary, provided by the University of Edinburgh consists of 110K word entries, common and proper English names. The great advantage of this dictionary is that it is transcribed in metaphonemes which allows the encoding of multiple accents of English (e.g. UK, US, Australian, etc.). Output is available in Sampa (Wells, 1997) or IPA (Handbook, 1999) phonetic alphabets. The LC-STAR dictionary (Hartikainen et al., 2003) includes 50K common words and 50K proper General American names. Each proper name is assigned with a label that indicates whether it is a geographic, person's or company name. The NETtalk (Sejnowski and Rosenberg, 1993) is publicly available dictionary of 20,008 words that were manually aligned by Sejnowski and Rosenberg for their experiments with neural networks (Sejnowski and Rosenberg, 1987).

Terms *target lexicon* will be used to describe the main system dictionary and *source lexicon* for the auxiliary lexicon that will be used to extend the target lexicon.

The target lexicon in all cases was set to be the LC-STAR lexicon. LC-STAR dictionary was chosen as the target dictionary because of its considerable size, reliability and presence of a large number of proper names. Besides, it has been previously used as the system lexicon for our TTS system Ogmios. The source lexica in this case were: CMU and Unisyn.

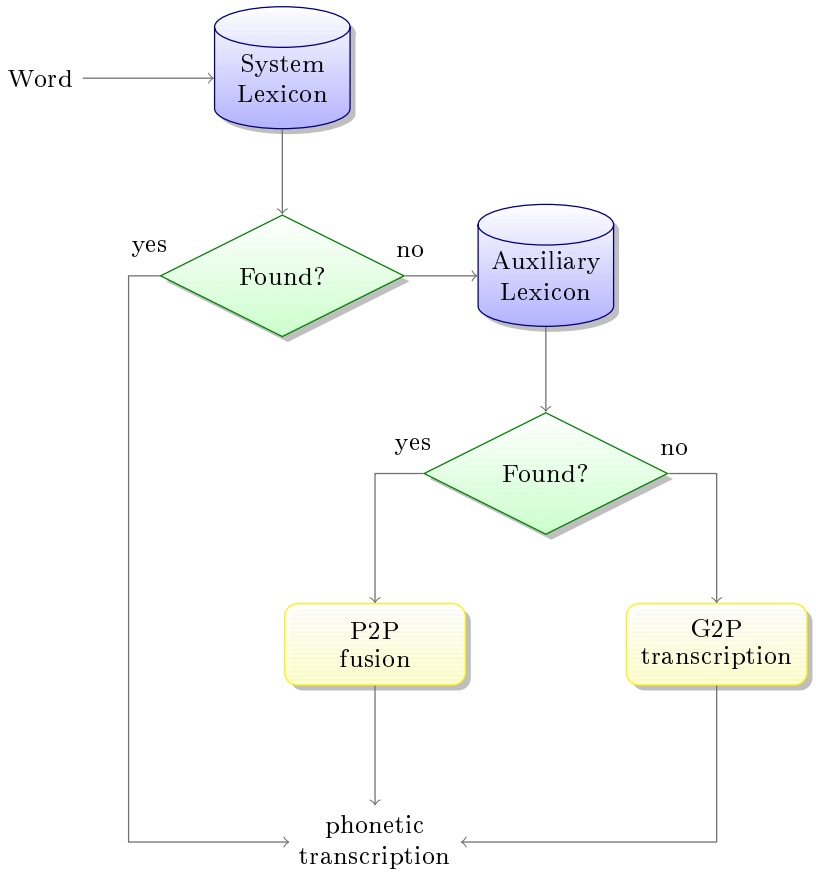


Figure 4.1: Scheme of the phonetic module in our TTS system.

4.3 Fusion of lexica in TTS

The goal of dictionary fusion is to be able to automatically update lexica with new entries from auxiliary dictionaries to improve the G2P results and therefore the overall speech quality. In this section, we apply the data-driven methods proposed in Chapter 3 to approach the G2P task to the P2P task of adapting the pronunciation of words found in auxiliary lexicon. This would improve the reliability and intelligibility of the resulting pronunciations.

The fusion diagram is illustrated in Figure 4.1. The TTS phonetic module including fusion first searches the system dictionary for the input word. If it is found, the tool outputs the corresponding pronunciation. If the input word is not found in the system dictionary, it is then searched in the auxiliary dictionary and in case it is found there, the fusion is performed. The transcription is obtained from the auxiliary lexicon and then passed to the P2P fusion module, responsible for adapting the pronunciation in question to the system

lexicon format and pronunciation criteria. In the last case, if the word is not found in any of the available lexicons, the transcription of the word is obtained by feeding its graphemic string to the G2P module.

The first proposal consist on applying some hand-derived rules so that the phonesets and the transcription criteria of both lexica are more consistent. However, we will see that the inconsistencies remain and more elaborated rules are required. Then, we propose to apply data-driven derived rules to transform the phoneme sequence of the auxiliary lexicon in a phoneme sequence consistent with the system lexicon. This is represented in the diagram by the P2P fusion block.

4.3.1 Results with almost direct merging

In this section we first consider direct dictionary merging with phoneset adaptation and then we propose to apply dictionary fusion algorithm based on P2P transformations.

First of all, almost direct dictionary merging was attempted to obtain baseline results. To validate the "direct" dictionary merging method previously used in TTS evaluations, words that appeared in both dictionaries of each source-target pair were compared. This comparison was performed directly, by selecting all common word entries for each pair of the lexica and counting the errors. CMU contained both proper names and common nouns, and, therefore, it had common entries with LC-STAR lexicon of common names and also LC-STAR lexicon of proper names. However, CMU dictionary was mapped to SAMPA phonetic alphabet using the mapping in Table Table 4.1.

For the Unisyn lexicon some phonetic alphabet adaptation was also necessary. For example, the rhotic `/@r r/` and `/3r r/` in the Unisyn lexicon were replaced by `/@r/` and `/3r/` according to the LC-STAR transcription system. Some examples of differences between dictionaries are shown in Table Table 4.2.

A total of 40,129 same word entries were found for common words from the CMU and LC-STAR dictionary pair. The CMU dictionary of proper names had less common entries with the LC-STAR dictionary of proper names: a total of 24,232 words appeared in both lexica. Unisyn and LC-STAR (common words) shared 43,606 common word entries.

Table 4.3 shows the word and phoneme accuracy for direct dictionary merging (CMU being the source and LC-STAR being the target lexicon) for both with and without stress marks. Because of a large number of proper names contained in the CMU dictionary, we were able to analyze proper names separately from the common words. This table shows that the compatibility between common names transcription in different lexica is about 12-13 percentage points higher than for proper names. There is no significant difference in stress placement. Table 4.4 shows the results for direct merging of Unisyn lexicon with

Vowels

CMU	SAMPA
AA	A
AE	{
AH	V
AO	O
AW	aU
AY	aI
EH	E
ER0	@r
ER1	3r
EY	e
IH	I
IY	i
OW	o
OY	OI
UH	U
UW	u

Consonants (I)

CMU	SAMPA
B	b
CH	tS
D	d
DH	D
F	f
G	g
HH	h
JH	dZ
K	k
L	l
M	m
N	n

Consonants (II)

CMU	SAMPA
NG	N
P	p
R	r
S	s
SH	S
T	t
TH	T
V	v
W	w
Y	j
Z	z
ZH	Z

Table 4.1: CMU to SAMPA mapping for vowels and consonants

word	LC-STAR	Unisyn
commuted	`/k @ m j u t @ d/`	`/k @ m j u t I d/`
jalapeno	`/h ae l @ p i n j o/`	`/h ae l @ p i n o/`
Taormina	`/ t A O r m i n @/`	`/t aU @r m i n @/`
McGary	`/m @ g E r i/`	`/m @ k g E r i/`
cure	`/ k j 3r/`	`/k j u r/`

Table 4.2: Differences in pronunciation between system and auxiliary dictionaries.

	Common Words		Proper Words	
stress	yes	no	yes	no
phonemes	93.37	94.17	86.95	88.03
words	69.10	71.06	57.23	58.28

Table 4.3: Phoneme and word coincidence (%) between CMU and LC-STAR dictionaries (common words and proper names).

LC-STAR lexicon. Only common words were analyzed. For this pair of dictionaries the compatibility rate is rather low. Only 31% of words were transcribed equally. Again there

stress	yes	no
phonemes	84.14	84.86
words	31.44	32.02

Table 4.4: Phoneme and word coincidence between Unisyn and LC-STAR dictionaries (common words).

is no significant difference in stress placement.

From Tables Table 4.3 and Table 4.4 we can observe a much higher level of similarity between the common words from CMU and LC-STAR lexicons, than between common words from Unisyn and LC-STAR. CMU seems to be rather similar to the target lexicon chosen for our framework, although it is still far from reaching 100% compatibility. The pronunciation of proper names seems to be quite different between the lexica. The Unisyn dictionary seems unsuitable for merging with LC-STAR lexicon without performing the pronunciation adaptation. The difference between the stress patterns for considered lexicon pairs is neglectable, but seems to be greater for proper names. From now on, only the results obtained with unstressed lexicons are considered.

4.3.2 Fusion method using P2P techniques

P2P methods allow to convert one phonetic form to another using data-driven techniques. When a word is found in auxiliary dictionary, its phonetic form is fed to the previously trained data-driven P2P converter in order to obtain the phonetic form consistent with the system lexicon. Most automatic G2P and P2P converters require prior alignment, so in order to train our phoneme-to-phoneme fusion algorithm we need a training corpus consisting on the left of phonetic forms found in the source dictionary and on the right of phonetic forms found in the target dictionary, for the words that are present in both of these. Then the alignment is learned between the source and target phoneme strings. Given that the length difference between source and target strings are unpredictable, it was necessary to have an alignment method able to introduce the nulls both into source and target strings. That is why the dynamic programming algorithm based alignment method similar to (Damper et al., 2004) was implemented. Since there is no way to predict where to place nulls in the source string, these were removed and the phonemes corresponding to the null letters were joined by an underscore with the previous phoneme, e.g.:

```
b o x _ / b A k s      →      b o x / b A k_s.
```

Once the alignment is obtained, a P2P converter is trained to perform the fusion. A similar, but much smaller test corpus is used to assess the performance of the algorithm.

The training and test corpora contain 90% and 10% of common words between source and target dictionaries correspondingly.

Two machine learning techniques such as decision trees (Black et al., 1998b) and finite-state transducers (Galescu and Allen, 2001) were used for training both the automatic G2P and P2P conversion systems. For the G2P case, in order to have more consistency in the alignments, the list of prohibited alignments was used (Black et al., 1998b). No vowel-consonant, consonant-vowel alignments were allowed. We assume that the fusion task is simpler than general G2P task. We consider that resuts obtained by one of the best G2P classifiers, FST, would validate the fusion idea. It was also interesting to compare FST results with those of a less powerful classifier such as DT. The detailed description of the DT and FST machine-learning methods are given in Chapter 3.

4.4 Fusion results

For the evaluation of the P2P converter we considered the common entries from each pair of dictionaries as in Table 4.3 and Table 4.4. The training set for each experiment consisted in 90% percent of the common entries from each pair of dictionaries and the test set of 10% accordingly. The fusion results for 3 different pairs of lexicon are given below. Table 4.5 shows the results for the conversion of common words and proper names from CMU to LC-STAR format, Table 4.6 from Unisyn to LC-STAR respectively.

lexicons	CMU → LCSTAR common		CMU → LC-STAR proper	
P2P classifier	DT	FST	DT	FST
phoneme acc.,%	96.80	97.26	88.91	88.40
word acc.,%	83.70	84.72	58.30	59.26

Table 4.5: Common words and proper names conversion CMU to LC-STAR dictionary

lexicons	Unisyn → LC-STAR common	
P2P classifier	DT	FST
phoneme acc.,%	96.36	96.72
word acc.,%	79.92	83.02

Table 4.6: Common words conversion results for Unisyn to LC-STAR dictionary

Table 4.5 gives phoneme and word accuracy after adapting pronunciation from the source lexicon CMU to the criteria of target lexicon LC-STAR. Word accuracy for common words obtained by both methods, DT and FST, is around 84%. Direct merging gave only 71%

lexicon	LC-STAR common		LC-STAR proper	
G2P method	DT	FST	DT	FST
phoneme acc.,%	94.22	96.11	87.34	89.91
word acc.,%	70.09	81.58	53.10	65.42

Table 4.7: G2P results for common and proper names for LC-STAR dictionary

(see Table Table 4.3). For proper names the results are not so remarkable. Table 4.6 gives the same results for Unisyn lexicon. The word accuracy for the FST fusion is around 83%. However in this case, the difference between word accuracy obtained by direct merging and dictionary fusion is even more remarkable because the word accuracy for the direct merging was ony about 32% for this lexicon (see Table Table 4.4). We can conclude that the fusion is able to reduce the number of inconsistencies existing between dictionaries, especially for common names, where the post-fusion word accuracy improvements range about 13% for the CMU dictionary and about 50% for the Unisyn dictionary. After the fusion the latter one reaches a similar level of compatibility with the reference that CMU. Nevertheless, there are still some different transcription criteria that the automatic methods were not able to capture.

To justify the importance of the dictionary fusion the G2P conversion was performed both for common and proper names form the LC-STAR dictionary. The G2P conversion results are shown in Table 4.7. The test set was the same as for the P2P conversion for CMU to LC-STAR scheme, which is 10% percent of the common words between those dictionaries, while the training set in each of the cases included all the remaining words in the corresponding dictionary.

The P2P accuracy for the fusion of CMU and LC-STAR dictionaries is higher than the G2P accuracy for the LC-STAR dictionary, therefore allowing for a conclusion to be drawn that including dictionary fusion algorithm as a part of phonetic module of the TTS could improve the speech quality.

Furthermore, we believed that G2P conversion errors could be more severe than differences between dictionaries, and for this reason, 50 erroneous common words for each of the 3 conversion schemes (CMU $\rightarrow$ LC-STAR , Unisyn $\rightarrow$ LC-STAR and G2P) were analyzed and "quickly" classified into three categories, according to the criteria from Table 4.8. The 3 categories were: L, M and S. The *"light"* errors "L", do not difficult the comprehension of a word, and may even pass unnoticed to a non-native English speaker. The *"medium"* errors, "M", make the word less recognizable but not unpleasant to hear, while *"severe"* errors or "S" can severely alter the intelligibility of a synthesized word, and may be the cause of very unpleasant acoustic artefacts.

Error category	Classification criteria	Example
Light (L)	shwa substituted by other short vowel	@ → I
	flap /4/ substituted by /t/ or vice versa	4 → t, t → 4
	short vowel substituted by a similar long one and vice versa	I → i, i → I
	a short vowel substituted by shwa	A → @, I → @, E → @
	consonant-consonant confusion (same articulation point)	s → z, t → d
	missing shwa	
Medium (M)	two or three errors of type "L"	
	missing consonant	
	affricate-fricative, fricative-plosive, confusions ,etc.	dZ → z, Z → z
	diphthong-vowel confusions	ae → aI , aI → i
Severe (S)	two or more errors of type "M"	
	vowel-consonant confusions	3r → r, A → r, V → v
	more than three errors of type "L"	

Table 4.8: Error classification, criteria and examples.

conversion type/ errors	Light	Medium	Severe
P2P (CMU → LC-STAR)	38	10	2
P2P (Unisyn → LC-STAR)	39	8	3
G2P	28	15	7

Table 4.9: Count of erroneous examples for each conversion method and category.

From Table 4.9 we can observe that in the case of common names by fusing the dictionaries we do not only obtain a higher overall pronunciation accuracy but also reduce the number of severe and moderate errors that really worsen the speech quality.

4.5 Conclusions

After carrying out all of the above described experiments, several conclusions can be made. The experiments confirm the existence of significant number of inconsistencies between different dictionaries; some of them appear due to the difference in transcription criteria employed by the experts while others are caused by the inconsistencies already existing in the dictionaries which, in its turn could be a result of using various unhomogenized sources for building the dictionary (as in the case if CMU dictionary). Some of these differences can be overcome by dictionary fusion procedure which consists deriving automatic P2P conversion rules from the words that the dictionaries have in common. In the case of the common words form LC-STAR dictionary it seems to be feasible to fuse them only with those CMU dictionary since this procedure gives the best results. For proper names the fusion does not give significant improvements due to the elevated difficulty of the proper names transcription problem even for human experts. The DT and FST methods give similar results in the fusion task. After performing fusion the number of severe transcription errors decreases, therefore guaranteeing a better quality of synthesized speech.

Chapter 5

Multilingual speech synthesis

In the modern world, speech technologies must be flexible and adaptable to any framework. Mass media globalization introduces multilingualism as a challenge for the most popular speech applications such as text-to-speech synthesis and automatic speech recognition. Mixed-language texts vary in their nature and when processed, some essential characteristics must be considered. In Spain and other Spanish-speaking countries, the use of Anglicisms and other words of foreign origin is constantly growing. A particularity of peninsular Spanish is that there is a tendency to nativize the pronunciation of non-Spanish words so that they fit properly into Spanish phonetic patterns. UPC participated in a project named "AVIVAVOZ". The main objective of this project was to create a speech-to-speech translation system capable of performing translation between four official Spanish languages (Catalan, Galician, Basque and Spanish). The phonetic alphabets of each one of them are different. The UPC was responsible for the translations from Basque, Galician and Spanish to Catalan. Thus, the main goal for TTS developers was to give the system the ability of reading correctly multilingual texts. For this purpose, a language identification system was developed, language-specific automatic G2P methods were implemented and a nativization system was proposed.

With the goal of improvement of pronunciation of "foreign" word and proper names of "foreign" origin the *nativization* was proposed. The term "foreign" is used to refer to the words that are not native to a particular language, with no reference to the country of origin.

Spain is a country of a remarkable linguistic patrimony, which is a cultural treasure but also represents an additional challenge in terms of speech technologies. In the framework of the rapidly expanding field of applications, speech tools must be adapted to the multilingual scope allowing a higher level of flexibility and answering the needs of modern users. Currently in Spain, it has become quite commonplace to hear proper names from all over

the world. Text-to-speech synthesis finds many important applications on the emerging market of speech technologies. Voices that can embrace more than one language are highly demanded in the era of mass media globalization. We wish to assign the term *nativization* to the pronunciation adaptation process.

The existence of several large bilingual regions in Spain make the problem more crucial. The autonomic languages have the same rights and are used with the same naturalness as Spanish. Focusing on the particular case of Catalonia, it could be said that fully bilingual conversations are rather commonplace in any social environment. Besides, written texts in Spanish are full of Catalan proper names and vice versa. Furthermore, the use of Anglicisms is also rapidly increasing. Multilingualism, being such a frequent phenomenon in mass media, social networks and other areas tightly related to communications, undoubtedly deserves special attention.

To maintain an up-to-date synthesizer, we need an ultimate automatic method for the derivation of the nativized pronunciation. The final goal of nativization is to be able to produce highly intelligible synthesized speech that would be well accepted by native speakers of the target language, those with some knowledge of the source language, as well fluent source language speakers. In the framework of this thesis, both English and Catalan were considered as source languages in different experiments, and the two target languages considered were Spanish and Catalan. For clarity of definitions, we will use the term *source language* to define the language of origin of a foreign word and the term *target language* to indicate the language to which that foreign word should be adapted.

We encounter numerous multilingual contexts in our daily life, if we take a look, for example, at two consecutive comments from the same news discussion forum of a popular local newspaper "Avui " the first comment maybe partially in English, partially in Catalan, while the second one could be entirely in Spanish. Email messages in Catalonia are usually at least bilingual. Any official email message circulating in our department contains at least two languages, Catalan and Spanish, English is often included for the international students and staff. Proper names form different languages very frequently appear together in newspaper articles talking about European Parliament sessions, international trade, etc.

There are, of course, many more examples of texts where the correct identification of the language is necessary to improve the pronunciation. Popular global social networks, such as Facebook, Twitter, Google +, blogs, forums, etc. represent an endless source of multilingual entries. Whether we consider two random or consecutive entries, they are unlikely to be written in the same language. A fully adaptable synthesizer must be able to meet the constantly growing needs of a global technological village.

This chapter focuses on multilingual scope, so we consider two different tasks, the task of language identification and that of nativization. Firstly, in Section 5.2 a baseline language

identification system is proposed and the results for four languages are obtained. Next, a baseline nativization system is used to enhance pronunciations and the results are obtained. Further along, were assume that the language of origin is already known and we focus strictly on improving the nativization. In Section 5.3.1 the differences between English and Spanish phonetic systems are explained. In Section 5.3.2 we describe the corpora creation for training and evaluation of the proposed automatic nativization system for English, which is described in Section 5.4. Section 5.5 gives a detailed analysis of the experimental results. Section 5.6 explains the nuances of nativization for Catalan and gives the analysis of the results obtained. A perceptual evaluation allows to draw conclusions on whether or not the nativized pronunciations produced are acceptable in the framework of this text.

5.1 Multilingual grapheme-to-phoneme system

In this chapter, we propose a baseline system to adapt the pronunciations of foreign words to the target language based on phoneme-to-phoneme nativization tables. After obtaining some preliminary, but quite promising results, a more sophisticated nativization system and manually crafted nativization corpora for two language pairs (English source and Spanish target; Catalan source and Spanish target) is proposed and evaluated. Both phonetic and orthographic information was included for better picture. For more accurate multilingual grapheme-to-phoneme conversion knowing the language of *each word* in the text can be rather helpful for quite obvious reasons. However, it is also important to have a tool capable of efficiently determining the language of the paragraph in mixed-language texts extracted from newspapers, online forums or social media, emails, scientific articles, technical support manuals, web pages, and other sources where the language can suddenly change from one paragraph to another.

Our multilingual G2P conversion system is configured to determine the language of the paragraph and then of each isolated word in that paragraph. By defining correctly the source and the target languages for nativization the synthesis quality can be improved considerably.

Some results on identification of the language of the paragraph and isolated words are reported in 5.2.1. Figure 5.1 shows our nativization system, used to adapt the pronunciation of foreign words that had been previously assigned a special label F_LANG, to the default language of the system - LANG or the target language.

Our pronunciation module consists of system lexica in several languages and corresponding language-specific grapheme-to-phoneme The first step is to determine whether the word in question is found in the system dictionary of the target language. If this occurs, the encountered pronunciation is validated. It is important to emphasize

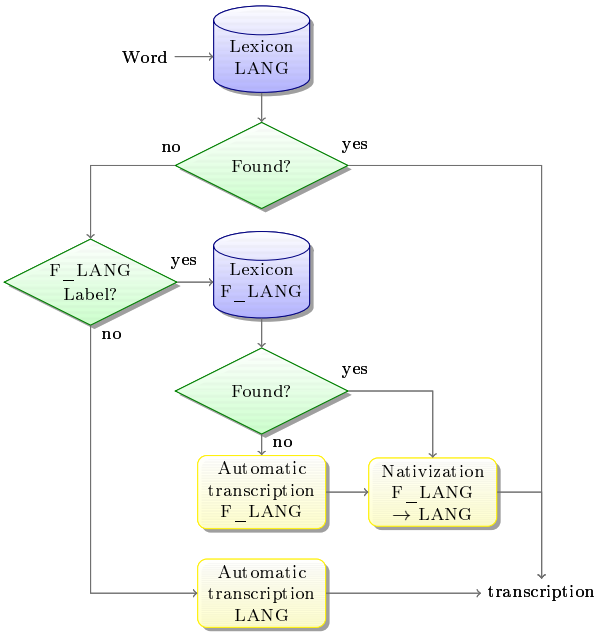


Figure 5.1: Scheme of a multilingual G2P system.

that if a foreign word is found in the target language dictionary, we consider it already to be nativized. Usually, dictionaries include the most common proper names. That is why there is no need to check the language before the first step. If the word is not in the target language dictionary, the next step is to determine if its language of origin is different from the target language (does it have an F_LANG label?). If no such label has been found, the pronunciation is derived using the automatic transcription system for the target language. For the words identified as foreign, the search continues in the corresponding source language dictionary. Before validating the pronunciation, if it is found in the dictionary, the nativization phoneme-to-phoneme converter is applied to the source. The output of the nativization module is the nativized pronunciation adapted to the target language. In the last case, if the word is also absent from the source language dictionary, its pronunciation is derived using the automatic transcription system for the source language, after which nativization is applied before validating the pronunciation.

5.2 First approach to multilingual grapheme-to-phoneme conversion

In this section, we introduce a baseline language identification system. The multilingual grapheme-to-phoneme system illustrated in Figure 5.1 obtains language information from

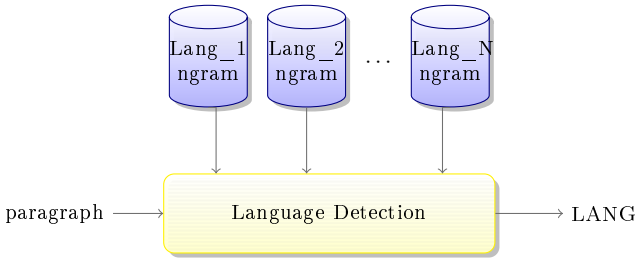


Figure 5.2: N-gram based language identifier

LANG and F_LANG labels. We can apply a n-gram-based language identifier to obtain these labels.

5.2.1 Language identification

The reasons why language identification is essential are the following: 1) improvement of the pronunciation of the foreign words and 2) adaptation of the pronunciation to the language of the paragraph. News reports, popular international sporting events, etc., mainly contain proper names of foreign origin, while forums, social networks, emails, SMS, have so much language mixture that, even for a human, it is difficult to identify the language, since it can change suddenly and without any reason at all.

As the language identifier we took the standard n-gram model, also previously used for the same task (Font Llitjos and Black, 2001), see Figure 5.2.

An n-gram is estimated for each language of interest. The n-grams include word beginning <s> and end </s> markers: <s>mi</s><s>*casa*</s>. Then the probability is calculated for each paragraph:

$$Lng^* = \underset{Lng}{\operatorname{argmax}}\, p(l_1 \dots l_n | Lng) \tag{5.1}$$

5.2.2 Nativization Tables (NatTAB)

Our first hypothesis was to approach nativization by creating nativization tables for each pair of source/target languages as a form of a preliminary study. The first step towards nativization was to choose the basic sound units or the phoneset to be used in the target language. For the AVIVAVOZ project(focused on four official languages spoken in Spain), only two target languages we chosen: Spanish and Catalan. Such a choice was due to the fact that in our group at the UPC there were no fluent speakers of Basque or Galician, the other two languages involved in the project (Polyákova and Bonafonte, 2008b).

For the case of Catalan as the target language, we defined a new enriched Catalan phoneset by adding such Spanish phonemes as /x/ and /T/ as well as the unstressed vowels /a/, /o/ and /e/ (all of which are nonexistent in Catalan) to the existing Catalan phoneset. The new phoneset was called *Catalan+*. In bilingual societies the use of co-called enriched phonetic alphabets is not surprising, in fact, they are quite commonplace. Bilinguals have a very similar fluency in both languages , "Spanglish", a rather well known mixture of Spanish and English frequently used in the United States could be taken as an example of such phenomenon. For Catalonia, choosing the Catalan+ phoneset seemed optimal because it has a larger variety of phonemes than Spanish.

For the case when Spanish was ser as the target language, however, the standard Spanish phoneset was used. This choice is due to the fact that professional speakers, whose voices were used in the database recording, were not bilingual and their phoneset was strictly limited to Spanish phonemes. In case of ambiguities, such as when the source pronunciation contained an /ə/, that does not exist in Spanish, the target language G2P system was triggered and the phoneme suggested by this system was chosen. For the word *talent* /ˈt æ l ə n t/, the table suggests that /ə/ should be nativized to a Spanish phoneme /a/, while the Spanish G2P system gives an /e/ for that position. Therefore, the resulting nativized pronunciation has an /e/ in the 4^{th} position.

5.2.3 Evaluation of the baseline system

The experiments were carried out for the language identification of the paragraph and of isolated proper names. For perceptual evaluation of the nativized transcription, the unit selection synthesizer Ogmios was used (Bonafonte et al., 2007). Several utterances were synthesized before and after the application of nativization rules. Although at this stage the resulting quality was evaluated by volunteers without much experience in speech synthesis, experimental guidelines instructed the participants to neglect possible TTS artefacts in order to give the most unbiased opinion.

In language identification parts of the experiment, the following five languages were considered: Catalan, Spanish, Basque, Galician and English. In the framework of the AVIVAVOZ project the statistical machine translation bilingual corpora were available for Catalan, Basque, Galician, and Spanish. The sentences and their translations were taken from the Eroski consumer's magazine. It included mainly common words. For the experiments with the proper names, we created a corpus of Galician names and surnames with 3613 words in it (provided by the University of Vigo), a corpus of Basque names and surnames of about 11,200 words (Real Academia Vasca), and took the existing set of proper names labeled as person's names from the LC-STAR dictionary (Hartikainen et al., 2003) available for Catalan, Spanish and American English, all counting about 20-27K words.

"Foreign" words were counted for different sources and the resulting statistics obtained as given below.

1. Corpus Eroski Catalan (2*500 words) 2,2% English words, 1,4% words from other languages 0,02% Spanish words; the majority of "foreign" words are proper names.
2. Corpus Eroski Spanish (2*500 words) 3,3% English words, 1,2% words from other languages 0,03% Catalan words; the majority of "foreign" words are proper names.
3. 20minutos.es entertainment and leisure (5*200 words) 14% English words, 42% of which are proper names.
4. Avui.cat entertainment and leisure (5*200 words) 10% p. English words, 40% of which are proper names.

For Catalan and Spanish, the language models were trained on 90% of the Eroski corpus and 10% was saved for the test. The language model for English was trained using the common words from the LC-STAR dictionary. The training and test percentages were the same for the proper names language models, only the unique proper names were considered(those that were native to the language of the dictionary where they appeared). The words that appeared only in one language helped to polish the models.

Table 5.1 is a confusion table for four Spanish autonomic languages, considered in the experiment and English.

Lang./Ident. lang.	ca	es	eu	ga	en
ca	95.2	1.2	2.1	1.5	-
es	1.9	90.9	3	0.1	-
eu	0.6	0.8	92.6	8	-
ga	1.9	0.8	2.1	89.2	-
en	0.6	0.3	0.3	1.2	-

Table 5.1: Language identification results for Eroski corpus.

The best results were obtained for Catalan and Basque and the worst for Galician, due to its similarity to Catalan and Spanish, that in their own turn share a lot of common words. The majority of errors are caused by the following factors:

1. Presence of foreign words in the corpus, e.g. Kodak, Kellogg's, Fuji, etc.
2. Some words or phrases can belong to more than one language, e.g "agua mineral natural".

3. Digits and abbreviations: 10g, rayos UVA, etc.

Table 5.2 shows the language identification results for proper names (200 proper names were randomly selected from each corpus).

Lang./Ident. lang.	ca	es	eu	ga	en
ca	77	33	5	25	9
es	46	98	12	29	11
eu	6	12	170	6	3
ga	37	24	6	121	9
en	31	33	7	19	168

Table 5.2: Results for language identification results of proper names using "polished" language models.

The best results were obtained for Basque followed up by English and Galician.

A preliminary perceptual evaluation was carried out in order to validate the nativization tables. For the evaluation purposes a special multilingual corpora was created from the available online sources (20minutos.es, Avui.cat, etc.). The intelligibility of speech synthesis in Catalan and Spanish using the language detection and nativization rules as shown in Figure 5.1 was evaluated by volunteers. In the first experimental setting Catalan was set as the main or the target language mixed with some Spanish and English inclusions. The second experimental setting had Spanish as the target language and the inclusions were in English and Catalan. Each coprus consisted of 1000 words, 10% of which were non-native inclusions. Therefore, the corpus in Spanish contained 50 English and 50 Catalan inclusions, Catalan, in its turn, had 50 Spanish and 50 English inclusions.

Since the listeners, who participated in the evaluation of system were fluent in Spanish, Catalan and English, but had no knowledge of Basque and Galician, the perceptual test was designed including only the first three languages. To assess the synthesis intelligibility in Catalan and Spanish according to the scheme in Figure 5.1, we randomly chose and synthesized 10 utterances, 5 per experimental setting. F_LANG tags were added by hand to the non-native words. The synthesized utterances were evaluated in terms of intelligibility and naturalness. A total of 10 listeners volunteered to take part in this preliminary evaluation. For Spanish utterances, 70% of the listeners considered the nativized utterances more natural and 62% more intelligible. Sixteen percent did not find any significant differences between utterances in terms of naturalness and 28% percent judged the utterances equally intelligible. Fourteen percent of the volunteers found the nativized utterances less natural and 12% less intelligible than the baseline synthesis. For 5 Catalan utterances the results were slightly worse. Out of the same ten listeners, 44% found the

nativized synthesis more natural than the baseline. Thirty-six percent thought that the naturalness did not vary between the nativized and baseline utterances, while other 20% thought that naturalness decreased. On the other hand, 26% of the listeners believed that the nativized speech was easier to comprehend, 48% did not find any difference, and 28% said that, in fact, the nativized utterances were more difficult to understand. The results for Spanish seems very encouraging, however, the results obtained for Catalan were less optimistic. Such a strong listeners' preference towards baseline synthesis in Catalan can be explained by the absence of certain diphones in Catalan speech database that caused the appearance of some unpleasant acoustic artefacts, a system's flaw independent from the phonetic transcription. Other errors could have been caused by imperfect alignments of the training dictionaries and/or by the fact that some nativization rules were not suitable for rare words.

Some examples of multilingual sentences used in the perceptual evaluation are given below.

1. The Return of the King, la tercera entrega de Lord of the Rings, filmada por Peter Jackson.

2. Los admiradores de Scarlett Johansson están de enhorabuena.

3. Tal com va passar amb el Mobile World Congress.

4. Josep Antoni Duran Lleida, s'ha reunit avui amb José Luis Rodríguez Zapatero.

Lang./Ident.	ca	es	eu	ga	en
ca	77	33	5	25	9
es	46	98	12	29	11
eu	6	12	170	6	3
ga	37	24	6	121	9
en	31	33	7	19	168

Table 5.3: Summary of the preliminary perceptual evaluation.

At this point, we can conclude that the language identification system works rather well for paragraphs and slightly worse for single words. A preliminary assessment indicates that knowing the language of the foreign words and using multilingual phonetic transcription module allows obtaining a considerable improvement in naturalness and intelligibility of the speech synthesized in Spanish and Catalan in the majority of the cases.

5.3 Further improvements of the nativization system

The rest of the chapter focuses on the case when Spanish is set as the target language. Both English and Catalan inclusions will be considered in Sections through After having concluded a preliminary study and the dimensions of the issue, our goal was to approach the nativization challenge by data-driven methods, so popular among speech researchers because they are flexible, transferable to other languages and do not drop in performance in comparison with explicit rules manually written by experts. Training and test corpora for nativization consisted of 1000 and 100 words respectively, and were crafted manually for each pair of source-target languages(EN-US>ES, EN-US>CAT). In the previous section 5.2.2 we used a table-based method for nativization of foreign words in Spanish that produced noticeable improvements in comparison with the results obtained by applying a Spanish G2P converter to pronounce English words directly. The nativization process can be compared to the task of G2P conversion for out-of-vocabulary words. Pronunciation by analogy, previously used in Marchand and Damper (2000) and Polyákova and Bonafonte (2009), proved to be one of the most efficient methods for G2P tasks. First, we are going to focus on English inclusions in Spanish, which seem to be more ambiguous. We believe that analogy between the nativized pronunciation and the original pronunciation can be inferred in an even more reliable and simple way because nativization of English words in Spanish is an easier task than finding the pronunciation of unknown English words. In fact, all human attempts to nativize foreign words depend on the analogy between known and unknown words. Very few databases containing non-native pronunciation are available,while the nativization corpora are simply non-existent.

Our nativization proposal is explained as follows: in Section 5.3.1 we explain the differences between English and Spanish phonetic systems. In Section 5.3.2 we describe the corpora creation for training and evaluation of the proposed automatic nativization system, described in detail in Section 5.4. Section 5.5 gives a detailed analysis of the experimental results.

5.3.1 Spanish phonetics vs. English phonetics

There are numerous phonetic differences between English and Spanish. We sought to examine consonants and vowels separately. The discrepancy between consonants and their orthographic representation in English is less significant than in the case of vowels.

Peninsular Spanish lacks English consonants such as /ʃ/, /v/ /ð/, /ʤ/, /ʒ/, /z/ and /ŋ/, and Latin Spanish also lacks the unvoiced /θ/. Bear et al. (2002), Yavas (2006) and Raynolds and Uhry (2009) reported that the most common substitutions for the missing consonant sounds in English by native Spanish speakers are: /θ/→/t/, /f/ (e.g., *thin/tin,*

bath/baf), /ð/→/d/ (e.g., *they/day, lather/ladder*), /v/→/b/ (e.g., *vote/boat*), /z/→/s/ (e.g., *zip/sip, prize/price*), /ʃ/→/ʧ/,/s/ (e.g., *shop/chop, wash/watch, she/see*), /ʤ/→/ʧ/ (e.g., *jeep/cheap*), and /ŋ/→/n/ (e.g., *hanged/hand, sung/sun*). In both English and Spanish phoneme repertoires, we find unvoiced stop consonants /p/, /t/, and /k/ and voiced /b/, /d/, and /g/. However, they have significant differences at the time of articulation. In English, voiced stop consonants /b/, /d/, and /g/ present loss of voicing during their production. In Spanish, however, /b/, /d/, and /g/ are fully voiced because voicing begins before the start of the vowel. In English, there is a small delay after unvoiced stop consonants /p/, /t/, and /k/ before the following vowel in stressed syllable-initial positions that is known as aspiration. Spanish stop consonants, on the contrary, are not aspirated. The phoneme /p/ in the Spanish word *pesos* sounds more like /b/ in the English word *basis* than /p/ in *paces* (Ladefoged, 2003), although this particular difference was not considered in this thesis. English has two different phonemes to represent the letters *b* and *v*, /b/ and /v/, respectively. Spanish also contains these letters, however, they are pronounced either with a bilabial approximant sound [β] or a stop /b/ that occurs at the beginning of the word. No labiodental /v/ is produced (Hammond, 2001). The English phoneme /ŋ/ finds its twin in the Spanish velar nasal allophone [ŋ] occurring before velar consonants in words or at word boundaries, e.g., *increíble* or *un gato*. English alveolar-voiced fricative /z/ also exists in Spanish only as an allophone [z] occurring at the end of a syllable before a voiced consonant, e.g., *abismo*, *desdén*. English dental fricative /ð/, is similar to Spanish dental approximant [ð] that occurs inside a word when it is not preceded by nasals /m/,/n/, lateral alveolar /l/ or a pause.

English and Spanish vowels are quite different. Spanish has 5 pure vowels while American English has 11 pure vowel sounds. Vowel transcription in English presents a special difficulty due to its deep orthography. For consonants, the length of the preceding vowel contains important information that helps to distinguish voiced consonants from unvoiced stop consonants at the end of a word; this is crucial for making distinctions between words. In Spanish, vowel length is not as variable and these small differences do not cause semantic changes (Fox et al., 1995). The list of Spanish and American English pure vowels is given in Table 5.4.

Native speakers of Spanish usually have trouble in perceiving and producing the variety of English vowels. For example, no distinctions are made between *ship/sheep* or *fool/full*. Besides, Spanish speakers tend to prefix English words beginning with *s*− consonant cluster with an /e/ sound, so that *school* becomes [e s ˈk u l]. Furthermore, some sound swallowing is typical when three or more consonants occur together, as in *next* [ˈn e k s] (Swan and Smith, 2001). These are the main observations that helped us to define the nativization criteria detailed in Section 5.3.2.

IPA symbol	Description	Example
/i/	close front	*pico* /ˈp i k o/
/e/	mid front-central	*pero* /ˈp e ɾ o/
/o/	mid back-central	*toro* /ˈt o ɾ o/
/u/	close back	*duro* /ˈd u ɾ o/
/a/	open central	*valle* /ˈb a ʎ e/

(a) Spanish (Conde, 2001)

IPA symbol	Description	Example
/i/	close front	*tree* /ˈt r i/
/ɪ/	near-close front	*rich* /ˈr ɪ ʧ/
/ei/	close-mid front	*cake* /ˈk ei k/
/ɛ/	open front	*bed* /ˈb ɛ d/
/æ/	near-open front	*had* /ˈh æ d/
/u/	close back	*lose* /ˈl u z/
/ʊ/	near-close back	*put* /ˈp ʊ t/
/oʊ/	close-mid back	*home* /ˈh oʊ m/
/ɔ/	open-mid back	*pause* /ˈp ɔ z/
/ʊ/	open-mid back	*cut* /ˈk ʊ t/
/ɑ/	near-open mid back	*dot* /ˈd ɑ t/

(b) American English (Wells, 1982)

Table 5.4: Pure vowels in Spanish and American English.

5.3.2 Database creation

In this section, we describe the nativization lexicon created for training and evaluation of nativization methods for English inclusions in Spanish. Rule-based approaches to phonemization require significant linguistic engineering, and they are always language-dependent, thus lacking flexibility. Data-driven approaches were proven to be more efficient than those based on the explicit linguistic modeling and they are undoubtedly superior in adaptability (van den Bosch and Daelemans, 1993). The main purpose of this work was to train a nativization model capable of converting English pronunciation to nativized Spanish. Data-driven techniques require training corpora, so a need for nativization training was apparent. For typical G2P conversion tasks, large pronunciation corpora of 100,000 words and their corresponding pronunciations are available. Since we did not find any existing nativization databases, we chose to create a minimalistic corpus that would not require hiring a highly qualified expert in linguistics.

Training data for English to Spanish nativization

For our task, due to the reduced sized of the training lexicon or *TrainingSet*, it was necessary to have it orthographically balanced. A greedy corpus-balancing tool was used for selecting words to be nativized from the available LC-STAR dictionary of American English (Hartikainen et al., 2003) with more than 50,000 entries, previously used by the authors in G2P conversion experiments (Polyákova and Bonafonte, 2009). To have all possible letter bi-grams in the corpus, we selected 1000 words. The original English transcriptions of these words were manually nativized according to the criteria described in 5.3.2. It is necessary to emphasize that the phoneme inventory used for nativization was limited to the Spanish phoneset including three allophones [ŋ], [ð] and [z]. The proportion of rare words in the resulting corpus was noticeable; however, a few non-English words were removed because their pronunciations did not obey English phonetics. Therefore, their presence in the nativization corpus could have introduced additional ambiguity. The TrainingSet consisted exclusively of common words. They were manually aligned during the nativization process.

Test data for English to Spanish nativization

To evaluate the nativization methods a test corpus was required. A specific test corpus was created in order to keep the full coverage of the TrainingSet. The test data was divided into two sets. The main one, named *CommonSet*, consisted of common words only. The words selected for CommonSet from the available online free daily newspaper *www.20minutos.es*, were rather frequently used common words. Such a choice was motivated by the fact that

the nativized pronunciation of the frequently used words is less ambiguous than that of the rare ones. In addition to the common words, it was interesting to evaluate the nativization algorithm on a set of frequently used people's names. Therefore a secondary evaluation set was defined. *ProperSet* contained people's names of English origin. The database entries for ProperSet were also collected from free online sources. None of the test words were present in the training lexicon. Both CommonSet and ProperSet contained 100 words each.

Nativization criteria and examples

In this thesis, we attempted to find a meeting point between a totally incorrect pronunciations of English words by Spanish speakers unfamiliar with English phonetics and almost correct pronunciations by those who are fluent in English. Since the goal of this project was to improve both naturalness and intelligibility of the synthesized speech, nativization was oriented at general Spanish-speaking auditory conventions. Nonetheless, the evaluation of the synthesized speech is a difficult task because its quality can only be defined by a listener and it varies from one listener to another (Black and Lenzo, 2004). With this goal TrainingSet, CommonSet and ProperSet were nativized using the criteria described further in this section. These criteria were based on the principles described in (Llorente and Díaz Salgado, 2004), however it was necessary to extend them to be able to transcribe the entire corpus and consider each case separately. Table 5.5 illustrates how some of the criteria were applied in particular cases. As it was already mentioned the non-English words were deleted from all sets since they could have hampered the generalization. In all cases the frequency of usage of a particular English word in Spanish was taken into account seeking better adaptation of its pronunciation to the language.

Absence of certain source language phonemes in the target language poses phonetic challenges for non-native speakers. For example, the English word *these* would be pronounced as ['d i: s] because the Spanish phoneset lacks the voiced dental fricative /ð/. When determining the best way to nativize a word, its level of assimilation to the target language plays a major role as well as the complexity of its orthography. Another question to be asked is "Is the word consonant-vowel pattern similar to that in the target language?" For instance, in Spanish, it is unnatural to have more than two consonants in a row at the beginning of the word. Additionally, Spanish does not typically allow more than three consonants sounds in a row in any position in the word, while Czech allows no-vowel words consisting of up to 4-5 consonants. There are even vowelless sentences such as "Strč prst skrz krk", where the nucleus of each syllable is a syllabic *r*, the phenomenon rather typical for Slavic languages. It was also important to ensure that no unusual consonant agglomerations in any of the word parts were encountered, even though sometimes it was inevitable due to the lack of vocalization in English. The case of two consonants *st−* at the beginning of the word particularly stands out because in Spanish a vowel is added before this consonant

cluster to smooth the agglomeration. The English [s t] is pronounced [e s t] in Spanish as it was already mentioned in Section 5.3.1.

The challenge of this task consisted of developing solid criteria for nativization, taking into account local specifications of certain words, pronunciation and word popularity factor, among others. Most certainly, it was found inappropriate to apply the same criteria to well-known words and to words with much lower occurrence rates. In the word *jazz*, the phoneme /ʤ/ was nativized to /jj/, while in *Egyptian*, the same phoneme was transformed to /ʧ/. In the word *logjam*, it was transcribed as [ð j], because the latter is a rare word and complete omission of the initial sound /d/ of the English phoneme /ʤ/ would cause important drawbacks in comprehension of the word (see Table 5.5 for more examples). The database nativization task was conducted using both source language orthographic form and pronunciation. In English to Spanish nativization, vowels were found much harder to transcribe systematically because their nativized pronunciation in Spanish is highly related to the word frequency-dependent English-to-Spanish orthographic analogy. Phonemes representing double sounds such as /ei/, /oʊ/, /aɪ/, /ɔɪ/ and /aʊ/ were transformed into corresponding double phonemes /e j/, /o w/, /a j/, /o j/, and /a u/. The stressed vowels were mapped to the closest match in the Spanish phone table, e.g., *agency* [e j ʤ ə n s i] to [ˈe ʧe n s i]. Most of the unstressed vowels and especially *schwa* /ə/ in the majority of cases were transcribed with a vowel closest to the letter as in *aimless* [ˈei m l ə s] to [ˈe j m l e s]. Additionally, we considered a specific extension of the Spanish phoneset. This decision was based on the hypothesis that conserving vowel length and word stress would contribute to the intelligibility of the nativized pronunciation. Thus, the /ɪ/ in *dip* was mapped to a short vowel [iˑ], /i/ in *deep* to a long [iː], /ɑ/ to a long vowel [aː], /ɚ/ to [eˑ r], /ɝ/ to [eː r], and /ə/ was mapped to the vowel corresponding to the letter but marked as short. For the consonants, as previously mentioned in Section 5.3.1, some difficulties were found when transcribing English /ʒ/ /ʤ/ and /ʃ/. The nasal /ŋ/, the voiced /ð/ and /z/ were conserved as they were present as allophones [ŋ],[ð], and [z] in our Spanish TTS system. The unvoiced /ʃ/, in most cases, was transcribed to /s/. The letter sequence *rr* corresponding to the Spanish vibrating phoneme /r/ in all nativized words was mapped to a Spanish alveolar tap /ɾ/ with reduced vibration, as well as /r/, usually corresponding to the letter *r* at the beginning of the word or after a pause (Llorente and Díaz Salgado, 2004). An illustrative review of the criteria used for nativization together with some exceptions is shown in Table 5.5. Some of the nativization rules were based on Canellada and Madsen (1987).

Word	Original pronunciation	Nativized pronunciation	Comment
airways	/ˈɛ r w e z/	[ˈe j ɾ w e j z]	In Spanish /e j r/ instead of [e r] is frequent.
basketball	/ˈb æ s k ə t b ɑ l/	[ˈb e s k eˑ t b oː l]	British pronunciation of *-ball* is widely used.
water	/ˈw ɑ t ɚ/	[ˈw oː t eˑ r]	/o/ in the 2nd position displays British tendency.
Egyptian	/ə ˈʤ ɪ p s ə n/	[eˑ ˈʧ iˑ p s j aˑ n]	/ʤ/ to /ʧ/ between vowels; /s j/ used to imitate /ʃ/.
comfortable	/ˈk ʊ m f t ɚ b ə l/	[ˈk a m f oˑ r t eˑ β oˑ l]	A short vowel inserted between 2 consonants.
dogfight	/ˈd ɑ g f aɪ t/	[ˈd oː ɣ f a j t]	British tendency for a frequent word part.
Aleutian	/ə ˈl j u ʃ ə n/	[a ˈl j u s j aˑ n]	/s j/ used to imitate /ʃ/.
awkward	/ˈɑ k w ɚ d/	[ˈoː k w eˑ ɾ ð]	British tendency observed for a frequent word.
bank's	/ˈb æ n k s/	[ˈb e n ɣ s]	Final /k s/ in Spanish tends to be converted to [ɣ s].
thanksgiving	/ˈθ æ ŋ k s g ɪ v ɪ ŋ/	[ˈθ e ŋ ‗ s ɣ iˑ β iˑ ŋ]	Deletion of /k/ to break-up 4 consonants.
American	/ə ˈm ɝ r i k ə n/	[a ˈm eː ɾ i k aˑ n]	[ɝ] turns into [eː] in frequent word.
bowman	/ˈb oʊ m ə n/	[ˈb o w m eˑ n]	*-man* transcribed as [m eˑ n] not [m aˑ n] for more intelligibility.
length	/ˈl ɛ ŋ θ/	[ˈl e ŋ k θ]	Insertion of /k/ after a nasal before fricative.
rainforest	/ˈr ei n f ɔ r ə s t/	[ˈr e j m f o ɾ eˑ s t]	/n/ before /f/ is converted to /m/.
straightjacket	/ˈs t r ei t ʤ æ k ə t/	[e s ˈt ɾ e j t j e k eˑ t]	[s] followed by a cons. at the word beginning to /e s/.
webcam	/ˈw ɛ b k æ m/	[ˈw e β k a m]	Very frequent usage of *cam* with an /a/.
jazz	/ˈʤ æ z/	[ˈjj a z]	Frequent word, /jj/ in 1st and /a/ in 2nd positions.
logjam	/ˈl ɑ g ʤ æ m/	[ˈl oː ɣ ð j e m]	/ʤ/ to /ð j/ after a consonant in a rare word and /ɑ/ to [oː] for more naturalness.
headquarters	/ˈh ɛ d k w ɑ r t ɚ z/	[ˈx e ð k w oː ɾ t eˑ ɾ s]	*-quart-* follows British tendency.
work	/ˈw ɝ k/	[ˈw oː ɾ k]	Frequently used form.
Haitian	/ˈh ei ʃ ə n/	[ˈx e j s j aˑ n]	/s j/ used to imitate /ʃ/.
Australian	/ə ˈs t r ei l i ə n/	[aˑ w s ˈt ɾ e l j aˑ n]	/a w/ corresponds to the orthographic form.
Nigerian	/n aɪ ˈʤ ɪ r i ə n/	[n a j ˈʧ iˑ ɾ j aˑ n]	/ʤ/ to /ʧ/ between a diphthong and a vowel.
Norwegian	/n ɔ r ˈw i ʤ ə n/	[n o ɾ ˈβ iː ʧ aˑ n]	/ʤ/ between vowels to /ʧ/.
Argentinean	/ɑ r ʤ ɛ n ˈt ɪ n i ə n/	[aː ɾ j eˑ n ˈt iˑ n j aˑ n]	/ʤ/ to /j/ after a consonant.
backgrounds	/ˈb æ k g r aʊ n d z/	[ˈb e k ɣ ɾ a w n ‗ z]	Deletion of /d/ before /z/ for more naturalness.
blindfold	[ˈb l aɪ n d f o l d]	[ˈb l a j m ‗ f o w l ð]	Deletion of [ð] before [f] for more naturalness; [n] before [f] to [m].
brainpower	/ˈb r ei n p aʊ ɚ/	[ˈb ɾ e j m p a w eˑ r]	/n/ before /p/ to /m/.
boyfriend's	/ˈb ɔɪ f r ɛ n d z/	[ˈb o j f ɾ eˑ n ‗ z]	Deletion of /d/ to avoid 3 consonants at the end of the word.
jeep	/ˈʤ i p/	[ˈd j i p]	[ʤ] to /d j/ at the beginning of the word.
Persian	/ˈp ɝ ʒ ə n/	[ˈp eː ɾ s j aˑ n]	/ʒ/ also transforms to /s j/, like /ʃ/.
father	/ˈf ɑ ð ɚ/	[ˈf aː ð eˑ r]	/ð/ is approximated by Spanish allophone [ð]

Table 5.5: Some examples of the nativization criteria application.

110

5.4 Nativization methods

In comparison with non-native speech, nativized speech is easier to manage in many aspects. Non-native speech is different from the native speech in articulation points, pause distribution, and diphone behaviour at word boundaries, and moreover, it is characterized by frequent pronunciation errors. Nativized speech, on the other hand, is more consistent in its definition, conserves the articulation point of the target language and does not contain important pronunciation errors, because its sole purpose is to mold the pronunciation of a foreign word to fit smoothly into target language utterances where foreign accented pronunciation would be unacceptable. Nativization is based either on a set of manually crafted or data-driven rules, all of which follow coherent criteria. Nativized speech does not contain mispronounced phonemes. The rest of the chapter is organized as folows: For English-to-Spanish nativization, all English phonemes were mapped to their closest analogs (see 5.2.2). This imperfect system, that considered no contexts and only a few exceptions that were left up to a language specific G2P converter, showed a significant improvement when compared to the transcriptions generated using Spanish G2P converter alone. In this approach, this method will be used as our baseline system. Pronunciation by analogy and learning from errors algorithms will be applied in order to improve the performance of the nativization system.

Section 5.4.1 gives a summary of the pronunciation by analogy algorithm and its application to the nativization problem. In this case, two nativization by analogy approaches were proposed: using information about the orthographic form and the original English pronunciation. Section 5.4.2 justifies the application of the transformation-based learning algorithm to improve the results obtained by the preceding methods using both orthographic and phonetic representations. Experimental results, error analysis and perceptual evaluation follow up in sections 5.5 5.5.5 5.5.6 correspondingly.

5.4.1 Nativization by analogy

The pronunciation by analogy algorithm, previously applied to G2P conversion (Marchand and Damper, 2000; Polyákova and Bonafonte, 2009), described in detail in Chapter 3, is applied to the task of nativization of English words in Spanish.

In (Polyákova and Bonafonte, 2009), we proposed six additional strategies for choosing the best candidate that in combination with the others outperformed the original strategies. The scoring strategies are based on the following parameters: frequency of appearance of a given phoneme arc in the dictionary; its length; and the actual phonemes that constitute the candidate. Different strategies work with different aspects of analogy. High arc frequency is considered a major advantage over low arc frequency. The frequency of suffixes and

prefixes are prioritized by different strategies. The final score for the candidate is directly proportional to the number of phonemes it shares with the others. If two candidates share the same pronunciation, both of them are prioritized. These measures are used separately or combined across the strategies. The strategies are explained in detail in Polyákova and Bonafonte (2009) and briefly in Table 5.6.

Nativization by analogy was attempted from two different viewpoints. The first approach, G2Pnat, is very similar to G2P conversion from letters to nativized phonemes. It makes sense to perform grapheme-to-phoneme nativization. In fact, most of the Spanish listeners are only familiar with the orthographic form of English words. However, when there is a phonetic transcription available in the source language, finding automatic correspondences between source and target (nativized) phonemes is a more consistent task than in the case of letters, being G2P conversion already a difficult task for English. That is one of the reasons why the second approach chosen for this task is the P2Pnat. The pronunciation by analogy method can be also applied to the phoneme-to-phoneme nativization. Since the input data consisted of phonemes there was a need for slight modifications in the dictionary processing part.

strategy mask	strategy meaning
10000000000	maximum arc frequency product
01000000000	minimum standard deviation of arc length
00100000000	highest same pronunciation frequency
00010000000	minimum number of different symbols
00001000000	weakest arc frequency
00000100000	weighted arc product frequency
00000010000	strongest first arc
00000001000	strongest last arc
00000000100	same symbols multiplied by arc frequency
00000000010	lowest count of different phonemes
00000000001	max. freq. product with most frequent same pron.

Table 5.6: Eleven scoring strategies for pronunciation by analogy.

5.4.2 Transformation-based error-driven learning (TBL)

Previous results obtained for grapheme-to-phoneme conversion using TBL to correct the errors (Polyákova and Bonafonte, 2006, 2008a), described in Chapter 3 encouraged us to consider this approach for our current work as well. In order to further exploit the possibilities for improvement of the nativized pronunciations using TBL, the algorithm was applied to the results obtained by P2Pnat and NatTAB. With the purpose of determining

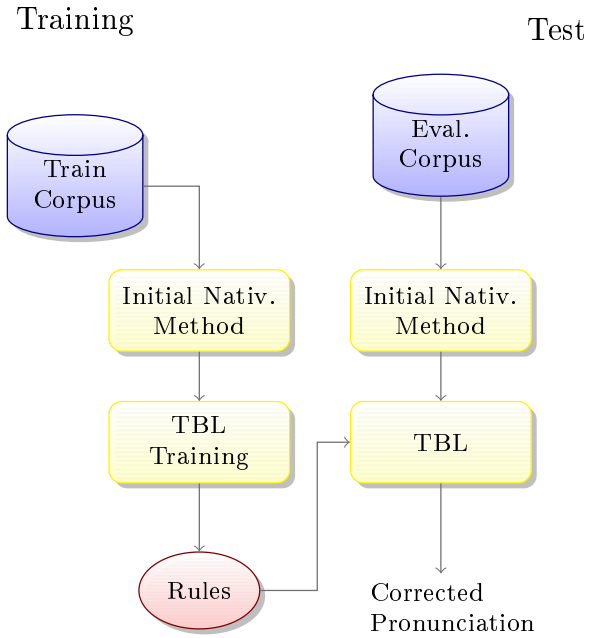


Figure 5.3: Scheme of combination of TBL with other nativization methods.

the generalization potential of the TBL algorithm itself, it was also applied to correct the results of a most-likely target phoneme prediction (ML). For this purpose, based on a lexicon aligned in a one-to-one manner each source phoneme was assigned the most-frequent target phoneme in the mapping.

Figure 5.3 shows the scheme of combination of the data-driven nativization methods with TBL. The transformation rules are derived from the errors in the initial nativized prediction needing improvements obtained by a previous classifier for the training data. The TBL algorithm not only allows correction of errors in the previous predictions but also an easy combination of different conditioning features for error correction. Here, we were able to use both orthographic and phonetic forms in the source language to improve accuracy of nativized pronunciations in the target language. The performance of the TBL algorithm highly depends on the size of the training corpus and the number of prediction errors obtained by the initial classifier. A higher error ratio in the training phase and a larger training corpus lead to better correction results. This was analyzed for the case of Mandarin polyphones prediction in Zheng et al. (2005).

5.5 Experimental results

In this section, we present the experimental results obtained for the nativization task using different methods. Pronunciations were derived according to the scheme shown in Figure 5.1. When an out-of-dictionary word was labeled as foreign (label F_LANG), its transcription was sought in the dictionary of the corresponding language (F_LANG dictionary). If the word was not in that dictionary, it was fed to a language-specific G2P system. In both cases, after the word pronunciation in a source language was determined, the nativization procedure was applied. First, in 5.5.1, we discuss the baseline results obtained with Nativization Tables (NatTAB). Next, in 5.5.2 and 5.5.3 we evaluate the proposed analogy-based approaches. Later, in 5.5.4, we describe an attempt to improve the best results obtained so far by applying the Transformation-based error-driven learning (TBL) algorithm. Furthermore, we evaluate the combination of TBL with simpler methods, such as Nativization Tables (NatTAB) or Most-likely phoneme assignation (ML). The latter combination allows to validate the performance of the TBL algorithm itself, given that the initial prediction, in this case, was very simple. Finally in 5.5.5, we compare the errors obtained by different nativization methods qualitatively.

5.5.1 Baseline results (NatTAB)

The NatTAB method carries out the nativization in a phoneme-to-phoneme manner, using hand-crafted nativization tables for the source-to-target phoneme transformations. The method based on the nativization tables was able to predict only 73.9% of phonemes and 23.8% of words correctly from CommonSet. However, these results are much better than those obtained for the same test data without using nativization, applying the Spanish G2P to derive the pronunciation of English words. Spanish G2P nativized only 61.2% of phonemes and 8.6% of words from CommonSet correctly. The only nativizations that this system predicted correctly were those pronounced very close to their orthography, e.g., *bed* → [b e ð] or *car* → [k a r]. The objective results obtained applying the table-based phoneme-to-phoneme mapping (NatTAB) for English to Spanish nativization were quite low in comparison with those reported for G2P conversion in many languages. Nevertheless, the results of the perceptual evaluation described in Polyákova and Bonafonte (2008b) showed that even such a simple nativization method had better acceptance among listeners than synthesized speech that implied no nativization at all and treated all words as if they were Spanish.

5.5.2 Grapheme-to-phoneme nativization (G2Pnat)

The first prediction method to be tested was the prediction of nativized pronunciation focusing on analogy in the orthographic word forms. Out of 11 strategies available for the PbA algorithm for choosing the best pronunciation candidate, it was necessary to determine the best strategy combination for our data. As we do not have any development data, an n-fold cross-validation was carried out on TrainingSet, leaving out each word at a time and using the remaining words for pronunciation lattice construction as described in 5.4.1. All possible strategy combinations were considered and compared. For G2Pnat, the resulting best strategy combination for TrainingSet was 10001001011 (1 meaning that the strategy corresponding to that position was included and 0 that it was omitted). The best n-fold results obtained for TrainingSet were 85.7% in phoneme and 45.6% in word accuracy. As already mentioned in 5.3.2, both training and test data contained lexical stress and vowel length information. However, the vowel length was not predicted at this time but will be addressed in the future. Firstly, it was important to evaluate the nativization accuracy without introducing any additional complexity to the task. For this reason, for the first experiment, the stress markers were removed. The results obtained with CommonSet using the best strategy combination were 84.2% phonemes and 43.8% words correct (Table 5.7). If we compare these results to the baseline results obtained with NatTAB, we can see that the word accuracy rate has almost doubled. See Figure 5.4 for an overview of the results.

The follow-up experiment, also carried out with CommonSet was aimed at prediction of the stress and nativized phonemes together. Stress inclusion increased the number of errors considerably, resulting in accuracy rates of 74.7% for phonemes and 20.0% for words. This further demonstrated that in English, stress prediction uniquely from the orthographic form is a difficult task (Black et al., 1998b).

method	test set	phon. acc. [%]	word acc. [%]
G2Pnat	common	84.3	43.8
	proper	74.8	31.5
P2Pnat	common	91.6	63.8
	proper	87.2	55.6

Table 5.7: Results obtained for G2P and P2P nativization by analogy with CommonSet and ProperSet.

Since stress prediction results were slightly discouraging, experiments on the ProperSet were performed discarding this additional feature. The word accuracy obtained for G2Pnat on ProperSet was about 12 percentage points lower than that for CommonSet (see Table 5.7). Such a loss in accuracy can be explained by the fact that even if the proper

names test set contained the most frequent and rather simple proper names of strictly English origin, their orthography is deeper than that of the common words.

5.5.3 Phoneme-to-phoneme nativization (P2Pnat)

For P2Pnat experiments the PbA algorithm was also applied. The training lexicon used was the phoneme-to-phoneme version of TrainingSet, with source phonemes on the left-hand side and nativized phonemes on the right-hand side. Similarly as for G2Pnat, the best strategy combination (11011000010) was determined performing n-fold cross-validation of all possible strategy combinations. The best n-fold results obtained for TrainingSet were 91.8% of phonemes correct and 61.3% of words correct. The accuracies obtained for CommonSet were 91.6% for phonemes and 63.8% for words respectively (Table 5.7). These results showed that P2Pnat outperforms G2Pnat by 20 percentage points in word accuracy terms. For ProperSet, the phoneme-to-phoneme results were also promising: 87.2% in phoneme and 55.6% in word accuracy beat by 23 percentage points the grapheme-to-phoneme nativization results for the same dataset(Table 5.7). Furthermore, this method is advantageous because it allows copying of the original accent to the nativized form with 99% accuracy for CommonSet. As both CommonSet and ProperSet datasets are rather small, the confidence interval of these results is relatively large. However, even such small sets allow obtaining statistically significant results at the $p = .05$ level on the basis of a binomial significance test.

5.5.4 Applying transformation-based learning to nativization

In view of the improvements previously obtained using TBL for the G2P task (Polyákova and Bonafonte, 2008a),in our next approach we applied the transformation-based learning in order to improve the results of other nativization methods, as mentioned in Section 5.4.2. The experimental results were evaluated on CommonSet. As it can be seen from Figure 5.3, to learn error-correcting rules the TBL algorithm requires an initial prediction both for the training and test sets. In this work, TBL was aimed at correcting the initial nativization prediction for three methods: 1)nativized pronunciations obtained by Phoneme-to-phoneme Nativization (P2Pnat), 2)Nativization Tables (NatTAB) and 3)Most-likely phoneme assignation (ML).

Before running the TBL algorithm, it was necessary to obtain the initial predictions for training and test data for all methods. For the P2Pnat method the initial prediction for TrainingSet was generated using n-fold cross-validation, leaving out each word at a time and using the rest of the lexicon to derive the nativization of the word in question, by analogy with the remaining words. The initial prediction for the test data (CommonSet) was obtained using the entire TrainingSet. To obtain the initial TrainingSet and CommonSet

predictions with NatTAB, English phonemes were mapped to the closest corresponding Spanish phonemes given in the nativization table for this pair of languages. And finally, for the last experiment, the most-likely nativized phoneme was assigned to TrainingSet and CommonSet as explained in Section 5.4.2.

The correction rules for all methods depended on such features as source letter, source phoneme and predicted phoneme, therefore, allowing combination of orthographic and phonetic knowledge from the source language. Different context types and lengths were considered (see Section 5.5.4). Table 5.8 shows phoneme and word nativization accuracies obtained for different initial predictions and contexts. The source letter context varied from 3 to 5, while the source phoneme context was considered in all cases but the first case, and its length varied from 1 to 3 phonemes. The predicted phoneme context was set to 3 for all experiments. The results are given for different methods (P2Pnat, NatTAB, most-likely phoneme (ML)) combined with TBL and for two different stopping thresholds: $t_1 = 1$ and $t_2 = 5$. The algorithm terminated when no rule with a score lower than the specified termination threshold was generated.

context type/methods	P2Pnat+TBL		NatTAB+TBL		ML+TBL	
	phoneme	word	phoneme	word	phoneme	word
Stopping threshold = 1						
orig_let = ± 3 orig_ph= 0	90.4	60.0	88.4	59.1	83.8	43.8
orig_let = ± 3 orig_ph= ± 1	91.1	62.9	88.4	59.1	83.5	40.0
orig_let = ± 3 orig_ph= ± 3	91.3	62.9	90.5	64.8	88.8	49.5
orig_let = ± 4 orig_ph= ± 3	91.5	64.8	90.2	63.8	88.5	47.6
orig_let = ± 5 orig_ph= ± 3	91.5	64.8	90.2	63.8	88.5	47.6
Stopping threshold = 5						
orig_let = ± 3 orig_ph= 0	92.0	63.8	87.7	59.1	78.6	32.4
orig_let = ± 3 orig_ph= ± 1	92.0	63.8	87.7	59.1	78.9	34.4
orig_let = ± 3 orig_ph= ± 3	92.7	66.7	90.0	64.8	87.0	43.8
orig_let = ± 4 orig_ph= ± 3	92.5	66.7	90.0	64.8	87.0	43.8
orig_let = ± 5 orig_ph= ± 3	92.5	66.7	90.0	64.8	87.0	43.8

Table 5.8: Phoneme and word accuracy (%) obtained by TBL in combination with different nativization methods as a function of letter and phoneme context used by the rules.

The best results of 66.7% words correct, were obtained for the largest source phoneme context and P2Pnat prediction. This was more than 2 percentage points higher than the result obtained using P2Pnat alone. The second best results, 64.8% of words correct, were obtained by applying the TBL method to the NatTAB prediction, and in this case, the initial word accuracy was improved by 20 percentage points. The results obtained by

NatTAB+TBL are quite good since they are slightly better than the performance of P2Pnat alone. TBL by itself proved capable of generalizing the nativization criteria when applied to correct the most-likely phone prediction, with a gain of about 24 percentage points in word accuracy in comparison to that obtained with NatTAB without TBL. However, the best TBL results are obtained when the best initial prediction is used, in this case P2Pnat. The results obtained by the combination of NatTAB+TBL and P2Pnat are quite similar and can be considered equal alternatives.

Even though no precise conclusions can be drawn, we can observe that larger letter and phoneme contexts appear to make a greater contribution to error correction. For training data containing less errors, as in the case of P2Pnat a higher stopping threshold seems to be more suitable.

The nativization results obtained on CommonSet using different methods are summarized in Figure 5.4. The differences are statistically significant in all cases except P2Pnat+TBL: we cannot ensure that TBL combined with P2Pnat gives better performance than P2Pnat alone. Such a small test corpus does not allow us to obtain statistically significant differences between best performing methods. Furthermore, for P2Pnat the number of errors available for rule learning is inferior to that obtained by other methods. Usually, good error correction rates are achieved for large lexicons of about 50K words and high error rates in the training prediction. If the initial prediction accuracy is rather high and the training corpus is rather small, the application of TBL may not give significant improvements. All improvements obtained by the TBL algorithm are consistent.

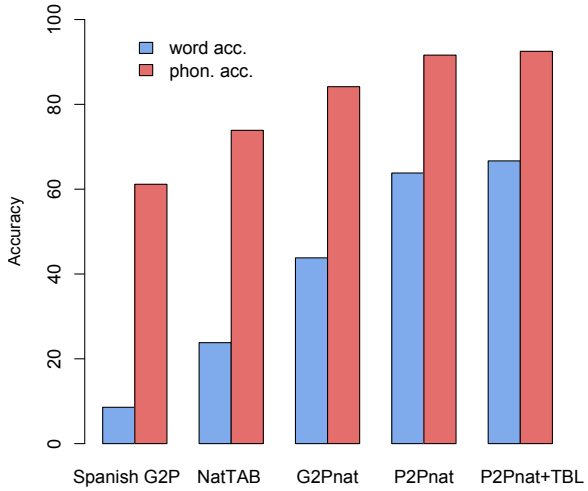


Figure 5.4: Word and phoneme accuracy obtained with: no nativization; hand-crafted nativization tables; grapheme-to-phoneme by analogy; phoneme-to-phoneme by analogy alone and combined with transformation-based-learning.

5.5.5 Error analysis

In terms of specific tasks such as nativization, an objective evaluation is insufficient to determine the validity of the results. Test results obtained with PbA using G2Pnat and P2Pnat were compared and exhaustively evaluated by the authors. Three types of errors were suggested. The term *severe errors* referred to the cases when the word was either unrecognizable and/or could be confused with another one. *Medium errors* referred to vowel confusion cases e.g. (a/e) (e/i), or (o/a). Vowel insertions and deletions together with similar consonant confusions (k/g, t/d, etc.) that did not drastically affect the intelligibility of the words were considered to be *light errors*. The results obtained using G2Pnat on CommonSet contained 22 severe errors affecting the intelligibility, while for the same test corpus using P2Pnat, only 10 severe errors were found. An example of a severe error is the pronunciation of the word *agency* nativized to [a ɣ e n s a j] or *general* to [ð j n e ɾ a l]. We considered the following nativization error for the word *agency*: [e j ʧu n s i] to be a medium error. An example of a light error would be the word *beautiful* nativized to [b j u ð i f u l]. Our experiments were performed using isolated words, making no pronunciation adjustments at word boundaries at this point.

It was also interesting to compare the errors obtained by more advanced nativization methods such as pronunciation by analogy with those obtained by the Spanish G2P converter. The most common severe errors obtained by the Spanish G2P converter on CommonSet that rendered words completely unintelligible were the following: in words that contained a combination of a *g* with either an *e* or an *i*, e.g., *girl* and *give*, the first phoneme /g/ was converted into /x/, and the final silent *e* in *give* was transcribed as /e/. In words that started with an *h*, the sound /x/ at the beginning was lost and *home* became [o m e]. Words such as *cool* and *need* were transcribed as [k o o l] and [n e e ð], respectively.

5.5.6 Perceptual evaluation

The last step to validate nativized pronunciations was to carry out a perceptual test with synthesized signals. The perceptual quality of the synthesized speech can be only defined by a listener and therefore is a difficult issue.

Thirty-eight volunteers were asked to evaluate 20 utterances synthesized using 3 different nativization methods. The utterances were produced using a concatenative unit-selection synthesizer (Bonafonte et al., 2008). The system concatenates diphones selected from a 10 hour speech database recorded by a professional speaker in a recording studio (Bonafonte et al., 2006a). Each of the 20 utterances contained 1 to 6 foreign words, excluding the articles and two-letter prepositions, grouped into maximum of 3 foreign word chunks. A few examples of the sentences offered to the listeners can be found below.

1. Los índices de *Wall Street* abren la sesión con ganancias. (The Wall Street index opens the session with gains).

2. *Microsoft* anunció hoy que sus beneficios cayeron un diez por ciento. (Microsoft announced today that its benefits dropped by ten percent).

3. *New York Stock Exchange* es el mayor mercado de valores del mundo. (New York Stock Exchange is the largest stock market in the world).

4. Su disco *Born to run* vendió quince millones de copias en Estados Unidos. (His album "Born to run" sold 15 millions of copies in the United States).

It is worth mentioning that the sentences varied in their difficulty and uncommon words at the beginning of the sentence could have been found less comprehensible due to the lack of preceding context. Anticipating this additional ambiguity issue we inserted a phrase opener that included the word "Frase" (sentence) followed up by its number in the list.

The listeners were given 20 sets of 3 randomly ordered utterances. In the group of 3, the possible choices represented 3 different nativization methods applied to the foreign words. These methods were: no nativization (Spanish G2P); our baseline system NatTAB (Section 5.5.1); and nativization by analogy P2Pnat (Section 5.5.3). For each group of 3 utterances the listeners were asked to choose the best and the worst of the 3. However,a "none" option was added to cover the cases when listeners could not clearly decide which utterance liked or disliked the most.

Listeners who volunteered for the experiment had different backgrounds in speech synthesis as well as in English and Spanish. Thirty out of 38 listeners were native speakers of Spanish, 2 were fluent and 6 claimed to have good knowledge of the language. Only 19 out of 38 were fluent in English, while the remaining half indicated to possess good knowledge of the language. Among the participants, 9 were experts in speech synthesis, 4 were experts in other speech technologies, 8 were occasional users of synthesis and the rest claimed no experience with synthesized speech whatsoever.

Overall evaluation results are shown in Figure 5.5. The graph shows the average number of times each method was chosen as best or worst independently of the sentence difficulty. From Figure 5.5 it is easy to see that the Spanish G2P method was voted worst in almost 45% of the cases, while the analogy-based method was voted best with a percentage close to 50%. Finally, the nativization by analogy had the lowest incidence of worst votes in comparison with other methods. The Nativization Tables (NatTAB) method received a similar number of best and worst votes. The percentage of indecision in both cases oscillated around 10%. The results allow to draw the same conclusions as the objective test: the analogy-based method (P2Pnat) performs much better than the table-based method (NatTAB),

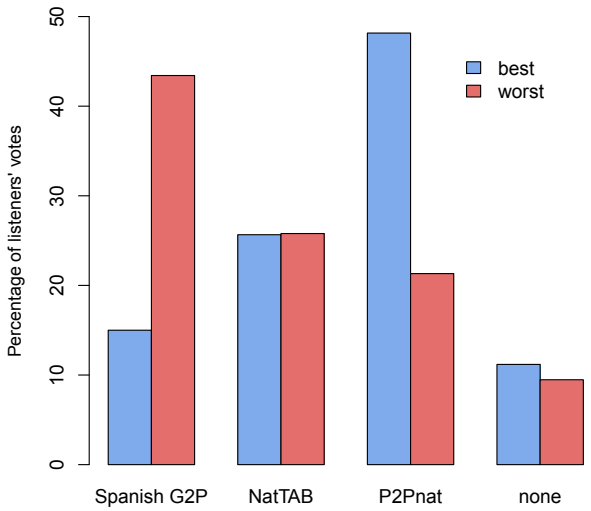


Figure 5.5: Perceptual evaluation of the TTS system using three different nativization methods.

which in its turn shows better performance than the original Spanish G2P. The results are statistically significant at the level $p = .05$ on the basis of a binomial test.

The separate analysis of the results based on the listeners' background showed that their previous experience with speech synthesis for this particular task had more influence on the results that their knowledge of Spanish or English. Experts in speech synthesis, thus, showed a stronger preference for the analogy-based method and a stronger rejection of the Spanish G2P-based method. Although the goal of this work was to evaluate nativization from the viewpoint of native Spanish speakers, due to the great enthusiasm towards the test shown by the non-native group members, we decided to include and analyze their contribution. The pattern of the non-native Spanish speakers' reactions to the test presented a higher variability as it was more difficult for them to recognize subtle differences between pronunciations generated by different nativization methods. However, they strongly preferred the analogy-based method, choosing it in 53% of the cases. This may be related to the fact that the analogy-based nativization sounded more correct from the point of view of English phonetics. For the rest of the subgroups the general tendency was similar to that shown for the overall results in Figure 5.5.

Curiously, nativization by analogy was voted as the worst method for inclusions such as *hangover* and *born to run* and the highly assimilated word *Microsoft*. In the first case, the main difference lies in the nativized pronunciation of the English phoneme /r/. The nativization by analogy method disregarded the position-dependent pronunciation of the grapheme *r*, and all English /r/ were converted to Spanish intervocalic /ɾ/ in the training corpus and, therefore, in the resulting synthesis. Consequently, the unit selection synthesizer

could not find any /ɾ/ at the beginning of the word or before a consonant using instead the demi-phonemes that sounded very close to the Spanish dental approximant [ð]. In the case of the word *Microsoft*, the most common pronunciations in Spain are [m i k ɾ o ˈs o f t] or even [m i k ɾ o ˈs o f]. Even [ˈm a j k ɾ o s o f t], the alternative offered by the table-based method, was found less pleasant and the pronunciation predicted by the analogy-based method [ˈm a j k ɾ o s a f t] was considered too foreign sounding. The assimilation-influenced accent displacement was not accounted for in this work.

At this point, we can conclude that the best received method was the phoneme analogy-based nativization and the worst was the Spanish G2P converter (absence of nativization). However, the frequency of word usage introduced variability and nuances.

5.6 Application of the nativization to Catalan inclusions in Spanish utterances

In Catalonia non-Spanish inclusions are abundant because there are two languages that represent the source of these, they are namely English and Catalan. The nativization of English inclusions has been thoroughly explained in the beginning of this chapter. As significant phonetical differences exist between Spanish and Catalan as well, nativization was also found to be necessary for this pair of languages. In order to train a nativization model to convert Catalan pronunciations to acceptable Spanish ones, the pronunciation of the phonemes that do not exist in Spanish, need to be adapted. Such factors as frequency of usage of a word and Spanish pronunciation rules, etc., need to be carefully taken into account to achieve a higher performance of TTS synthesizers. Both nativization by analogy scopes were exploited in Catalan to Spanish nativization task 1)training of a nativization model using source orthographic forms and nativized phonetic transcriptions and 2) usage of source and nativized pronunciations for training. In order to apply data-driven techniques to nativization a need for training and test data arises. For grapheme-to-phoneme conversion tasks large pronunciation corpora of 100 thousands words and their corresponding pronunciations are available. Since we did not find any existing nativization database for this pair of languages, we chose to manually create a minimalistic corpus that would not require expert linguistic knowledge. For our task the Catalan training corpus *CTraining Set*, extracted from the LC-STAR lexicon Hartikainen et al. (2003) was orthographically balanced in order to have all possible letter bi-grams in the corpus, with a total of 1000 words. The original phonetic transcriptions of these words were manually nativized according to the criteria described in the book of styles for one of the Spanish TV channels (Llorente and Díaz Salgado, 2004). It is necessary to emphasize that the phoneme inventory used for nativization was limited to Spanish phoneset. The test data for Catalan consisting of a 100 words, *CCommonSet*, was manually collected from the available on-line

sources. Since a thousand words was selected for training, it was found appropriate that the test data comprised 10% of the training corpus. None of the test words were present in the CTrainingSet. It was intended that the test words were frequently used and with simple meanings in order for the results to be unbiased by other factors. Here are some examples of train *agredolça, boirumós, migjorn* and test words *enllaç, desig, forjar*. The next section focuses on the phonetic differences between Spanish and Catalan and on the elaboration of nativization criteria, based on these differences.

5.6.1 Spanish phonetics vs. Catalan phonetics

The sounds of a language are defined by a phoneme inventory or phoneset. A phenomenon called extension of the phoneset often occurs in bilingual communities and speakers; however, it is impossible to study foreign word pronunciation at the level of each individual. It is much easier to observe general tendencies in bilingual societies. In the particular case of Catalan, both nativization and phoneset extension phenomena occur. It is curious to note that Spanish words in Catalan are pronounced using regular Spanish phoneset, due to the fact that the majority of Catalan speakers are perfectly fluent in Spanish. For example, the Spanish name *Jaime* in Catalan is pronounced `/'x a i m e/` and not `/'dZ a i m @/` as Catalan phonetics would stipulate. Even though the phoneme `/x/` is absent from Catalan, it is widely used for Spanish proper names and other inclusions, e.g. quotations. On the contrary, the pronunciation of Catalan words in Spanish is adapted according to Spanish pronunciation rules and the phoneset extension phenomenon is rather rare. Spanish and Catalan have several major phonetic differences which depend on the dialect of the latter. Most varieties of Catalan possess seven stressed vowels that are: `/a/`, `/e/`, `/o/`, `/u/`, `/i/`, `/E/`, `/O/`,`/@/`. Open vowels `/a/`, `/E/` and `/O/` as well as the unstressed `/@/` do not occur in Spanish. In Spanish, medium vowels can be realized as open only in particular contexts, while in the rest of the cases all vowels are articulated as closed. In Catalan, however there is an important phonological difference between open and closed vowels, which can not be attributed to the context and therefore is not predictable. For example, homographs *seu (yours) vs. seu (headquarters)* `/s 'e u/` vs. `/s 'E u/` have different meanings depending on the vowel articulation point. A diagram of Catalan vowels can be found in Figure 5.6 As well as in Spanish, in Catalan there are six plosives `/b/`, `/d/`, `/g/`, `/p/`, `/t/`, `/k/` (3 voiced and 3 unvoiced) at three different articulation points. Catalan does not have any dental, uvular or velar fricative consonants sounds, but has two alveolo-palatals `/Z/` (voiced) e.g. *vigent* `/b i Z 'e n/` and `/S/` (unvoiced) e.g. *caixa* `/k 'a S a/`. The labiodental `/v/` exists in Catalan as a result of sonorization of any `/f/` before a voiced consonant or a vowel at the beginning of the word. Besides, in Catalan, all unvoiced fricatives are sonorized if followed by a voiced consonant. The fricative `/z/` which is very frequent in Catalan, exists in Spanish as an allophone but not as a phoneme. Catalan has four affricates, 3 more than Spanish,

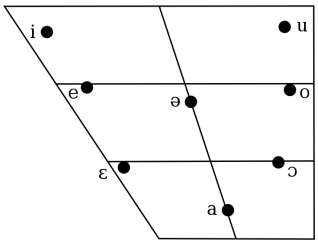


Figure 5.6: Vowels of standard Eastern Catalan

the voiced affricates `/dz/`, `/dZ/` and the unvoiced `/ts/`, `/tS/`. The phonemes `/ts/` and `/dz/` arise mainly from compounding such as in *potser* /p u ts 'e/, but may as well occur at any other position as in *dotze* `/d 'o dz @/`. Similarly to Spanish, the nasals are adapted to the articulation point of the following consonant, however, in Catalan both `/m/` and `/J/` can occur at the end of the word, e.g. *any* `/'a J/`. There are two laterals in Catalan, as well as in Spanish, the alveolar `/l/` and the alveolo-palatal `/L/`. Additionally Catalan has the double *l.l* that is pronounced as a prolonged alveolar articulation `/l:/`. Both languages possess 2 trills, the simple `/r/` and the multiple `/rr/`. In contrast with Spanish, the `/r/` in Catalan can only appear at the intervocalic position or after a plosive or fricative that forms part of the same syllable e.g. *frau* `/ f r 'a uw/` or *cara* `/k 'a r a/` (Planas, 2005). A complete set of Catalan consonants can be found in Figure 5.7.

	Bila-bial		Labio-dental		Inter-dental		Dent.-alv.		Alveolar		Alveol.-palatal		Palatal		Velar	
Ocl.	p	b					t	d							k̟k k̲	g
Fric.			f	v			s̪	z̪	s	z	ʃ	ʒ				
Afric.									t͡s	d͡z	t͡ʃ	d͡ʒ				
Aprox.		β̞				ð̞								j̞		ɣ̞
Nasal		m		ɱ				n̪		n		nʲɲ				ŋ
Later.								l̪		l		lʲʎ				ɬ
Vibr. simp.										ɾ ɹ̝						
Vibr. múlt.										r						

Figure 5.7: Catalan consonants (Planas, 2005).

Taking into account most of the above mentioned phonetic differences, we developed nativization criteria in order to find the best pronunciation for Catalan words in Spanish utterances.

5.6.2 Nativization criteria for Catalan

The challenge of this task consisted in developing solid criteria for nativization, taking into account local specifications for certain words, pronunciation and word popularity factor, among others. Some of the criteria could not be easily formulated that is why using a training corpus clearly has an advantage over the rule-based approach. Several examples of the criteria used are described below. First of all, let us take a look at the vowels. All open vowels were mapped to the closed ones, while the unstressed `/@/` was mapped to `/a/` in most of the cases, except for those words that are were similar to Spanish where it was transcribed as `/e/` e.g. *adreçarà* from `/a D r a s a r 'a /` to `/a D r e s a r 'a/` in the nativized form. For consonants, some difficulties were found when transcribing `/Z/` andn`/S/`. Their nativization depended both on letter and phoneme context. The voiced fricative `/Z/` at the beggining of the word was nativized to `/jj/` e.g *jutge*, to `/tS/` before a nasal e.g. *taronja*, and to `/j/` in other cases as in *vorejar*. The unvoiced fricative `/S/` was transcribed or to `/j s/` when it corresponded to the digraph *ix* e.g. *coix*; to `/tS/` when it corresponded to the same phoneme in a similar Spanish word e.g *anxoves*(*cat.*) vs. *anchoas*(*sp.*); or to `/s/` in the rest of the cases. The affricate `/dZ/` was nativized to `/tS/` as in the word *migdia*. Affricates `/ts/` and `/dz/` were mapped to the corresponding double phonemes `/t s/` and `/d z/`. Multiple trill `/rr/` was conserved only in the cases when it coincided with the Spanish phonetic rules, in all other cases it was changed to the simple trill `/r/`. The nasal `/N/` and the voiced `/z/` were conserved as they were present in our voice database. Silent *r* at the end of the verb in a compound verb-pronoun construction such as $afegir - n'hi$, was restored in the nativized form for the sake of comprehension. The database nativization task was carried out by the authors using both the source language orthographic form and pronunciation. Data-driven methods were used to approach the nativization for Catalan.

5.6.3 Experimental results

Nativization tables were able to convert 79.74% phonemes and 21.78% words correct from the CCommonSet described in Section 5.6.1. These results are much better than those obtained without using nativization, applying the Spanish G2P to derive the pronunciation of Catalan words, Spanish G2P scored only 33.97% correct in phoneme and 3.96% in word nativization on the same 100 word test corpus. The words nativized correctly by this primitive system were those that are pronounced very closely to Spanish orthography, for example *aquí to* /a k 'i/ or *sac to* /s 'a k/.

Grapheme-to-phoneme nativization

As well as for English to Spanish nativization, all possible strategy combinations were considered and compared. For Catalan grapheme-to-phoneme nativization (G2Pnat), the resulting best strategy combination was the following: 0000101101. The best results obtained on CComonSet were 86.51% in phoneme and 38.61% in word accuracy. When each strategy was considered individually, the best results were obtained for the eleventh strategy that combines the frequency product with the frequency of the same pronunciation. The lowest scoring strategy is seventh strategy that prioritizes the candidates with very frequent first arc. The results for each single strategy and the best strategy combination are given in Table 5.9

strategy mask	ph. acc.	word. acc.
10000000000	85.80	35.64
01000000000	83.96	31.68
00100000000	83.82	31.68
00010000000	83.79	31.68
00001000000	84.62	32.67
00000100000	85.65	33.66
00000010000	83.38	30.69
00000001000	85.53	33.66
00000000100	83.96	35.64
00000000010	83.69	30.69
00000000001	85.86	36.63
00001011011	86.51	38.61

Table 5.9: Single strategy results for G2Pnat and best strategy combination.

Phoneme-to-phoneme nativization

Phoneme-to-phoneme nativization P2Pnat makes a lot of sense in case of Catalan as well, in fact, non-Catalan speakers apply Spanish grapheme-to-phoneme rules when reading Catalan; however, in order to find the best pronunciation for Catalan phonemes absent from Spanish the phonetic transcription available in the source language may be quite helpful. Finding automatic correspondences between source and target (nativized) phonemes is a more consistent task than in the case of letters, being G2P conversion already a challenging task for Catalan especially for such a reduced training corpus. The best strategy combination (11010101010) as in case of G2Pnat was determined performing n-fold cross-evaluation of all possible strategy combinations. The results obtained on 100 word CCommonSet were 92.09% phonemes and 56.44% words correct. These results show

that P2Pnat nativization outperforms G2Pnat nativization by 22% percentage points in word accuracy terms. Performing single strategy experiments for Phoneme-to-phoneme Nativization (P2Pnat) we can also observe that the best scoring strategies are the sixth and the eight one, while the worst place belongs to the ninth. For more results see Table 5.10.

strategy mask	ph. acc.	word. acc.
11000000000	91.51	53.47
10000000000	91.51	52.48
01000000000	90.09	48.51
00100000000	90.51	48.51
00010000000	90.95	50.50
00001000000	90.79	49.50
00000100000	91.65	54.46
00000010000	89.94	46.53
00000001000	91.51	54.46
00000000100	89.18	42.57
00000000010	90.38	48.51
00000000001	90.92	50.50
11010101010	92.09	56.44

Table 5.10: Single strategy results for P2Pnat and best strategy combination

Although subjective evaluation has not been performed, the magnitude of the ojective improvements in nativized words accuracy from 21.18% (baseline) to 56.44% (best resultP2Pnat),similar to the objective results obtained English to Spanish nativization, proves the nativization for Catalan to be as effective.

5.7 Conclusions

Multilingual speech synthesis module is a must for any up-to-date synthesizer. With the unstoppable growth of social networks usage and globalization, multilingualism has become contagious and wide-spreading. In fact, we can encounter mixed languages not only in rather informal social networks, but also in well-known conservative mass media. Online versions of these feature comments in many languages as well as comments in mixed languages because multilingualism has now become a modern trend. An efficient automatic module capable of handling multilingual input to the TTS, therefore was needed. Spain is known for its linguistic heritage. Four official languages are spoken here. These are: Spanish, Catalan, Galician and Basque. This chapter focuses on two different tasks applicable in multilingual speech synthesis: language identification and nativization. Language identification is made part of the multilingual TTS in order to improve the pronunciations of words of foreign

origin, including both proper and common nouns. Two different types of multilingual texts were considered, texts that contained foreign word inclusions and those that contained paragraphs written in different languages. Thus, the first task that required attention was the identification of the language of the paragraph. For those paragraphs that contained words foreign to the paragraph itself the language of the paragraph was also important because it allowed to determine the main language in which the speech should be produced for the paragraph in question. For single-word foreign inclusions, the language of origin helped to adapt the words to the new main (or target) language environment correctly. This adaptation or as it will be further referred to, *nativization*, is important because it allows to smooth out the harsh contrast between pronunciations in two non-similar languages. It is a clear need in case of Spanish text with foreign inclusions, because Spanish, lacking some affricate and fricative consonants typical in other widely spoken languages, has rather low tolerance for foreign sounds.

Nativized pronunciations are more tolerant to vowel and consonant substitutions. There is no gold standard for nativization, and some exceptions that occur in highly assimilated pronunciations increase the difficulty of the problem. However, these exceptions are created by humans and obey the analogy both in orthographic and phonetic forms. Simple mapping rules were proven to be insufficient for the task.

First, a baseline multilingual synthesizer that included a baseline language identification system and a simple nativization method, was developed. n-gram based language identifier has shown rather good results for the five languages considered (English, Spanish, Catalan, Basque and Galician. Nativization tables, that mapped phonemes in one language to their nativized analogue in the other, were created to adapt the pronunciation of foreign words. A preliminary perceptual evaluation was carried out to validate the effectiveness of nativization tables. Ten volunteers were asked to evaluate utterances containing foreign inclusions in terms of naturalness and intelligibility. The results show that in the majority of cases, even such a baseline nativization was preferred to using Spanish G2P to derive the pronunciation of foreign words.

Later in this chapter, we proposed to use pronunciation by analogy for nativization of English and Catalan inclusions in Spanish. Analogy was considered both in orthographic and phonetic domains. Having found no suitable nativization database for these pairs of languages, we proposed to create minimalistic, phonetically-balanced training corpora, consisting of 1000 words each, for both languages. Different criteria were considered while creating nativized pronunciations for the corpora. Simple mapping rules were proven to be insufficient for the task because some of the criteria could not be easily formulated. The test corpora were not included in the corresponding training corpus and contained 100 frequently used words each.

The nativization results for English, obtained by using analogy only in the orthographic domain were much better that the baseline, however, leaving a lot of possibilities for improvement (43.8% words correct for the CommonSet) due to deep orthography of the English language. We believe, and the numbers back us up, that the G2Pnat results obtained in the experiments carried out in this chapter are better than G2P results could have been obtained for the same minimalistic corpus of only 1000 words. It is worth mentioning that even in the case of G2Pnat, the results show very significant improvements in comparison to those obtained by direct phoneme-to-phoneme table-based mapping (NatTAB). The method based on analogy in the phonemic domain, P2Pnat, gave an improvement of approximately 12-14 percentage points (both for Spanish and Catalan) with respect to the orthographic analogy, thus showing the tight connection between the pronunciation in the source language and the nativized one. TBL algorithm applied in combination with other methods produced rather good results. NatTAB+TBL performed slightly better than P2Pnat. The best results were achieved using P2Pnat enhanced by the TBL algorithm (66.7%) allowing to incorporate additional information about the orthography in the source language. As this improvement was not statistically significant, the TBL algorithm was not considered for Catalan.

A perceptual test was conducted for English in order to determine the rate of acceptance of speech containing nativized utterances by the listeners. Thirty-eight volunteers who participated in this subjective evaluation were asked to rank three nativization methods, namely Spanish G2P rules, nativization tables and P2Pnat. The utterances were randomly put into 20 groups of three and the listeners were given three choices to evaluate them:*best*, *worst*, *none*. The latter option was added for the cases where a listener could not decide on the quality of the synthesized utterances. Spanish G2P rules were clearly voted to be the worst method for this task. It received almost 45% of the *worst* votes. Nativization tables were better accepted by the listeners with 25% of *worst* and *best* votes. Finally, P2Pnat also showed a significant advantage over the nativization tables with 50% of *best* votes. For Catalan the results were quite similar and equally encouraging. G2Pnat outperformed NatTAB by almost 17 percentage points, while P2Pnat, in its turn, outperformed G2Pnat by another 17 percentage points. A summary of the experimental results, obtained in this chapter can be found in Table 5.11.

Both objective and subjective evaluation have shown the effectiveness of the proposed nativization method and the good rate of acceptance of the nativized speech by the listeners with different background in speech synthesis.

In the future, it would be interesting to tackle the reverse problem (Soonklang et al., 2008) because Spanish inclusions in English utterances could result in quite unintelligible pronunciations simply by applying English G2P, being that Spanish is a vocalic language with transparent letter-to-sound rules. For example *Jorge Casacubierta*

Method	Results for English	Results for Catalan.
Spanish G2P	8.6	3.96
NatTAB+TBL	23.8	21.78
G2Pnat	43.8	38.61
P2Pnat	55.6	56.44

Table 5.11: Summary of the experimental results for nativization of English and Catalan inclusions in Spanish.

would be pronounced as [dʒ o r ʒ k a z ə k j u b ə t ə] if the English G2P rules were applied directly. This pronunciation would be rather difficult to understand for both Spanish and English native speakers. Nativization would significantly improve the intelligibility of this proper name.

Chapter 6

Conclusions

This thesis is dedicated to exploring way of improving the quality of the speech produced by the TTS through enhancing its first component, the phonetic module responsible for pronunciation generation for the input text. Text-to-speech (TTS) is widely used for over three decades but new challenges are continuously arising, especially regarding the pronunciation module. Modern life, full of newest revolutionary technological inventions is more dynamic than ever. If speech technologies were important a decade ago, now they are simply indispensable, the requirements to quality and reliability are also becoming more demanding as time goes by, logically because we delegate a great deal of our everyday decisions to the devices. They read our emails, tell us the fastest way to the hospital, the best way to get through the myriad of highways, and lately smart helpers on our "smart devices" even keep us company by talking to us and advising on our everyday life or at least trying to do so the best they can. You can talk to your phone, tablet PC or even watch. A huge number of technologies is, of course, involved in the correct and reliable functioning of these innovative systems and undoubtedly, the role of a reliable and intelligible TTS is huge. The modern TTS not only have to be reliable, but also adaptable to the changing and rapidly developing conditions of the today's technological market. The global vocabulary is in continuous expansion like the universe itself so we need methods capable to embrace these changes. And last, but definitely not the least is the multilingual aspect of the rapidly developing global village. Mobility, media, technologies have created a need for a highly intelligible and adaptable and multilingual speech synthesizer.

This dissertation is mainly focused on these three issues: adaptability, reliability, multilingualism.

6.1 Adaptability

Chapter 3 is mainly focused on the adaptability of the phonetic module of modern TTS systems. Obviously in such a world where everything is changing so rapidly the functionality of the phonetic module cannot be limited to a fixed set of LTS rules. Moreover, data-driven systems have been proven to show much better performance for languages with deep orthography such as English. Several data-driven systems are applied to the task of automatic pronunciation generation and the improvements were obtained.

The adaptability issue has been approached through focusing on different machine-learning techniques for grapheme-to-phoneme conversion module of a TTS synthesizer and comparing their performance. The same evaluation techniques and datasets were used to allow a more precise comparison. Such machine-learning techniques as Decision Trees (DT), Finite State Transducers (FST) Hidden Markov Models (HMM), Pronunciation by Analogy (PbA) and Transformation-based error-driven learning (TBL) have been applied and compared. These techniques are data-driven, language-independent and corpus-based. Their performance as well as their flexibility and portability match the ones required for high quality grapheme-to-phoneme conversion for mono- and multilingual synthesis.

The experiments were performed using three lexica of American English: Unisyn, NETtalk and LC-STAR. The first series of the experiments performed compare two state of the art G2P conversion techniques, namely DT and FST. The results are obtained both for phoneme prediction with and without lexical stress. As expected, the number of errors obtained for the phonemes without stress marks was significantly inferior. Furthermore, FST gave better performance in both cases.

At the next step, the PbA was developed and applied to the task of G2P conversion. New scoring strategies were proposed and the improvements were obtained based on these strategies. The 1.5-2.5 percentage points of error reduction was obtained in comparison with the original work of Marchand and Damper (2000).

For all three lexica evaluated,one of the proposed strategies (the eleventh strategy) was found to be the best. Furthermore, it is present in all top 5 strategy combinations together with some other newly proposed strategies. The performance of each strategy was analyzed for different word lengths; the eleventh strategy performs best both for words and phonemes for all word lengths. Further improvements were achieved when these baseline results were enhanced by means of applying a set of transformation rules acquired through learning from errors. The transformation rules were learned automatically from a training corpus previously labeled using four different data-driven classifiers. The rule templates are highly adaptable; being language-independent they can be easily used to generate transformation rules for any language. The combination of any method with transformation-based error-driven algorithm has significantly improved the results obtained

by the same method alone. This is especially relevant for worst-performing classifiers. The best G2P results were obtained for the combination of FST with TBL algorithm. The error-transformation rules were also trained and applied to a simple prediction obtained by assigning the most-likely phoneme to each letter based on letter-phoneme pairs seen in the lexicon.The results obtained proved the effectiveness of the transformations rules. In fact, these results were better than those obtained by the widely used DT and HMM and closely comparable to those obtained by FST and PbA before the application of TBL. The transformation-based learning algorithm was also applied to improve the prediction of the PbA and the results were analyzed. New strategy combination methods were considered and slight improvements attained. The fact that application of error-correction rules did not give significant improvements allows concluding that the PbA method is quite capable of capturing the regularities in English orthography by itself.

The pronunciation was also derived for other languages. For Spanish, the obtained results were high as expected, since for languages with shallow orthography the pronunciation of common names can be as easily inferred by a small set of simple rules as by MLR techniques. In the case of proper names and neologisms the simple rules might have difficulty to predict the pronunciation. The most difficult languages in G2P conversion task were found to be Slovenian and German.

The overall conclusion that can be drawn from the experimental results obtained is that pronunciation by analogy algorithm including the new strategies gives the best results for G2P conversion for all English lexicons, although it is closely followed by the performance of the finite-state transducers enhanced by the transformation-based learning algorithm. The size of the training corpus as well as the method used to align the training data have major influence on the system performance.

Furthermore, we analyzed different factors that could influence error rates in grapheme-to-phoneme conversion. The error distribution obtained for the LC-STAR corpus indicates that mainly 9-letter long words contribute to the total error rate, if the optimal model parameters are chosen for training of the system.

The case of connected speech was also considered in the framework of Blizzard 2007 and 2008 challenges. Phonetic weak forms were introduced to our TTS synthesizer in order to find the best match between speaker's dialect and vocalization style. A more precise phonetic transcription positively influences the resulting synthesis quality.

Having obtained these improvements we have showed a great degree of adaptability of grapheme-to-phoneme conversion systems. They are easily transferrable to other languages and highly adaptable to the expanding dictionaries.

The results obtained in this chapter can be found in the following articles:

T. Polyákova and A.Bonafonte. Main issues in grapheme-to-phoneme conversion. In *Actas del XXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, (SEPLN2005)*, pages 29–34, Granada, España, September 2005. URL `http://gps-tsc.upc.es/veu/research/pubs/download/780.pdf`

Tatyana Polyákova and Antonio Bonafonte. Learning from errors in grapheme-to-phoneme conversion. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 2442–2445, Pittsburgh, USA, September 2006

A. Bonafonte, J. Adell, P.D. Agüero, D. Erro, I. Esquerra, A. Moreno, J. Pérez, and T. Polyákova. The upc tts system description for the 2007 blizzard challenge. In *Proc. of the 6th ISCA Speech Synthesis Workshop (SSW6)*, Bonn, Germany, August 2007

Tatyana Polyákova and Antonio Bonafonte. Further improvements to pronunciation by analogy. In *Actas de las V Jornadas en Tecnologías del Habla*, pages 149–152, Bilbao, Spain, November 2008a

A. Bonafonte, A. Moreno, J. Adell, P.D. Agüero, E. Banos, D. Erro, I. Esquerra, J. Pérez, and T. Polyakova. The UPC TTS system description for the 2008 blizzard challenge. In *Proc of the Blizzard Challenge, Brisbane, Australia*, September 2008

Tatyana Polyákova and Antonio Bonafonte. New strategies for pronunciation by analogy. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4261–4264, Taipei, Taiwan, April 2009

6.2 Reliability

Chapter 4 mainly focuses on the improvement of reliability of pronunciations obtained by the phonetic module. Even though G2P results have been very promising, even a few pronunciation errors can cause serious damage to the speech quality. The main idea lies in the fact that pronunciations previously validated by human experts are more reliable than those generated by artificial intelligence. Previously validated pronunciations are merged together to create a bigger and more reliable database, the issues arising in the process are discussed and solutions proposed.

After carrying out the experiments aimed at the improving of the reliability of the phonetic transcription results, several conclusions can be made. The existence of significant number of inconsistencies between different dictionaries has been confirmed; these inconsistencies have different origins, some of them are caused by discrepancies in expert opinions, others can be a result of using different unhomogenized sources for lexica generation. For instance, taking the words common to two of the lexical, the coincidence of pronunciations varies between 30 and 70%, depending on the lexicon. A

great part of these discrepancies can be overcome by applying the proposed dictionary fusion algorithm which automatically derives Phoneme-to-phoneme (P2P) conversion rules in order to get the adapted target pronunciation. The discrepancies that appear because different transcription criteria are used are not mistakes, however the TTS requires the pronunciations to be consistent. Dictionary fusion expands the target lexicon by adding new words from source lexica after they have undergone an adaptation. It has been carried out for both proper names and common words. The application of fusion algorithm gives important improvement for common words (more than 16 p.p.) Fusions seems more feasible and performs much better for common words that for proper names. This can be explained by the fact that proper names transcription per se is a very challenging task even for human experts taking into consideration different origins of these names and different assimilation degrees reached within a language. A proper name that isn't very frequent in a language is unlikely to be given the same exact phonetic transcription by different linguists. The experiments involved three lexica: CMU, Unysin and LC-STAR. The latter was selected as the target dictionary to be expanded via fusion. Fusion has been attempted by applying two data-driven algorithms: DT and FST. The results obtained by these methods differ less than for grapheme-to-phoneme conversion. Fusion errors have been classified into three groups: light, medium and severe; depending on their impact on intelligibility of the synthesized speech. Fusion has shown a significant decrease in the number of severe errors that seriously hamper the intelligibility of the synthetic speech. This important improvement increases the reliability of the phonetic transcriptions used, therefore guaranteeing a better quality of speech at the output of the synthesizer.

This results of this chapter can be found in the following article:

Tatyana Polyákova and Antonio Bonafonte. Fusion of dictionaries in voice creation and speech synthesis task. In *Proceedings of the International Conference on Speech and Computer*, Moscow, Russia, October 2007

6.3 Multilingualism

Chapter 5 focuses on the high priority challenge of all up-to-date TTS systems, the multilingualism. Multilingualism is common all around the world, however, the particular study territory for the multilingual phenomenon was chosen to be Catalonia, Spain, as our university is located in Barcelona. Many steps were taken toward achieving the goal of intelligible multilingual synthesis, specific for the study area. Non-native inclusions (English and Catalan) in daily conversation, TV programs and written press have been thoroughly studied taking into the account not only the origin of the word but also its frequency in the language. A multilingual system including nativization was designed. As the first step a quick approach was chosen and a baseline language identification system was proposed. The

results that were obtained at this step were promising and motivating towards obtaining further improvements. The baseline nativization method involved nativization tables, that were found insufficient to cover multiple aspects of nativization. The need for a data-driven nativization method base on corpora has arisen. As no suitable corpora was found, new nativization corpora were developed for both English and Catalan inclusions in Spanish based on the adaptation criteria derived from the study. New automatic nativization system based on nativization corpora was proposed and significant improvements were obtained and confirmed by the perceptual evaluation.

A synthesizer cannot be up-to-date if it is not prepared to deal with the challenge of multilingualism. Thriving linguistic diversity and mixing languages have gained a lot of importance in our lives. This phenomenon is spread through numerous media sources, social networks and even by polyglots who use several languages in one sentence, just to name a few. The growing range of TTS applications demands the systems to be more reliable and adaptable than ever, especially in the multilingual scope. Multilingual speech needs to be smooth and intelligible for speakers of all languages in question. An automatic multilingual TTS system which complies with the above mentioned requirements has been introduced. A special focus has been given to language identification and nativization. Language identification is an important part of any multilingual TTS because the pronunciation of foreign words highly depends on their origin. Language detection task has to consider language identification of a paragraph in a multilingual text and also of isolated words. For those paragraphs with foreign inclusions language of origin was important because it allowed to determine the main language in which the speech should be produced for the paragraph in question. For single-word multilingual inclusions, the language of origin helped to find the best way to adapt the pronunciation of this word to the paragraph language. This pronunciation adaptation or *nativization*, allows to smooth out the harsh contrast between pronunciations of words with different language origins. Spanish language in comparison to languages like English or Catalan is lacking some of the voice affricate and fricative consonants, etc. Its tolerance for foreign phonemes, is comparatively low.

First, a baseline multilingual synthesizer with a baseline language identification system, using a simple table-based nativization method, was developed. Five languages (English, Spanish, Catalan, Basque and Galician) were considered for the languages identification task and the baseline letter n-gram based language identifier has performed rather well. Nativization tables, mapped phonemes in one language to their nativized analogue in the other, were created to adapt the pronunciation of foreign words. The effectiveness of nativization tables was validated by a perceptual evaluation that consisted in ten volunteers rating the intelligibility and naturalness of the sentences with foreign inclusions nativized using this baseline method. The listeners agreed in most cases that this baseline nativization

was preferable to using Spanish G2P without making a distinction between foreign and native words.

Having obtained promising results with our baseline approach to nativization, a data-driven nativization was attempted by applying the pronunciation by analogy algorithm.It seemed to be the most appropriate methods as people make a wide use of analogy in order to figure out the best suitable pronunciation for an 'intruder' word in their language. This analogy is based on the orthographic form of the word as well as on its frequency of usage, pronunciation, among other factors. Analogy was explored both in orthographic and phonetic domains. In order to automatically derive analogy patterns for pronunciation a database containing nativized pronunciation was needed.

This was the reason for creation of minimalistic, phonetically-balanced training corpora for both English and Catalan nativized inclusions consisting of 1000 words each. Two additional corpora of 100 frequently used words each were created for test purposes.

The nativization results for English, obtained by using analogy only in the orthographic domain were much better that the baseline results obtained with nativization tables. However, there are stilla lot of possibilities for improvement (43.8% words correct for the English common words test set). This can be explained by the deep orthography of the English language. Even in this case, the results show very significant improvements in comparison to those obtained by direct phoneme-to-phoneme table-based mapping (NatTAB). We believe, and the numbers back us up, that the G2Pnat results obtained in the experiments are also much better than G2P results that would be obtained for the same corpus.The method based on analogy in the phonemic domain, P2Pnat, gave an improvement of approximately 12-14 percentage points (both for English and Catalan) with respect to the orthographic analogy, demonstrating the connection between the pronunciation in the source language and in the nativized one. To further explore the possibility of improvement of our multilingual system we applied the TBL algorithm, based on learning from errors, in combination with previously used nativization methods (G2Pnat, P2Pnat and NatTAB). NatTAB+TBL performed slightly better than P2Pnat. The best results were achieved using P2Pnat enhanced by the TBL algorithm (66.7%) allowing to incorporate additional information about the orthography in the source language. As this improvement was not statistically significant, the TBL algorithm was not considered for Catalan.

A perceptual test was conducted for English in order to determine the rate of acceptance of speech containing nativized utterances by the listeners. The participants were asked to rank three nativization methods, namely Spanish G2P rules, nativization tables and P2Pnat. P2Pnat showed a significant advantage over the nativization tables with 50% of *best* votes. For Catalan the results were quite similar and equally encouraging.

G2Pnat outperformed NatTAB by almost 17 percentage points, while P2Pnat, in its turn, outperformed G2Pnat by another 17 percentage points.

Both objective and subjective evaluation have shown the effectiveness of the proposed nativization method and the good rate of acceptance of the nativized speech by the listeners with different background in speech synthesis.

As a result of the work done in this thesis, steps were taken towards the solution of three important issues in up-to-date TTS systems, adaptability and reliability were improved and the multilingual challenge in the framework of Catalonia had given very promising results.

The results obtained in this chapter can be found in the following publications:

Tatyana Polyákova and Antonio Bonafonte. Transcripción fonética en un entorno plurilingüe. In *Actas de las V Jornadas en Tecnologías del Habla*, pages 207–210, Bilbao, Spain, November 2008b

Tatyana Polyákova and Antonio Bonafonte. Introducing nativization to spanish tts systems. *Speech Communication*, 53(8):1026 – 1041, October 2011. ISSN 0167-6393. doi: DOI:10.1016/j.specom.2011.05.009. URL `http://www.sciencedirect.com/science/article/pii/S0167639311000744`

6.3.1 Future work

In the future within the same line of work several issues may be studied. In this thesis we worked with standard formal pronunciation, however there are numerous speech styles such as video games, conversations between friends or family, slang, etc... Different speech styles use different phoneme definitions (allophones) and phonetic transcription.

In order for the new technologies to be able to reproduce these speech styles, the pronunciation should be specifically adapted to the speech style in question. Related to that issue we also need to adapt the pronunciation to the training voice. The algorithms developed in this thesis could be extended to use acoustic evidence from the training corpus.

As for multilingualism, so far only English and Catalan inclusions in Spanish were considered, however, the opposite problem exists and it renders English speech with Spanish inclusions even more unintelligible that English utterances with Spanish or Catalan inclusions were before the nativization was proposed. A quick example could be a GPS system that speaks English while helping you navigate through the streets of Barcelona, where some of the street names are Spanish and some are Catalan.

Another important direction can be focused on more than single language inclusions in our native language utterances. Two issues should be addressed here: how to nativize each type of inclusions without loosing in intelligibility and also study if any additional adaptation is needed due to the presence of several languages.

Bibliography

Jordi Adell, Pablo D. Agüero, and Antonio Bonafonte. Database pruning for unsupervised building of text-to-speech voices. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 889–892, Toulouse, France, May 2006.

W. Ainsworth. A system for converting english text into speech. *IEEE Transactions on Audio and Electroacoustics*, 21(3):288–290, 1973.

O. Andersen, R. Kuhn, A. Lazaridès, P. Dalsgaard, J. Haas, and E. Noth. Comparison of two tree-structured approaches for grapheme-to-phoneme conversion. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 1700–1703, Philadelphia, USA, October 1996.

P.C. Bagshaw. Phonemic transcription by analogy in text-to-speech synthesis: Novel word pronunciation and lexicon compression. *Computer Speech & Language*, 12(2):119–142, 1998.

G. Bakiri and T.G. Dietterich. Achieving high-accuracy text-to-speech with machine learning. In *In Data mining in speech synthesis*. Chapman and Hall, 1997.

D.R. Bear, S. Templeton, L.A. Helman, and T. Baren. *English learners: Reaching the highest level of English literacy*, chapter Orthographic development and learning to read in different languages, pages 71–95. International Reading Association, 2002.

J.R. Bellegarda. Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy. *Speech Communication*, 46(2):140–152, 2005.

M. Bisani and H. Ney. Investigations on joint-multigram models for grapheme-to-phoneme conversion. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 105–108, Denver, USA, September 2002.

M. Bisani and H. Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451, 2008.

A. W. Black, P. Taylor, and R. Caley. The festival speech synthesis system, 1998a. URL `http://www.cstr.ed.ac.uk/projects/festival.html`.

A.W. Black and K.A. Lenzo. Multilingual text-to-speech synthesis. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 761–764, May 2004.

A.W. Black, K. Lenzo, and V. Pagel. Issues in building general letter to sound rules. In *SSW3*, pages 77–80, 3rd ESCA Workshop on Speech Synthesis, pp. 77-80, Jenolan Caves, Australia, November 1998b.

A. Bonafonte and J. B. Marino. Language modeling using x-grams. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 394–397, Philadelphia, PA, USA, October 1996. URL `http://gps-tsc.upc.es/veu/research/pubs/download/Bon96b.ps.gz`.

A. Bonafonte, H. Höge, I. Kiss, A. Moreno, U. Ziegenhain, H. van den Heuvel, H.U. Hain, X.S. Wang, and M.N. Garcia. TC-STAR: Specifications of language resources and evaluation for speech synthesis. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, pages 311–314, May 2006a.

A. Bonafonte, J. Adell, P.D. Agüero, D. Erro, I. Esquerra, A. Moreno, J. Pérez, and T. Polyákova. The upc tts system description for the 2007 blizzard challenge. In *Proc. of the 6th ISCA Speech Synthesis Workshop (SSW6)*, Bonn, Germany, August 2007.

A. Bonafonte, A. Moreno, J. Adell, P.D. Agüero, E. Banos, D. Erro, I. Esquerra, J. Pérez, and T. Polyakova. The UPC TTS system description for the 2008 blizzard challenge. In *Proc of the Blizzard Challenge, Brisbane, Australia*, September 2008.

Antonio Bonafonte, Harald Höge, Herbert S Tropf, Asunción Moreno, Henk van der Heuvel, David Sündermann, Ute Ziegenhain, Javier Pérez, Imre Kiss, and O Jokisch. Tts baselines and specifications. *Deliverable D8 of the EU project TC-STAR "Technology and corpora for Speech to Speech Translation"(FP6-506738)*, 2005.

Antonio Bonafonte, Pablo D. Agüero, Jordi Adell, Javier Pérez, and Asunción Moreno. Ogmios: The UPC text-to-speech synthesis system for spoken translation. In *Proceedings of the TC-Star Workshop on Speech to Speech Translation*, Barcelona, Spain, June 2006b.

L. Breiman. *Classification and regression trees*. Chapman & Hall, 1984.

E. Brill. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565, 1995.

P. Brown and D. Besner. *The assembly of phonology in oral reading: A new model.* Lawrence Erlbaum Associates, Inc, 1987.

M.J. Canellada and J.K. Madsen. *Pronunciación del español.* Editorial Castalia, 1987.

D. Caseiro, L. Trancoso, L. Oliveira, and C. Viana. Grapheme-to-phone using finite-state transducers. In *Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002.*, pages 215–218, Santa Monica, USA, September 2002.

CELEX. The celex lexical database, 1995. URL `http://www.kun.nl/celex`.

S.F. Chen. Conditional and joint models for grapheme-to-phoneme conversion. In *Proc. the of European Conference on Speech Communication and Technology*, pages 2033–2036, Geneva, Switzerland, September 2003.

Y. Chen, J. You, M. Chu, Y. Zhao, and J. Wang. Identifying language origin of person names with N-grams of different units. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Toulose, France, May 2006.

N. Chomsky and M. Halle. *Sound Pattern of English.* Harper and Row, New-York, 1968.

X.F. Conde. Introducción la Fonética y fonología del Español. *Ianua. Romance Philology Journal*, Sup04, 2001. ISSN 1616-413X.

W. Daelemans and A. Van Den Bosch. Language-independent data-oriented grapheme-to-phoneme conversion. *Progress in speech synthesis*, pages 77–89, 1996.

W. Daelemans, A. Van Den Bosch, and T. Weijters. Igtree: Using trees for compression and classification in lazy learning algorithms. *Artificial Intelligence Review*, 11(1):407–423, 1997.

R.I. Damper, editor. *Data-Driven Techniques in Speech Synthesis.* Kluwer Academic Publishers, 2001.

R.I. Damper and J.F.G. Eastmond. Pronouncing text by analogy. In *Proceedings of the 16th conference on Computational linguistics*, pages 268–273, Morristown, NJ, USA, 1996.

R.I. Damper and J.F.G. Eastmond. Pronunciation by analogy: Impact of implementational choices on performance. *Language and Speech*, 40(1):1, 1997.

RI Damper, Y. Marchand, MJ Adamson, and K. Gustafson. Comparative evaluation of letter-to-sound conversion techniques for english text-to-speech synthesis. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, pages 53–58, 1998.

R.I. Damper, Y. Marchand, J.D. Marseters, and A. Bazin. Aligning Letters and Phonemes for Speech Synthesis. In *Proc. of the 5th ISCA Speech Synthesis Workshop (SSW5)*, pages 209–214, Pittsburgh, USA, June 2004.

M. De Calmès and G. Pérennou. Bdlex: a lexicon for spoken and written french. In *Proceedings of 1st International Conference on Langage Resources & Evaluation*, 1998.

B. De Mareüil, C. D'Alessandro, G. Bailly, F. Béchet, M.N. Garcia, M. Morel, R. Prudon, and J. Véronis. Pronunciation of proper names by four french grapheme-to-phoneme converters. In *Proc. the of European Conference on Speech Communication and Technology*, pages 1521–1524, Lisboa, Portugal, September 2005.

MJ Dedina and HC Nusbaum. PRONOUNCE: a program for pronunciation by analogy. *Computer speech & language*, 5:55–64, 1991.

S. Deligne, F. Yvon, and F. Bimbot. Variable-length sequence matching for phonetic transcription using joint multigrams. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 169–172, Detroit,USA, May 1995.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

Mark Donohue and Johanna Nichols. Does phoneme inventory size correlate with population size. *Linguistic Typology*, 15(2):161–170, 2011.

T. Dutoit. *An introduction to text-to-speech synthesis*. Springer Netherlands, 1997.

EF EPI. *English Proficiency Index*. Education First, 2011. URL `http://www.ef-uk.co.uk/sitecore/__/~/media/efcom/epi/pdf/EF-EPI-2011.pdf`.

H. Elovitz, R. Johnson, A. McHugh, and J. Shore. Letter-to-sound rules for automatic translation of english text to phonetics. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(6):446–459, 1976.

Susan Fitt. *Documentation and User Guide to Unisyn Lexicon and Post-Lexical Rules*. Centre for Speech Technology Research, University of Edinburgh, 2000.

A. Font Llitjós. Improving Pronunciation Accuracy of Proper Names with Language Origin Classes. Master's thesis, Masters Thesis (Technical Report: CMU-LTI-01-169), 2001.

A. Font Llitjos and A.W. Black. Knowledge of Language Origin Improves Pronunciation Accuracy of Proper Names. In *Proc. the of European Conference on Speech Communication and Technology*, pages 1919–1922, Aarborg, Denmark, September 2001.

RA Fox, JE Flege, and MJ Munro. The perception of English and Spanish vowels by native English and Spanish listeners: a multidimensional scaling analysis. *The Journal of the Acoustical Society of America*, 97(4):2540–2551, 1995.

L. Galescu and J. Allen. Bi-directional conversion between graphemes and phonemes using a joint n-gram model. In *Proc. of the 4th ISCA Speech Synthesis Workshop (SSW4)*, Perthshire,Scotland, September 2001.

AC Gimson and A. Cruttenden. *Gimson's pronunciation of English*. A Hodder Arnold Publication, 2001.

R.J. Glushko. *Interactive processes in reading*, chapter Principles for pronouncing print: The psychology of phonography, pages 61–84. Lawrence Erlbaum, 1981.

J. Häkkinen, J. Suontausta, S. Riis, and K.J. Jensen. Assessing text-to-phoneme mapping strategies in speaker independent isolated word recognition. *Speech Communication*, 41 (2-3):455–467, 2003.

R.M. Hammond. *The Sounds of Spanish: Analysis and Application (with Special Reference to American English)*. Cascadilla Press, 2001.

IPA Handbook. *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press, 1999.

E. Hartikainen, G. Maltese, A. Moreno, S. Shammass, and U. Ziegenhain. Large Lexica for Speech-to-Speech Translation: From Specification to Creation. In *Proc. the of European Conference on Speech Communication and Technology*, pages 1529–1532, Geneve, Switzerland, September 2003.

Sh. Hunnicut. Phonological rules for a text-to-speech system. *American Journal of Computational Linguistics*, pages 1–72, 1976.

Irina Illina, Dominique Fohr, Denis Jouvet, et al. Grapheme-to-phoneme conversion using conditional random fields. In *Proc. of Interspeech*, pages 2313–2316, Brighton, UK, sep 2011.

F. Jelinek. *Statistical methods for speech recognition*. the MIT Press, 1997.

Sittichai Jiampojamarn and Grzegorz Kondrak. Online discriminative training for grapheme-to-phoneme conversion. In *Proc. of Interspeech*, pages 1303–1306, Florence, Italy, aug 2009.

L. Jiang, H.W. Hon, and X. Huang. Improvements on a trainable letter-to-sound converter. In *Fifth European Conference on Speech Communication and Technology*, pages 605–608, Rhodes, Greece, September 1997.

P. Kingsbury, S. Strassel, C. McLemore, and R. MacIntyre. Callhome american english lexicon (pronlex). *Linguistic Data Consortium, Philadelphia*, 1997.

P. Ladefoged. *Vowels and consonants*. Blackwell Publishing, 2003.

S. Lewis, K. McGrath, and J. Reuppel. Language identification and language specific letter-to-sound rules. *Colorado Research in Linguistics*, 17(1):1–8, 2004.

M. Liberman. *The intonational system of English*, volume 24. Garland Pub., 1979.

A. Lindström. *English and other foreign linguistic elements in spoken Swedish: studies of productive processes and their modelling using finite-state tools*. PhD thesis, University Linköping, Linköping, Sweden, 2004.

J. Llisterri. Las tecnologías del habla para el español. *Seminario "Ciencia, Tecnología y Lengua Española: La terminología científica en español". Madrid*, 11, 2003.

J.M Llorente and L.C. Díaz Salgado. *Libro de estilo de Canal Sur TV y Canal 2 Andalucía*. Radiotelevisión de Andalucía, 2004.

Y. Marchand and R.I. Damper. A multistrategy approach to improving pronunciation by analogy. *Computational Linguistics*, 26(2):195–219, 2000.

Y. Marchand and R.I. Damper. Can syllabification improve pronunciation by analogy of english? *Natural Language Engineering*, 13(01):1–24, 2007.

J. B. Marino. Avivavoz: tecnologias para la traduccion de voz. In *Actas de las IV Jornadas en Tecnologías del Habla*, pages 285–90, Zaragoza, Spain, November 2006.

J.B. Mariño and H. Rodríguez. Proyecto aliado: Tecnologías del habla y el lenguaje para un asistente personal. *Procesamiento del lenguaje natural*, (31):305–306, 2003.

N. McCulloch, M. Bedworth, and J. Bridle. Netspeak–a re-implementation of nettalk. *Computer Speech & Language*, 2(3-4):289–302, 1987.

T. Mitchell. Decision tree learning. *Machine learning*, 414, 1997.

R. Mitten. Computer-usable version of oxford advanced learner's dictionary of current english, 1992. URL `http://ota.adhs.ac.uk`.

Grace Ngai and Radu Florian. Transformation-based learning in the fast lane. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics, 2001.

K.U. Ogbureke, P. Cahill, and J. Carson-Berndsen. Hidden markov models with context-sensitive observations for grapheme-to-phoneme conversion. In *Eleventh Annual Conference of the International Speech Communication Association*, Makuhari, Japan, sep 2010.

V. Pagel, K. Lenzo, and A.W. Black. Letter to sound rules for accented lexicon compression. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 2015–2018, Sydney, Australia, December 1998.

B. Pfister and H. Romsdorfer. Mixed-Lingual Text Analysis for Polyglot TTS Synthesis. In *Proc. the of European Conference on Speech Communication and Technology*, pages 2037–2040, Geneva, Switzerland, September 2003.

A.M.F. Planas. *Así se habla: nociones fundamentales de fonética general y española: apuntes de catalán, gallego y euskara*. Horsori Editorial, Sl, 2005.

T. Polyákova and A.Bonafonte. Main issues in grapheme-to-phoneme conversion. In *Actas del XXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, (SEPLN2005)*, pages 29–34, Granada, España, September 2005. URL `http://gps-tsc.upc.es/veu/research/pubs/download/780.pdf`.

Tatyana Polyákova and Antonio Bonafonte. Learning from errors in grapheme-to-phoneme conversion. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 2442–2445, Pittsburgh, USA, September 2006.

Tatyana Polyákova and Antonio Bonafonte. Fusion of dictionaries in voice creation and speech synthesis task. In *Proceedings of the International Conference on Speech and Computer*, Moscow, Russia, October 2007.

Tatyana Polyákova and Antonio Bonafonte. Further improvements to pronunciation by analogy. In *Actas de las V Jornadas en Tecnologías del Habla*, pages 149–152, Bilbao, Spain, November 2008a.

Tatyana Polyákova and Antonio Bonafonte. Transcripción fonética en un entorno plurilingüe. In *Actas de las V Jornadas en Tecnologías del Habla*, pages 207–210, Bilbao, Spain, November 2008b.

Tatyana Polyákova and Antonio Bonafonte. New strategies for pronunciation by analogy. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4261–4264, Taipei, Taiwan, April 2009.

Tatyana Polyákova and Antonio Bonafonte. Introducing nativization to spanish tts systems. *Speech Communication*, 53(8):1026 – 1041, October 2011. ISSN 0167-6393. doi: DOI:10.1016/j.specom.2011.05.009. URL `http://www.sciencedirect.com/science/article/pii/S0167639311000744`.

L.B. Raynolds and J.K. Uhry. The invented spellings of non-Spanish phonemes by Spanish–English bilingual and English monolingual kindergarteners. *Reading and Writing*, pages 1–19, 2009.

Real Academia Española. *Diccionario de la lengua Española*. Espasa Calpe, 1992.

T Robinson. Beep- british example pronunciations, version 1.0, 1997. URL `ftp://svr-ftp-eng.cam.ac.uk/pub/comp.speech/dictionaries/beep.tar.gz`.

D.E. Rumelhart, G.E. Hinton, and Williams R.J. Learning internal representations by error propagation. *Parallel distributed processing*, 1:318–362, 1986.

R.E. Schapire. Theoretical views of boosting. In *Proceedings of the 4th European Conference on Computational Learning Theory*, pages 1–10, 1999.

T.J. Sejnowski and C.R. Rosenberg. Parallel networks that learn to pronounce English text. *Complex systems*, 1(1):145–168, 1987.

T.J. Sejnowski and C.R. Rosenberg. Nettalk corpus, 1993. URL `<ftp://svrftp.eng.cam.ac.uk/pub/comp.speech/dictionaries>`.

B. Sigurd, M. Eeg-Olofsson, and J. van Weijer. Word length, sentence length and frequency-Zipf revisited. *Studia Linguistica*, 58(1):37–52, 2004.

T. Soonklang, R. Damper, and Y. Marchand. Multilingual pronunciation by analogy. *Natural Language Engineering*, 14(04):527–546, 2008. ISSN 1351-3249.

R. Sproat. Multilingual text analysis for text-to-speech synthesis. *Natural Language Engineering*, 2(4):369–380, 1996.

R. Sproat, A.W. Black, S. Chen, S. Kumar, M. Ostendorf, and Richards C. Normalization of non-standar words. *Computer Speech and Language*, 15:287–333, 2001.

M. Swan and B. Smith. *Learner English: A teacher's guide to interference and other problems*. Cambridge University Press, 2001.

P. Taylor. Hidden Markov Models for Grapheme to Phoneme Conversion. In *Proc. the of European Conference on Speech Communication and Technology*, pages 1973–1976, Lisboa, Portugal, September 2005.

Paul Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

E.L. Thomdike and I. Lorge. The teacher's word book of 30,000 words. *New York: Teachers College, Columbia University*, 1944.

L.M. Tomokiyo, A.W. Black, and K.A. Lenzo. Foreign accents in synthetic speech: Development and evaluation. In *Proc. of Interspeech*, pages 1469–1472, Lisboa, Portugal, September 2005.

T. Tomokiyo. Applying maximum entropy to english grapheme-to-phoneme conversion. Technical report, CMU, May 2000.

K. Torkkola. An efficient way to learn english grapheme-to-phoneme rules automatically. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 199–202, Minneapolis, USA, April 1993.

I. Trancoso, C. Viana, I. Mascarenhas, and C. Teixeira. On deriving rules for nativised pronunciation in navigation queries. In *Proc. the of European Conference on Speech Communication and Technology*, pages 195–198, Budapest, Hungary, September 1999.

Isabel Trancoso. Issues in the pronunciation of proper names: the experience of the Onomastica project. In *In Proceedings of Workshop on Integration of Language and Speech*, pages 193–209, Moscow, Russia, November 1995.

A. Van Den Bosch and W. Daelemans. Data-oriented methods for grapheme-to-phoneme conversion. In *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, pages 45–53, 1993.

Antal van den Bosch and Walter Daelemans. Data-oriented methods for grapheme-to-phoneme conversion. In *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, pages 45–53, Utrecht, The Netherlands, April 1993.

H. Van den Heuvel, B. Réveil, and J.P. Martens. Pronunciation-based ASR for names. In *Proc. of Interspeech*, pages 2991–2994, Brighton, UK, September 2009.

T. Vitale. An algorithm for high accuracy name pronunciation by parametric speech synthesizer. *Computational Linguistics*, 17(3):257–276, 1991.

R.L. Weide. The carnegie mellon pronouncing dictionary, 1998. URL `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

S. Weisler and S.P. Milekic. *Theory of language*. Bradford books, 2000.

J.C. Wells. *Accents of English: an Introduction*. Cambridge Univ Pr, 1982.

J.C. Wells. *Handbook of Standards and Resources for Spoken Language Systems*, chapter SAMPA computer readable phonetic alphabet. Mouton De Gruyter, 1997.

Melvin Stanley Whitley. *Spanish/English contrasts: A course in Spanish linguistics*. Georgetown University Press, 2002.

M.S. Yavas. *Applied English Phonology*. Blackwell Publishing, 2006.

F. Yvon. Grapheme-to-phoneme conversion using multiple unbounded overlapping chunks. In *Conf. on New Methods in Natural Language Processing*, pages 218–228, Ankara, Turkey, 1996a.

F. Yvon. *Prononcer par analogie: motivation, formalisation et evaluation*. PhD thesis, Paris Telecom, 1996b.

H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, and K. Tokuda. The hmm-based speech synthesis system (hts) version 2.0. In *Proc. of Sixth ISCA Workshop on Speech Synthesis*, pages 294–299, Germany, aug 2007.

M. Zheng, Q. Shi, W. Zhang, and L. Cai. Grapheme-to-phoneme conversion based on a fast TBL algorithm in mandarin TTS systems. *Fuzzy Systems and Knowledge Discovery*, pages 600–609, 2005.