UNIVERSIDADE DO ALGARVE

Revisão do módulo de transcrição fonética para implementação no sintetizador de fala da empresa Verbio Technologies SL

Manoela Ramalho Dias

Dissertação de

Mestrado em Processamento de Linguagem Natural e Indústrias da Língua

Trabalho efetuado sob a orientação de:

Dra. Marta Estrada Medina

Dr. Jaume Padrell

Dr. Jorge Baptista

2013

Dissertation submitted as part of the study program for the award of the Master degree in Natural Language Processing and Human Language Technology, and supported by a grant from the European Commission, Education & Culture, under the Erasmus Mundus Master Courses Program (ref. EMMC 2008-0083).

Resumen

El objetivo del presente trabajo es aportar mejoras al módulo fonético del sintetizador del habla de la empresa *Verbio Technologies SL*. La metodología del estudio consistió en analizar de manera detallada dos variantes lingüísticas del portugués brasileño: la variante de Rio de Janeiro y la variante de São Paulo, regiones económicamente desarrolladas y dónde la aplicación de un sintetizador se justificaría.

El aporte de mejoras al sistema de reglas del transcriptor de grafemas en fonemas posibilita que la salida del sintetizador sea lo más natural posible. Para eso, fue necesaria la revisión de las reglas ya existentes en el módulo fonético y también la incorporación de nuevas reglas que permiten la realización de determinados procesos fonológicos frecuentes en el portugués brasileño, como es el caso de la epéntesis vocálica.

Además del análisis de los fenómenos fonológicos que ocurren en portugués brasileño, este trabajo también contempla la Nueva Ortografía de la Lengua Portuguesa. El nuevo acuerdo ortográfico que deberá ser implementado a partir del año 2012 provoca una reestructuración de la ortografía de la lengua portuguesa. Así, muchas palabras sufrieron modificaciones y tales alteraciones deberán ser contempladas por el transcriptor grafema-fonema del sistema de síntesis del habla.

La fundamentación teórica y la revisión bibliográfica de este estudio están basados en los trabajos de los siguientes autores en el ámbito de las tecnologías del habla: Braga (2008), Braga et al. (2003), Llisterri y Martí (2002), Llisterri et al. (1999 y 2004), Pachès *et al*. (2000). En el ámbito de la nueva ortografía de la lengua portuguesa serán consideradas las bases de los cambios ortográficos decretados en 2009 y también el trabajo de Tufano (2008).

Palabras claves: 1. Fonética 2. Síntesis del habla 3. Portugués Brasileño.

Resumo Estendido

O objetivo deste trabanlho é contribuir para a melhoria da qualidade do sistema de conversão de texto em fala elaborado para o Português do Brasil e desenvolvido pela empresa *Verbio* Technologies SL. Tais modificações foram possíveis a partir da revisão minuciosa e das consequentes modificações no módulo de transcrição fonética do sintetizador.

Devido às alterações introduzidas pela Nova Ortografia do Português foram feitas modificações nas regras de transformação dos grafemas em fonemas, parte integrante do transcritor fonético que compõe o sistema desenvolvido pela empresa. O novo acordo ortográfico consiste na reestruturação ortográfica da língua portuguesa, deste modo, muitas palavras sofreram modificações e, tais alterações deverão ser abarcadas pelo transcritor grafema-fonema do sistema de síntese de fala.

Além das novas regras da ortografia portuguesa, também foi utilizado um dicionário desenvolvido pelo Centro de Pesquisa e Desenvolvimento em Telecomunicações (CPqD), versão 1.4 de maio de 2003. Este dicionário foi usado como ponto de partida para a definição dos fonemas e do subsequente desenvolvimento das novas regras.

A metodologia de estudo consistiu na análise detalhada de duas variantes linguísticas do português brasileiro: a variante falada no Rio de Janeiro e a variante falada São Paulo, regiões economicamente desenvolvidas e onde a aplicação de um sintetizador se justica.

Além da incorporação das novas regras de ortografia da língua portuguesa, foram definidas também algumas regras que contemplam determinados processos fonológicos frequentes no português brasileiro, como é o caso da epêntese vocálica.

Para a realização deste trabalho foi realizado um estágio de sete meses na empresa *Verbio* Technologies SL. de dezembro de 2011 a junho de 2012. *Verbio* Technologies SL é uma empresa especializada em tecnologia de fala (síntese de voz, reconhecimento de voz, biometria de voz e processamento de sinal acústico), localizado em Barcelona. Trata-se de uma empresa global com mais de 20 anos de pesquisa e desenvolvimento na área de tecnologia de síntese de fala e reconhecimento de voz. Além da Espanha, existem filiais na Argentina, Brasil, Chile, Colômbia, México, Paraguai, Peru e Uruguai. Desde 2001, a tecnologia desenvolvida pela *Verbio* foi enriquecida com novos recursos e produtos, tais como: *Verbio* XML, *Wordspotting*,

Verificação e Identificação, *Call Steering, Speech Analytics* e Transcrição de voz para texto.

O sintetizador desenvolvido pela *Verbio* converte automaticamente texto escrito em fala em 11 idiomas diferentes (castelhano, catalão, valenciano, basco, galego, inglês, francês, português europeu, português brasileiro, variedades do espanhol do México e da Argentina).

O software desenvolvido pela empresa utiliza a técnica de concatenação de unidades como método para a síntese do texto em fala. Desta forma, é possível obter naturalidade elevada no *output* do sistema, além de baixo custo computacional.

A técnica de concatenação de unidades utiliza um conjunto fixo de polifones (segmentos sonoros de dois ou mais fonemas) armazenados em um dicionário composto por fragmentos de audio extraídos das gravações. O inventário de unidades acústicas é composto de fragmentos de diferentes tamanhos: desde semi-sílabas até seqüências de cinco fragmentos, totalizando 82.000 unidades.

O sistema foi programado em C + + para a plataforma Windows (NT, 2000, XP, 2003 e Vista) e Linux. A interface gráfica também foi desenvolvido em C + +. Os requisitos de memória são: motor de síntese (Vox Server) 10 MB; módulo de voz 8kHz cerca de 60 MB, módulo de voz 16 KHz aproximadamente 120 MB, 30 MB de RAM, no mínimo, locutor.

Para o processamento do sinal de voz foi utilizada a técnica PSOLA (*Pitch Synchronous Overlap Add Method*), que modifica os valores de duração e frequência fundamental através da multiplicação, redução, compressão ou expansão dos períodos realizados pela abertura ou fechamento da glote.

A arquitetura do sistema possui três módulos distintos: *front-end, back-end* e *voice font*. O pré-processamento textual (*front-end*) do sintetizador é composto por: (i) um sistema de normalização textual, (ii) o transcritor grafema a fonema e (iii) módulo de processamento prosódico. Para sintetizar uma frase, o software integra suprasegmentos e segmentos do componente fônico. O módulo *back end*, por sua vez, é formado pelo motor de síntese. Por fim, o banco de dados de voz (*voice font*) está vinculado ao processo de seleção do locutor e da gravação em estúdio.

A análise textual é composta pelo processo de normalização, pelo separador de frases (informação útil para a posterior geração dos modelos prosódicos) e pelo separador de palavras. O módulo de análise fonética é composto pelo conversor grafema

a fonema e por dois dicionários diferentes: o dicionário de exceções e dicionário de abreviaturas e siglas.

O dicionário de exceção apresenta as transcrições das palavras estrangeiras que não seguem os mesmos processos fonológicos da língua em questão. O segundo dicionário apresenta as expansões das siglas e abreviaturas.

Em caso de dúvidas quanto à ortografia das palavras em português brasileiro, foi consultado sempre que possível o vocabulário VOLP (Vocabulário Ortografia da Língua Portuguesa), elaborado pela Academia Brasileira de Letras, em consonância com o novo Acordo Ortográfico.

A transcrição fonética da *Verbio* foi desenvolvido utilizando dois sistemas diferentes: (i) sistema SAMPA e (ii) sistema Segre. O alfabeto SAMPA (*Speech Assessment Methods Phonetic Alphabet*), desenvolvido por Wells (1996), foi utilizado para as transcrições fonéticas. A metodologia desta pesquisa baseia-se, principalmente, no trabalho feito por Pachès *et al*. (2000), que desenvolveu um transcritor fonético para as quatro variantes de Catalão. Segre é uma ferramenta que permite a transcrição fonética automática do texto de entrada. As regras de conversão são processados por arquivos da tabela ASCII, que, por sua vez, especificam regras segundo uma sintaxe determinada pelos contextos adjacentes de cada fonema. La transcripción fonética automática toma como entrada cualquier texto escrito en portugués brasileño y genera como salida una cadena de carateres del tipo ASCII.

A revisão das regras aplicada ao módulo *front-end* do sintetizador da *Verbio* revela que as taxas de acerto do novo transcritor fonético são de 93,18% para as palavras e 98,90% para os fonemas.

A fundamentação teórica e a revisão bibliográfica deste estudo estão baseadas nos trabalhos dos seguintes autores no âmbito das tecnologias da fala: Braga (2008), Braga et al. (2003), Llisterri y Martí (2002), Llisterri et al. (1999 y 2004), Pachès *et al*. (2000). No âmbito da nova ortografia da língua portuguesa serão consideradas as bases das mudanças ortográficas decretadas em 2009 e o trabalho de Tufano (2008).

Palavras-chaves: 1. Fonética 2. Síntese de fala 3. Português Brasileiro.

Agradecimientos

Agradezco primeramente a las personas que depositaron confianza en este proyecto de trabajo, apoyándome e incentivándome constantemente: mi directora académica la Dra. Marta Estrada Medina y al Dr. Jaume Padrell, mi director de prácticas en la empresa *Verbio Technologies SL*. Agradezco igualmente a Dr. Jorge Baptista.

Me gustaría agradecer también a la Dra. Angels Catena Rodulfo, Coordinadora del Máster en Procesamiento del Lenguaje Natural y Tecnologías Lingüísticas de la Universidad Autónoma de Barcelona por las ayuda con los procedimientos burocráticos. Quiero agradecer también al programa *Erasmus Mundus* el apoyo financiero concedido para el desarrollo de este trabajo, con especial mención a Gabriel Sekunda y a la Dra. Sylviane Cardey-Greenfield.

De forma muy particular, quiero agradecer a Antonio Terradas – director general de la empresa *Verbio Tecnologies SL* – por permitirme la realización de las prácticas en la empresa. Me gustaría agradecer también a Jayme Nigri, director de *Verbio* en Brasil por la ayuda. Agradezco también a Huc Castells y a Guillem Vila por su rápida respuesta a mis dudas y preguntas, por la colaboración en las discusiones sobre el trabajo, por indicarme material bibliográfico y por las palabras de ánimos. De igual manera, me gustaría agradecer a todos los compañeros de trabajo Daniel Parera Pérez, Sonia Carbona, Arnau Padró y David Font por ayudarme cuando necesité.

Agradezco también a mis compañeros de curso Katherin Pérez Rojas, Sara Rodríguez Fernández, Gabriel André Iglesias, Cristina Marzo, Joan Pahisa, Kalina Mihaylova, Natacha Komarova, Iuliana Zodila, Cláudia Dias de Barros, Camila Rizzotti, Bozidar Bukilick, Aneta Rafajlovska, Jacopo Ottaviani, Iacer Calixto, Elizabeth Rodrigues, MariaSol Ferrer, José Maca, Lucas Nunes, Pedro Balage, Soledad López Gambino y Gaia Paixão por la ayuda prestada. Además, quiero agradecer igualmente a los amigos Lucero Munguía, Veronica Cordoba, Nathalie Marcela Cerón, Claudia Trejo, Reden Valencia Libo-on, Cristina Sorocovici por escucharme cuando más lo necesité. Y a todos los demás compañeros que me han acompañado en este periodo.

Finalmente, agradezco a toda mi familia y especialmente a mi madre Ana Carolina, a mis primos Gabriel, Amélia, Daniel y a mis amigos de Brasil: Flávio, Fernanda, Adriana, Érika por ayudarme y animarme continuamente a pesar de la distancia.

Contenido

Lista de figuras

Lista de gráficos

Lista de tablas

Lista de siglas y símbolos

ASCII – *American Standard Coding for Information Interchange*

CTH – Convertidor de texto en habla

G2P – *Grapheme-to-phoneme*

HMM – *Hidden Markov Models*

IPA – *International Phonetic Alphabet*

PB – Portugués Brasileño

PE – Portugués Europeo

PLN – Procesamiento de lenguaje natural

POS – *Part-of-speech*

PSOLA – *Pitch Synchronous Overlap Add Method*

SAMPA – *Speech Assessment Methods Phonetic Alphabet*

SP – São Paulo

TTS – *Text-to-speech*

TTP – *Text-to-phonem*

T2P – *Text-to-phonem*

RJ – Rio de Janeiro

VOLP - Vocabulario Ortográfico de la Lengua Portuguesa

Otras convenciones

C – Consonante

V – Vocal

'V – Vocal acentuada

<a> – Representación ortográfica

/a/ – Representación fonémica

[a] – Realización fonética

Capítulo 1

1. Introducción

Actualmente, el núcleo de las tecnologías lingüísticas está constituido por dos procesos distintos: el procesamiento del habla y el procesamiento del lenguaje natural (PLN). El procesamiento del habla es un área de trabajo en el campo de las tecnologías lingüísticas y se orienta al análisis del procesamiento de la señal del habla.

El procesamiento de lenguaje natural (PLN), a su vez, es una disciplina definida por la utilización de saberes lingüísticos, tanto para establecer la comunicación entre humanos y sistemas operacionales como para aportar mejoras en la comunicación entre humanos (Santos, 2001). Algunos ejemplos de aplicaciones en las que intervienen el PLN y el procesamiento del habla son los sistemas de síntesis de texto en habla, los sistemas de diálogo y los reconocedores del habla.

La era digital ha convertido los sistemas de conversión de texto en habla, CTH – también conocidos como *Text-to-Speech*, TTS – en uno de los sistemas más significativos en la interfaz de comunicación entre hombre y máquina. Las mejoras científicas para alcanzar el objetivo de desarrollar *software* que se aproxime al habla humana de la manera más natural posible son continuas. A parte de la conversión de texto a habla, cabe destacar igualmente los sistemas de reconocimiento del habla y los sistemas de diálogo.

Un sistema de síntesis del habla puede ser definido como un modelo computacional que genera habla automáticamente a partir de un texto de entrada, es decir, sistema que parte de un soporte escrito de una determinada lengua con el propósito de generar el componente oral. Para que este proceso sea realizado es necesaria la integración entre los conocimientos lingüísticos y los conocimientos sobre el sistema informático.

El proceso de generación de TTS se lleva a cabo desde la interdisciplinariedad de las ciencias del habla, aplicando los conocimientos lingüísticos al lenguaje de programación. Así, por ejemplo, las expresiones regulares estandarizan el lenguaje informático y, a partir de este patrón, es posible traducir las cuestiones relacionadas con la fonética acústica, la fonología, la sintaxis, la morfología, la ortografía en un lenguaje legible por el sistema informático.

La integración de los conocimientos lingüísticos e informáticos es necesaria para la implementación de un sistema que pueda convertir de texto en habla de calidad. El

modelo de lenguaje y la gramática son las bases para el desarrollo de dichos sistemas y dependen de un análisis previo de la lengua en cuestión.

Llisterri et al. (2004:146) señalan que:

> *[…] el lingüista tiene un papel central en el diseño y en el desarrollo de un convertidor, pues es quien aporta datos específicos de cada lengua, propone soluciones con conocimiento de causa a los problemas propios de su ámbito y, en las etas finales, evalúa el resultado en función de su familiaridad con los patrones habituales que esperan encontrar los usuarios de una determinada comunidad lingüística.* Llisterri *et al*. (2004:146)

Los sistemas de conversión de texto en habla generan habla sintética a partir de textos producidos por un usuario, o por un aplicativo, de modo que el sistema realiza la emisión de sonidos del habla a partir de una representación textual.

1.1 Objetivos

El objetivo de la presente investigación es contribuir a la mejora de la calidad del sistema de conversión de texto en habla en portugués brasileño desarrollado por la empresa *Verbio Technologies SL*[1] aportando propuestas de modificación en el transcriptor fonético y en el módulo prosódico del sistema de síntesis de Verbio. Debido a los cambios ortográficos introducidos por la nueva normativa ortográfica del portugués fueron realizadas modificaciones en las reglas que transforman grafema en fonema del sistema de sintesis de Verbio. En la salida del sintetizador fueron tratadas dos variedades distintas del portugues brasileño: el habla estándard de las capitales São Paulo y Rio de Janeiro.

Además de la incorporación de las nuevas reglas ortográficas de la lengua portuguesa, fue contemplado el analisis del modulo prosodico. En esta investigación nos interesa las variedades diatópicas, es decir, las variaciones que se producen en determinadas regiones geográficas. Estas variantes pueden ser definidas como realizaciones lingüísticas de agrupaciones humanos que pueden estar asociadas al acento, al ritmo del habla y a la elección lexical (Zágari: 2005).

[1] Puede obtenerse más información sobre *Verbio Technologies SL* en http://www.*verbio*.com/.

Para la realización de este trabajo fueron realizadas prácticas en la empresa *Verbio Technologies SL* desde diciembre de 2011 hasta junio de 2012. *Verbio Technologies SL* es una empresa de tecnologías del habla (síntesis del habla, reconocimiento del habla, biometría del habla y procesado de señal acustica) ubicada en Barcelona. La empresa, actualmente, colabora con la Universidad Autónoma de Barcelona (UAB) – además de otras universidades – y dispone de un equipo autónomo de I+D.

1.2 Motivación

El interés comercial y científico que las aplicaciones de las tecnologías del habla han despertado en el mercado actual constituye la principal motivación de este trabajo. En efecto, el desarollo creciente de la economía brasileña ha comportado que empresas e instituciones de investigación científica busquen cada vez más aportar mejoras en los *software* de tecnologías del habla producidos en lengua portuguesa.

Otro criterio para la inversión de las tecnologías del habla en portugués brasileño es el elevado número de habitantes en Brasil. Según datos colectados por el Instituto Brasileño de Geografía Estadística (IBGE) del año 2010, consta en el país una población de 190.732.694 personas. La región sudeste es la más poblada de Brasil, con 80.353.724 personas. De entre las unidades de la federación, São Paulo lidera con 41.252.160 habitantes, Minas Gerais ocupa la segunda posición con 19.595.309 personas y Rio de Janeiro ocupa la tercera posición con 15.993.583 personas.[2] Estas estadísticas asociadas al desarrollo económico brasileño indican el crecimiento del poder adquisitivo de la población y el consecuente consumo de estas nuevas tecnologías del habla.

La metodología de esta investigación está enfocada al analisis de las variantes lingüísticas de Rio de Janeiro y de São Paulo, regiones económicamente desarrolladas y con alta tasa de la población, dónde la utilización de un sintetizador del habla en portugués brasileño se justifica.

Otro incentivo para la realización de este estudio es la falta de trabajos más amplios en portugués brasileño en el ámbito de la síntesis del habla, especialmente cuando se compara con la gran cantidad de estudios realizadas para el inglés o francés (Braga, 2008: 4). Este hecho puede explicarse por razones históricas, políticas y

[2] La información detallada sobre los datos del IBGE se encuentra en http://www.ibge.gov.br/home/estatistica/populacao/censo2010/sinopse/default_sinopse.shtm [01/06/ 2012].

económicas del inglés y francés, que no se dan en el caso del portugués o el español, aunque sean los idiomas más hablados el mundo occidental como lengua materna y oficial, después del inglés. A pesar del avance científico en el ámbito de la síntesis del habla en portugués brasileño, persiste todavía una falta de recursos, de inversión económica y de formación.

1. 3 Aplicaciones

Según Llisterri y Martí (2002: 20) “las tecnologías del habla fueron proyectadas para facilitar la interacción entre hombre y máquina”. Una de las utilidades descritas por los autores es la realización de tareas simultáneas, liberando las manos y la vista del usuario, como es el caso, por ejemplo, de la producción oral de textos escritos, que permite escuchar los textos en lugar de leerlos.

Las posibilidades de aplicación de los sistemas de síntesis del habla, son muchas y van desde las interfaces para ordenadores domésticos, terminales portátiles, como: teléfonos móviles, *Personal Digital Assistant* (PDAs), *Tablets*, *Smartphones*, *E-books*, *GPS* hasta las centrales telefónicas.

En las utilizaciones domesticas, el sintetizador puede reproducir en habla sintética mensajes personalizados en páginas *web*, como: recordatorios, instrucciones de ayuda, correos electrónicos, noticias, informaciones meteorológicas, páginas amarillas, información sobre itinerarios, citas médicas, horóscopos, sitios web bancarios (*e-banking*), *chats* de charla virtual o en los sistemas la búsqueda y recuperación de informaciones. Los sintetizadores del habla pueden ser aplicados también en los sistemas de GPS (*Global Positioning System*) instalados en los coches para ubicación exacta de un determinado sitio sea realizada por información auditiva, lo que puede configurar mayor seguridad en la conducción.

Además de las utilizaciones domésticas, hay también las aplicaciones comerciales e industriales. De entre las aplicaciones comerciales e industriales pueden ser mencionadas: telefonía, multimedia, Internet, procesos de enseñanza y aprendizaje de lenguas, discapacitación, terminales portátiles.

En el área de telefonía existen, actualmente, inúmeras compañías de atención al cliente, los llamados *call centers* especializadas en la venta, instalación, manutención y prestación de servicios telefónicos. Muchas de estas empresas utilizan el sistema de que CTIs (*Computer Telephony Integration*) también conocidos como "integración de telefonía informática", canal de comunicación entre empresa y clientes a través de un

sistema informático destinado a la interacción entre una llamada telefónica y un ordenador. Un ejemplo de este tipo de interacción es la llamada *respuesta de voz interactiva* (***I**nteractive **V**oice **R**esponse,* IVR*)* o unidad de respuesta de voz (*Voice Response Unit,* VRU). Se trata de sistemas telefónicos en los cuales el ordenador haz el reconocimiento de respuestas simples, del tipo "sí", "no" a partir de preguntas realizadas por el usuario. Es un sistema automatizado de respuesta interactiva, orientado a entregar o capturar las informaciones por medio del teléfono, permitiendo el acceso a determinados servicios u operaciones.

Otro ejemplo de aplicación son las centrales telefónicas del tipo PABX (sigla de *Private Automatic Branch eXchange*) utilizados en ambientes corporativos donde las conexiones son usadas solo internamente por los usuarios de determinada empresa.

En los recursos multimedia, el sistema puede ser empleado (i) en las presentaciones corporativas, escolares o académicas; (ii) en los servicios de asistencia virtual, en la cual el habla aparece concatenada a un rostro generado por ordenador; (iii) en los quioscos digitales o portales de voz que ofrecen informaciones turísticas, direcciones, ubicación de hoteles, restaurantes, estaciones del metro, trenes, aeropuertos de entre otros; (iv) en CD's de presentación genérica y cambiante y (v) catálogos.

El internet también es un medio de aplicación de un sintetizador de habla, indicado por ejemplo para la lectura de los mensajes personalizados en páginas *web* con información de última hora. Otras aplicaciones serian los recordatorios, las locuciones de instrucciones de ayuda, lectores de correo electrónico o noticias. Otros ejemplos son proveedores de internet que fornecen servicios del tipo Voip o Voz sobre IP (*Voice over Internet Protocol*), es decir, tecnologías que permiten el enrutamiento de las conversaciones humanas por intermedio de la internet o de una red de ordenadores.

En el ámbito escolar o académico puede ser de utilidad también como soporte en los procesos de enseñanza y aprendizaje de lenguas, como: traductores automáticos, diccionarios electrónicos (bilingües o monolingües) que auxilian en la correcta pronunciación de las palabras del idioma estudiado. Además, un sintetizador del habla puede ser una herramienta en el proceso de alfabetización de niños o adultos cuando es utilizado en la lectura de textos.

El uso de sintetizadores del habla también permite incrementar la accesibilidad del ordenador para usuarios con habilidades restringidas o discapacitados, como es el caso del físico Stephen William Hawking, que utiliza un sintetizador del habla para seleccionar palabras, formando frases y, así, comunicarse. Además, los sistemas de

síntesis del habla pueden facilitar la inclusión digital de las personas discapacitadas visualmente.

El trabajo desarrollado por Condado (2009: 4) es un ejemplo de la aplicación de un sistema de síntesis del habla que permite a los discapacitados en el habla efectuar llamadas telefónicas. El sistema desarrollado por Condado engloba diferentes tecnologías de síntesis del habla, Voz sobre IP (VoIP) y métodos de interacción para las personas con limitaciones motoras, como es el caso de los portadores de parálisis cerebral que además de los problemas del habla también presentan dificultades motoras. Tales personas necesitan de una interfaz con teclado virtual, métodos de aceleración de la escritura como *trackballs, joysticks, touchpads,* pantallas táctiles, micrófonos, *Sip-and-puff, switches, webcams,* sistemas de *eye-track*ing, sistemas de *head-tracking* y sistemas de reconocimiento del habla.

Capítulo 2

2. Conceptos y Revisión Bibliográfica

En este capítulo se presentará la arquitectura general y el estado de la cuestión en el ámbito de la síntesis del habla, así como los diferentes métodos utilizados para la conversión de texto en habla. Serán discutidas también las descripciones de las variantes de São Paulo y Rio de Janeiro.

2.1 Arquitectura general

Tal como lo señalan Lliterri *et al*. (1999: 459) todo sistema de síntesis debe comprender un módulo destinado al tratamiento lingüístico del texto que se pretende convertir. Del funcionamiento de este módulo depende, en gran medida, de la calidad del habla sintetizada resultante del proceso (Texeira, 2003). Así,

> *Aunque la arquitectura concreta de cada sistema puede variar, el tratamiento linguistico se realiza en los módulos de pre-procesamiento del texto, transcripción fonética automática, analisis morfológico y sintáctico y en el módulo prosódico.* (Lliterri *et al*., 1999: 459)

A este primer bloque común a todos los sistemas de conversión conocido como el módulo de pre-procesamiento textual o (*front-end*), se añaden dos módulos más: el motor de síntesis (*back-end*) y la base de datos del habla (*voice font*) (Braga, 2008: 25).

En el bloque *front-end* se realiza el etiquetado fonético del texto partiendo de su análisis lingüístico, por lo que se trata de un módulo altamente dependiente de la lengua estudiada. A su vez, el *front-end* está compuesto por tres componentes distintos: (i) el análisis textual; (ii) el análisis fonético y (iii) el análisis y generación prosódica.

El pre-procesamiento textual (o normalización) adapta el input textual para su posterior síntesis (Trilla, 2009: 1). Para que un sintetizador pueda operar sobre un texto escrito es necesario que la representación ortográfica sea “pronunciable”. Así, en este módulo, deben tratarse aspectos como la expansión de abreviaturas, siglas, fechas, horas, números de teléfono, correos electrónicos, números cardinales, ordinales etc.

El análisis fonético convierte el texto en etiquetas de base fonética partiendo de análisis lingüísticos. Este bloque, por lo tanto, está relacionado intrínsecamente con las especificidades de la lengua estudiada. El input de un sistema de síntesis está

constituido, por lo tanto, por la conversión de la ortografía para el sonido (Braga *et al*., 2003: 1356).

El componente de análisis y generación prosódica controla la intensidad, la duración, la entonación (contorno de la frecuencia fundamental, F0), además de la inserción de pausas para el establecimiento del ritmo.

Por otra parte, en el bloque denominado *back-end* tiene lugar la conversión del etiquetado fonético en señal acústica, proceso que como destaca Braga (2008: 172) resulta muy costoso dada la lentitud del procedimiento y el alto grado de especialización de las tareas implicadas como la selección del locutor, de los técnicos de grabación y de edición, entre otras.

Finalmente, la base de datos de voz o *voice font* depende exclusivamente de las técnicas de síntesis utilizadas. Braga (2008: 172) afirma que el *voice font* es un proceso lento que abarca varias tareas especializadas, como la selección del locutor, técnicos de grabación y edición. Se trata de un recurso dispendioso que está relacionado exclusivamente de las técnicas de síntesis utilizadas. La síntesis final trata de reproducir el comportamiento fonético del hablante tanto en los aspectos segmentales como en los aspectos suprasegmental.

En la figura 1. (Teixeira *et al*., 2003) se representan los procesamientos existentes en la arquitectura general de cualquier sistema de síntesis del habla:

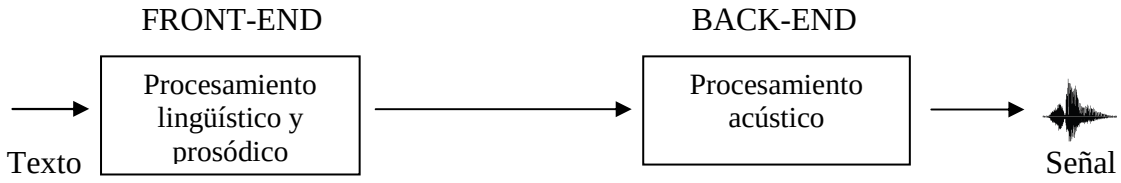

Figura 1. Módulos del procesamiento de un sistema TTS

En este trabajo se prestará mayor atención al procesamiento prosódico incluido en el módulo lingüístico-prosódico (el *front-end*). Según el trabajo de Teixeira *et al*. (2003) esta etapa de un sistema TTS determina la información segmental y la información suprasegmental a partir del texto. La información segmental está asociada a la cadena sonora compuesta por las representaciones abstractas (los fonemas). La información suprasegmental, a su vez, está asociada a los parámetros prosódicos intensidad, duración, ritmo y frecuencia fundamental (F0).

Los módulos del procesamiento lingüístico-prosódico están organizados de la manera siguiente (Teixeira *et al*., 2003 y Hentz, 2009):

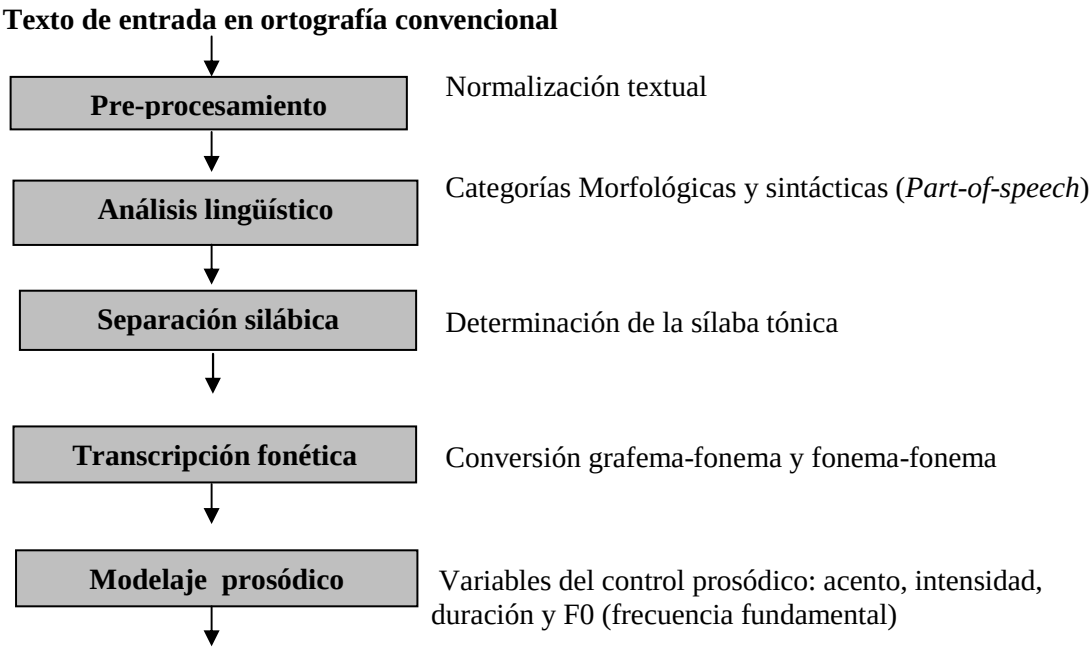

Figura 2. Módulos del procesamiento lingüístico-prosódico

De acuerdo con la figura 2. los bloques del procesamiento lingüístico-prosódico son los siguientes: (i) el pre-procesamiento del texto y (ii) el análisis lingüístico, (iii) la transcripción fonética y (iv) el procesamiento prosódico.

La naturalidad de las expresiones del habla producida por los módulos de procesamiento de señales está estrechamente vinculada a la realización de los módulos anteriores del procesamiento del texto.

A pesar de haber muchos trabajos desarrollados para la síntesis del portugués brasileño, todavía hay problemas en el *front-end* en cuanto a la lectura de extranjerismos, la resolución de ambigüedad de palabras homógrafas y en el proceso de conversión grafema-fonema.

2.1.1 Pre-procesamiento textual

Según Trilla (2009:1) el proceso de pre-procesamiento del texto (normalización textual) adapta el input textual a fin de ser sintetizado. Llisterri *et al*. (2004: 4) reiteran que para un convertidor de texto en habla operar sobre un texto escrito es necesario que este texto sea normalizado, es decir, que la representación ortográfica sea leida.

Entre los aspectos relacionados a la normalización textual pueden ser citados: la expansión de las abreviaciones, de las siglas y de las expresiones numéricas.

El primer paso para la conversión de un número en su respectiva expansión es reconocer si se trata de un número cardinal, de un número ordinal, de un número de teléfono, de una fecha, de una hora, de un número que forma una sigla. También es necesario el reconocimiento de las abreviaciones (como: Sr. Profª) y de las siglas (como: ONU, para la expresión "Organización de las Naciones Unidas").

Según Braga et al. (2003: 1354) para la identificación de los elementos referidos es necesaria una técnica de etiquetaje previa, es decir, es necesario la elaboración de algoritmos cuyo *input* debe ser el elemento (la expresión numérica, la sigla o abreviatura) y el *output* debe ser una representación por extenso. Para eso, es común la elaboración de tablas que relacionan la representación simbólica con su respectiva expansión "pronunciable", para después sean aplicadas las reglas de la gramática.

2.1.2 El análisis lingüístico del texto

Para el desarollo de un sintetizador del habla es necesario realizar un análisis linguistico del texto de entrada en el ambito morfológico, sintáctico, semántico que auxilian en la desambiguación del texto.

Hay mejoras en el proceso de conversión de texto en habla si al sistema es incorporado un categorizador y un procesador morfológico (también llamados POS *taggers*). La función de estos descriptores morfológicos es asignar una clase de palabra (adjetivo, verbo, nombre, adverbio, preposición etc) a cada *token* (unidad etiquetada) y reconocer la estructura interna de las palabras por medio del análisis de los morfemas que componen la palabra. Los etiquetadores de la parte del discurso tienen que tratar las palabras que no constan en el vocabulario y las palabras con etiquetas ambiguas.

La tokenización (segmentación) es el proceso por lo cual se realiza la separación de las unidades (palabras) que componen el texto. Normalmente el texto es dividido a partir de los espacios en blanco y de las marcas de puntuación de la frase. Este proceso es realizado con éxito con un programa de análisis lingüístico.

Retomando las palabras de Llisterri et al. (2004: 8) la principal utilidad de incorporación de un procesador morfológico en un sistema de conversión de texto en habla es que algunas palabras dependen de la estructura morfológica para la determinación de su representación fonética.

2.1.3 Separação silábica

En esta etapa, hay la división silábica para determinar la el acento de cada palabra. Las palabras que contienen tildes ortográficas ya tienen su sílaba particular. Por la otra parte, es necesario el uso de un conjunto de reglas, cuando las palabras no llevan tilde gráfica. Además de determinar la sílaba acentuada de cada palabra, es necesario clasificar las palabras según el tono de la frase, teniendo en cuenta, principalmente, en las palabras de contenido como verbos, sustantivos y adjetivos.

2.1.4 La conversión grafema-fonema

El módulo de transcripción fonética trata de la conversión de grafema en fonema. En esta etapa hay la asignación del conjunto fonético más adecuado en la secuencia de unidades (*tokens*).

Según Llisterri et al. (1999: 459) "un transcriptor es un algoritmo que transforma una cadena de caracteres[3] ortográficos en una cadena de caracteres fonéticos". En este módulo del procesamiento lingüístico se incluyen las descripciones de las relaciones entre los grafemas (caracteres ortográficos) y los fonemas (representación acústica).

Silva (2002: 126-127) distingue entre fones y fonemas, considerando estos últimos como "sonidos de una misma lengua que tienen valor distintivo" y los primeros como "segmentos encontrados en el cuadro fonético" (Silva, 2002: 126-127). Los fones deben ser transcritos entre corchetes [m] y los fonemas son transcritos entre barras inclinadas, por ejemplo /m/. Por otro lado, cuando dos segmentos no son caracterizados como distintivos, hay que buscar evidencias para caracterizarlos como alófonos. Estos alófonos identificados por medio de la distribución complementar.

En la etapa de transcripción textual existen dos procedimientos complementarios: la formulación de las reglas y el empleo del diccionario, cuando no hay la correspondencia entre el grafema y la adecuada pronunciación de un sistema lingüístico (Llisterri et al. 2004: 159). Esta falta de correspondencia entre grafema e fonema se da principalmente, en la transcripción de palabras extranjeras. Para que las

[3] *Token* es el término utilizado en el procesamiento del lenguaje natural para una cadena de caracteres.

unidades de síntesis sean seleccionadas correctamente, es necesaria la definición del modelo de lengua que debe ser empleado y la pronunciación exacta de estos extranjerismos (Llisterri et al., 2004: 159).

Braga et al. (2003: 1351) considera que pese a la ortografía del portugués brasileño tenga sido inicialmente basada en una relación alfabética univoca entre fono y grafema, es correcto decir, que actualmente esa unicidad no se verifica, pues en la lengua oral se observan algunos procesos fonológicos que algunas veces son desconsiderados por el carácter conservador de la escrita. Por consecuencia de esta no unicidad entre grafema y fonema surgen algunos desafíos que deben ser solucionados en proceso de conversión de texto en habla.

2. 1.5 El módulo prosódico

Según Braga et al. (2003: 1356) el input de un sistema de síntesis también es constituido por la conversión de la ortografía del texto para el sonido correspondiente. Uno de los componentes fundamentales para la generación del habla sintética es la prosodia, campo de estudios que abarca los procesos suprasegmentales, es decir, la intensidad, la duración, la entonación (por medio del contorno de la frecuencia fundamental, F0), además de la inserción de pausas para el establecimiento del ritmo.

Las funciones morfológicas y de las clases de palabras (POS) determinadas anteriormente son esenciales en este módulo, pues el acento léxico es definido en la mayor parte de las veces por la sílaba pesada (fenómeno fonológico lo cual trata determinados tipos de sílabas como más pesadas que otras).

El análisis lingüístico del texto es uno de los módulos de la manipulación prosódica en sistemas de síntesis del habla, conforme atesta Braga et al. (2003: 1356).

2.2 Cronología

Según Braga (2008:14), la historia de la síntesis del habla en el siglo XX depende de la evolución de los métodos de descodificación del texto en voz. Los métodos empleados en el área de síntesis del habla cambian de manera significativa al pasar de los años, conforme al desarrollo de nuevas técnicas.

Barbosa (1999: 25) considera que hay dos líneas de investigación[4] muy definidas para la posibilidad de realización de la síntesis del habla partiendo del texto. La primera línea llamada *faire-semblant* (o “hacer lo más parecido posible”) tiene como objetivo

[4] Las líneas de investigación presentadas son una adaptación de los términos propuestos por el *Institut de la Communication Parlée* (IPC, 1994, apud Barbosa, 1999).

principal reproducir una señal acústica de manera que se parezca a la señal del habla. El segundo enfoque llamado *faire-comme* (o "hacer como si fuera") busca obtener la señal acústica reproduciendo el mecanismo de fonación que permita imitar su funcionamiento en los seres humanos.

Así, en los años 80 existían los llamados métodos de primera generación de la síntesis por formantes y de síntesis articulatoria. A partir de los años 90, los llamados métodos de segunda generación, la mayoría de los sintetizadores empiezan a basarse en la técnica de concatenación de unidades siendo la más recién técnica elaborada es la síntesis mediante HMM, basados en métodos estadísticos. Braga cita Taylor (2007: 415, apud Braga, 2008) y explica que la principal diferencia entre las dos técnicas es que las primeras son construidas partiendo de los corpus de habla, al contrario de las técnicas de segunda generación, que no se generan a partir de corpus.

De entre los sintetizadores comerciales de segunda generación, pueden ser citados: Nuance,[5] IBM,[6] Acapela,[7] AT&T Labs,[8] Cepstral,[9] Microsoft,[10] y Cereproc.[11] De entre los sintetizadores desarrollados en Brasil por empresas o instituciones brasileñas podemos citar: Aiuruetê,[12] LINSE,[13] HMMs84 (Maia *et al.*, 2006; Maia, 2006) y VOCALIZE[14].

Aiuruetê es un sistema de conversión de texto en habla por concatenación de unidades, producido en el Laboratorio de Procesamiento Digital del Habla del Departamento Comunicaciones de la Universidad Estatal de Campinas (LPDF-DECOM-UNICAMP). Este *software* fue desarrollado desde 1991 hasta 2003, por los Profs. Eleonora Cavalcante Albano, Fábio Violaro y Plinio Barbosa. De entre las publicaciones sobre podemos citar: Violaro & Böeffard (1994 e 1998); Gomes (1998); Barbosa *et al.* (1999); Simões *et al.* (2000).

El inventario de unidades del sintetizador AIURUETÊ está compuesto por unidades de distintos tamaños: desde semi-sílabas hasta secuencias de 5 fragmentos,

[5] Disponible en: http://www.nuance.com/index.htm [03/04/2013].
[6] Disponible en: http://www.research.ibm.com/tts/ [03/04/2013].
[7] Disponible en: http://www.acapela-group.com/text-to-speech-interactive-demo.html [03/04/2013].
[8] Disponible en: http://www.research.att.com/projects/Natural_Voices/index.html?fbid=9TY9-e4V73W [03/04/2013].
[9] Disponible en: http://www.cepstral.com/ [03/04/2013].
[10] Disponible en: http://www.microsoft.com/en-us/Tellme/technology/ [03/04/2013].
[11] Disponible en: http://www.cereproc.com/es [03/04/2013].
[12] Disponible en: http://www.decom.fee.unicamp.br/lpdf/ [03/04/2013].
[13] Disponible en: http://www.linse.ufsc.br/skel1.php?parent=desenvolvimento&section_id=26&language=pt-BR [03/04/2013].
[14] Disponible en: http://www.e-vocalize.com.br/ [03/04/2013].

totalizando 2500 unidades. Estudios recientes han enseñado que los polifonos no son la opción más adecuada en la implementación prosódica de un convertidor TTS.

Barbosa et al. (1999, apud Silva 2004: 12) relata que el módulo de transcripción grafema-fonema del AIURUETÊ es formado por un pre-procesador, responsable por la expansión de las siglas, abreviaturas y el Ortofon, que convierte la salida del pre-procesador en una representación fonética simplificada del texto. Según los autores, este módulo presenta una tasa de error de cerca de 4%, que puede ser tratado por medio del uso de diccionarios de excepciones.

El AIURUETÊ determina la duración de cada fono a partir de un modelo de duración basado en el modelo de Campbell (1992, *apud* Silva, 2004:13). La propuesta de Campbel (1992, *apud* Silva: 13) utiliza unidades de tamaño silábico como unidades básicas de programación rítmica. Hay también una red neuronal que obtiene la duración de las unidades a partir de una representación fonológica del texto. En seguida, la duración de cada fonema es generada por medio de un modelo de repartición que sigue un principio de elasticidad fuerte. Para el contorno de F0 se utiliza una simple declinación declarativa.

El sintetizador concatenativo AIURUETÊ ha pasado por sucesivos desarrollos desde el proyecto inicial propuesto en 1991. De entre los trabajos realizados pueden ser destacados: Violaro & Böeffard (1998), Barbosa *et al* (1999) y Simões *et al* (2000).

El sistema de síntesis del habla de LINSE (Nicodem *et al.*, 2005; Nicodem *et al.*, 2007), llamado ORADOR es basado en unidades de tamaño variable, estas unidades pueden estar constituidos de un solo fonema o unidades aún mayores, como sílabas, palabras, oraciones cortas y frases. El sintetizador HMMs84 desarrollado en conjunto con el *Nagoya Institute of Technology* (Maia *et al.*, 2006; Maia, 2006).

Vocalize – *Speech and Language Technology Solutions* es una spin-off de la Universidad Estatal de Campinas, SP, Brasil (UNICAMP). Se trata de una empresa con especialización en tecnologías del habla y lenguaje, aportando soluciones al mercado en la conversión de texto en habla, en el reconocimiento de voz y en el diálogo entre el hombre y la máquina.

De entre los sintetizadores de habla desarrollados para el PE pueden ser citados el sintetizador por formantes DIXI (Oliveira, 1996), actualmente llamado Tecnovoz y el Multivox (Teixeira, 1995; Teixeira & Freitas 1998). En el ámbito académico fueron desarrollados el sintetizador de base articulatoria (Teixeira, 2000), los sintetizadores por

concatenación de unidades (Barros, 2001; Carvalho *et al.*, 2003) y el sintetizador basado en HMMs (Barros *et al.*, 2005).

Serán presentados a continuación los fundamentos teóricos acerca de los métodos desarrollados para los sistemas de síntesis del habla.

2.2.1 La síntesis por formantes

La síntesis basada en reglas, según Barbosa (1999: 26), también utiliza un vocabulario ilimitado para la generación de un sonido a partir de un texto cualquiera.

Ese método parte de una descripción detallada de las reglas que dirigen los movimientos de los formantes (sobre todo durante las transiciones entre segmentos) presentes en la señal del habla que se genera, lo que caracteriza acústicamente la dinámica de la fonación. La señal de habla es posteriormente generada por un sintetizador de formantes.

2.2.2 La síntesis articulatoria

Conforme Barbosa (1999: 25-26) el *faire-comme* o "hacer como si fuera" es realizado por la síntesis articulatoria. La síntesis articulatoria tiene como objetivo obtener el mensaje sonoro de la manera como es realizado por el tracto vocal.

El desarrollo de este tipo de tecnología es posible gracias a los estudios de dinámica de los articuladores abarcados por la fonación de las fuentes sonoras (a partir del movimiento de las cuerdas vocales, de los efectos de turbulencia, de los ruidos de fricción) y por la consecuente simulación de estos fenómenos en el ordenador.

La implementación de este método solamente es posible por el avance en los estudios de la dinámica de los articuladores involucrados con el proceso de fonación, como el control de los movimientos de las cuerdas vocales y de los ruidos de fricción. Además, se toman en consideración el papel de la percepción en la selección de los gestos articulatorios y consecuentemente la simulación digital de esos fenómenos.

2.2.3 La síntesis por concatenación

Según Klatt (1987, apud Barbosa 1999) la síntesis concatenativa en el dominio del tiempo (*Time Domain Synthesis*) es un método que utiliza vocabulario ilimitado, es decir, hay se favorece la generación de un sonido a partir de un texto de entrada. La concatenación de unidades, por lo tanto, es una técnica que depende de un corpus para la generación del habla sintética.

Esta técnica, también conocida como síntesis basada en corpus, requiere la formación de grandes bases de datos fonéticas y un análisis con el que se identifican las partes constituyentes de habla (segmentos, sílabas, palabras).

La propuesta de la síntesis concatenativa está basada en la generación de la señal de habla por concatenación de porciones de señal pre-almacenadas y ordenadas en un diccionario. Esas porciones de señal son recuperadas por un generador segmental que realiza la alineación, constituyendo una señal concatenada. Las porciones de señal son formadas por difonos, que contiene solamente una transición de segmento a segmento. De manera general, las porciones de señal pueden también ser formadas por polifonos, que contienen transiciones más complejas.

Mediante este procedimiento se genera habla sintetizada cercana a la producida por un locutor humano, puesto que las porciones de habla que forman un nuevo enunciado son segmentos de habla natural previamente almacenados.

Según Braga (2008: 15) la cualidad obtenida por la técnica de concatenación de unidades justifica éxito de las industrias del habla desde los años 80. Esas industrias surgen como una respuesta a la busca cada vez mayor por soluciones para equipos del tipo *hands-free* y *eyes-free*.

2.2.4 La síntesis por HMM

La síntesis por HMMs (*Hidden Markov Model*) constituye el más nuevo paradigma de los métodos de síntesis (Tokuda et al., 1995; Tokuda, 2004). Este sistema es capaz de producir una habla sintética de alta calidad a partir de bases de datos muy reducidas. Por ejemplo, Maia *et al.*, en el trabajo desarrollado en 2003, utilizaron 80 frases para el entrenamiento del sistema. En un trabajo más reciente publicado en 2006, Maia *et al.* utilizaron 221 frases foneticamente balanceadas de portugues brasileño para la implementación del sistema de conversión basado en HMM. El funcionamiento y la arquitectura del sintetizador de habla por HMMs aplicado al PB son descritos de manera minuciosa en el trabajo de Maia (2006).

De manera general, en la síntesis por HMMs hay una fase previa al entrenamiento de los modelos que tienen como input la señal del habla y la información lingüística, que constituye la base del funcionamiento del sistema.

Después del entrenamiento, el texto es procesado por el *front-end* que produce etiquetas fonéticas que, a su vez, pasarán por el módulo de selección y concatenación de HMMs. Las fases subsecuentes son: módulo de determinación de las duraciones y

módulo de generación de los parámetros acústicos. El módulo de generación, a su vez, está dividido en dos tipos: el módulo que genera la fuente (la señal de los impulsos de la glotis) y el módulo que genera el filtro (que simula las constricciones del tracto vocal).

En el funcionamento del sistema de síntesis por HMM se pueden distinguir dos procesos: entrenamiento del sistema y síntesis. La primera parte consta de tres fases: la extracción de los parámetros a partir de la base de datos del habla; la conversión de la información lingüística de los enunciados de la base de datos en etiquetas contextuales de HMMs y entrenamiento de los HMMs. La segunda parte está compuesta por: la generación de las etiquetas a partir de la información de las frases; la selección y la concatenación por HMMs; la determinación de los parámetros y control de la excitación y del filtro.

2.3 Técnicas de generación de la señal acústica

Según Barbosa (1999: 28-29) hay tres técnicas desarrolladas para los sistemas de síntesis del habla: PSOLA, LPC o el sintetizador por formantes.

El autor describe que la técnica PSOLA opera sobre la señal del habla, modificando los valores de duración y frecuencia fundamental de la señal concatenada por medio de la manipulación de los periodos de la apretura y cierre de la glotis.

En la técnica LPC, la señal del habla es producida por una fuente sonora (pulsos de la glotis o ruidos de fricción) aplicada a un filtro (el tracto vocal). El filtro es implementado por un conjunto fijo y pre-definido de parámetros que permiten la obtención de muestras del señal del habla a partir de muestras anteriores.

El sintetizador por formantes, a su vez, es un método de síntesis por reglas y genera la señal acústica a partir de la información, a una tasa de muestreo pre-definida de los siguientes parameros acusticos: cuatro primeros formantes, ancho de banda, amplitud, frecuencia fundamental, duración y intensidad.

2.4 Estado de la cuestión

Braga (2008: 28) considera que el estado de la cuestión de la síntesis del habla actual está dominado por los métodos de concatenación y, más recientemente, por la llamada síntesis mediante HMM. Los sistemas de conversión de texto en habla que emplean la concatenación dependen de largos bancos de grabaciones y de algoritmos que seleccionen unidades, es decir, por fragmentos de sonidos.

Los temas más estudiados en el actual estado de la cuestión en síntesis del habla son los siguientes: la síntesis por HMMs, la síntesis por emociones/expresiones, la evaluación, la síntesis multi-lengua, la síntesis audio-visual (o también llamada multimodal) y las nuevas aplicaciones.

Según Braga & Dias (2009), de manera general, hay tres enfoques principales para el estado de la cuesión del módulo *front-end*: la sintesis basada en reglas, la sintesis basada en modelos estadisticos y la sintesis basada en modelos híbridos. Según los autores, la sintesis basada en reglas son sistemas más robustos que requieren menos memoria, pero neceitan conocimientos lingüísticos avanzados. La sintesis basada en modelos estadisticos, por su vez, requieren gran cantidad de memoria computacional.

Actualmente, los principales temas debatidos en la comunidad academica, sobre el estado de la cuestión en el desarollo de sintetizadores del habla son:

(i) Sintesis de las emociones (Cabral, 2006; Cabral & Oliveira, 2006) y de la expresividad (Cahn, 1990; Bulut *et al.*, 2002; Hamza *et al.*, 2004; Eide *et al.*, 2004; Lee *et al.*, 2006; Schroder, 2006). Las empresas Loquendo y Cereproc son ejemplos de sistemas de síntesis con emociones;

(ii) Sintesis multi-modal (Martino, 2005; Raimundo *et al.*, 2007);

(iii) *Voice conversion* (Weiss *et al*., 2007);

(iv) Síntesis por HMM (Maia, 2006);

(v) Evaluaciones de sistemas (Black & Tokuda, 2005; Bennet & Black, 2006; Fraser & King, 2007; Weiss *et al.*, 2007) y

(vi) Prosódia y las interfaces con la fonologia y sintaxis (Barbosa, 2006; Teixeira, 2004; Braga & Marques, 2004; Seara *et al.*, 2007).

Las referencias de las investigaciones más recientes se encuentran en el periodico Speech Comunication[15] y también en los trabajos publicados en los congressos internacionais *Interspeech* e *Speech Prosody*.

Uno de los puntos de discussión en este trabajo es la revisión del módulo de trabscripcion fonética del sintetizador de *Verbio.* Aunque el tema del desarrollo del módulo de transcripción fonética sea muy discutida en el marco de las tecnologías del habla (síntesis y reconocimiento del habla), este tópico sigue siendo un reto por

[15] Los artículos se encuentran en: http://www.journals.elsevier.com/speech-communication/.

resolver, sobre todo en lo que respecta a los cambios de ortografía y actualizaciones que se producen en cualquier idioma.

Segun Braga (2008: 134-135), en al ambito teorico, podemos destacar las siguientes propuestas de resolución de los problemas verificados en los sistemas de síntesis:

(i) Arboles de decisión (Lucassen & Mercer, 1984; Oliveira *et al*., 2001),

(ii) Enfoques basados en diccionarios (Coker *et al.*, 1990),
Enfoques basados en reglas lingüísticas (Kaplan & Kay, 1994; Oliveira et al.,1991; Oliveira, 1996; Teixeira *et al*.,1998; Teixeira, 2004),

(iii) Modelos híbridos (Meng *et al.*, 1994),

(iv) Enfoques por redes neuronales (Sejnowski &. Rosenberg, 1987; Trancoso *et al.*,1994),

(v) Enfoques basados en estados finitos (Roche & Schabes, 1995; Caseiro & Trancoso, 2002; Caseiro *et al*., 2003; Oliveira *et al*., 2004),

(vi) Enfoques basados en HMM (Taylor, 2005),

(vii) Modelos estatísticos (Chotimongkol & Black, 2000) y

(viii) *Machine Learning* (Teixeira *et al*., 2006a, 2006b)

Según Braga (2008:135), el enfoque basado en diccionario consiste en una lista de palabras con su correspondecia transcripción fonética. Esta técnica es aplciada comumente a lenguas que no tienen isomorfismo 1:1 entre fonemas y grafemas, como es el caso del inglés y del portugués. Este enfoque tiende a fallar cuando aparecen palabras que están especificadas en el diccionario, como es el caso de los neologismos y de los extranjerismos. Al contrario del enfoque basado en diccionarios, los sistemas mixtos son basados en reglas lingüísticas y modelos estadisticos para generar las transcripciones fonéticas.

Sobre los sistemas desarollados para el portugués brasileño, cabe destacar el transcriptor Ortofon desarrollado para el sintetizador AIURUETÊ (Violaro *et al*.,1999). El Ortofon es un convertidor de grafema a fonema utilizando un modelo hibrido basado en dos diccionarios (el diccionario de abreviaturas y acrónimos, el diccionario de excepciones) y de reglas lingüísticas.

El sintetizador desarrollado por *Verbio* es un sistema mixto basado en reglas lingüísticas y en diccionarios, pues es un enfoque más económico con relación a la memoria computacional.

Para la revisión de los algoritmos desarollados para el transcritor fueron considerados los estudios en Fonética y Fonologia del portugués brasileño, propuestos por Silva (2002) y Caglari (2007).

2.5 Descripción fonética de las variantes de São Paulo y Rio de Janeiro

Según Silva (2002: 37-40), los siguientes segmentos consonantales ocurren de manera uniforme en todos los variantes del portugués brasileño: oclusiva bilabial sorda [p], oclusiva bilabial sonora [b], oclusiva alveolar sorda [t], oclusiva alveolar sonora [d], oclusivo velar sorda [k], oclusiva velar sonora [g], fricativa labiodental sorda [f], fricativa labiodental sonora [v]. fricativa alveolar sorda [s], nasal bilabial sonora [m], nasal bilabial sorda [n], lateral alveolar sonora [l], consonante lateral alveolar palatalizada sonora [l^j] y la consonante nasal palatal [ɲ].

La africada palatoalveolar sorda [tʃ] y la africada palatoalveolar sonoro [dʒ] son rasgos fonéticos tipicos de la pronunciación de la región sudeste de Brasil.

La fricativa alveolar sorda [s] es uniforme en el inicio de sílaba en todos los variantes del portugués brasileño. La fricativa alveolar sonora [z] es un fonema uniforme en inicio de sílaba en todas las variantes del portugués brasileño. En final de sílaba, la fricativa alveolar sonora [z] es un rasgo de la variedad de São Paulo.

La fricativa palatoalveolar sorda [ʃ] es uniforme en inicio de sílaba en todas las variantes del portugués brasileño. En coda silábica, la fricativa palatoalveolar sorda [ʃ] es un rasgo de la variedad de Rio de Janeiro.

La fricativa palatoalveolar sonora [ʒ] es uniforme en inicio de sílaba en todas las variedades del portugués brasileño. En coda silábica, la fricativa palatoalveolar sonora /ʒ/ es un rasgo de la variedad de Rio de Janeiro.

La fricativa velar sorda [χ] la pronunciación típica del dialecto del Rio de Janeiro y se caracteriza por la friccíon audible en la región velar. Es un fonema consonántico que ocurre en inicio de palabra, inicio de sílaba precedida por vocal y inicio de sílaba precedida por consonante. En la variante de Rio de Janeiro, este fonema ocurre en final de sílaba cuando seguido por consonante sorda y en final de sílaba que coincide con final de palabra.

La fricativa velar sonora [ɤ] es un fonema típico de la variante del Rio de Janeiro, se caracteriza ocurre fricción audible en la región velar. Ocurre en final de sílaba seguida de consonante sonora.

El tepe alveolar sonoro [ɾ] es uniforme en posición intervocálica y cuando seguido por consonante en todas las variedades del portugués brasileño. Este fonema fue escogido como la prononciación típica de la variedad de São Paulo para el desarrollo del módulo de transcripción fonética del sintetizador de Verbio cuando aparece en final de sílaba seguido por consonante sorda y en final de sílaba que coincide con final de palabra.

En portugués brasileño hay 26 grafemas para las consonantes. Según Silva (2002) existen en la lengua portuguesa de Brasil doce segmentos vocálicos plenos (siete orales y cinco nasales): /a/, /e/, /i/, /o/, /u/, /ɛ/, /ɔ/, / ĩ/, /ẽ/, /ã/, /õ/, /ũ/.[16] Sin embargo, hay solamente cinco grafemas vocálicos: <a>, <e>, <i>, <o>, <u>.

Las vocales reducidas representadas en IPA como [I] y [ʊ] son alófonos de los fonemas /i/ y /u/, es decir, son unidades de variantes posicionales que se relacionan con la manifestación fonética de los fonemas. Las vocales reducidas suelen ocurrir en silabas pre-tónicas o pos-tónicas. Así, para el análisis de un fonema vocálico en PB es importante considerar la posición del segmento vocálico con relación al patrón acentual. Además, las semi-vocales [j] e [w] también son representadas por los grafemas <i> e <u>, hecho que aumenta la ambigüedad gráfica.

2.6 Coarticulación entre palabras

Para el desarrollo de un sintetitzador cuyo resultado sea lo más próximo posible al habla natural, además del estudio de los segmentos consonánticos y vocálicos que constituyen la lengua estudiada, es importante también realizar la revisión de las reglas de los fenómenos de coarticulación entre palabras. Tales reglas se encuentran en el módulo de transcripción fonética del *front-end* y son establecidas de fonema a fonema, después de la conversión de los grafemas a fonemas.

Como referencia de estudio sobre los fenómenos de coarticulación entre palabras, puede ser citado el trabajo elaborado por Tenani (2002), que analiza procesos de sándi externo con el objetivo de identificar evidencias de la estructura prosódica del

[16] Según el IPA (*International Phonetic Alphabet*).

PB por intermedio del enfoque de la y jerarquía prosódica como proponen Nespor & Vogel (1986), dentro de la perspectiva de la Fonología Prosódica.

La autora considera los contextos segmentales y acentuales que favorecen cada uno de los procesos de coarticulación: (i) sonorización de fricativas, (ii) *tapping*, (iii) degeminación, (iv) elisión, (v) diptongación y (vi) haplología.

A continuación se pueden observar algunos ejemplos elaborados por Tenani (2002) para cada proceso citado:

1. Arroz amarelo (*Arroz amarillo*)
2. Açúcar amarelo (*Azúcar amarillo*)
3. Laranja amarela (*Naranja amarilla*)
4. Laranja holandesa (*Naranja holandesa*)
5. Pêssego amarelo (*Melocotón amarillo*)
6. Faculdade dinàmica (*Facultad dinámica*)

Capítulo 3

3. Metodología

En este capítulo se presenta la metodología utilizada durante la práctica en la empresa *Verbio technologies SL* para la realización del trabajo.

3.1 El sintetizador de Verbio Speech Technologies SL

El sistema de conversión de texto en habla con el que se ha trabajado en esta investigación ha sido desarrollado por la empresa *Verbio Technologies SL*.[17] Se trata de una compañía global con más de veinte años de investigación y desarrollo en el área de tecnologías de síntesis del habla y reconocimiento de voz. Además de España, existen sucursales en Argentina, Brasil, Chile, Colombia, México, Paraguay, Perú y Uruguay. A partir de 2001, *Verbio* ha enriquecido su tecnología con nuevas funcionalidades y productos, como: Verbio XML, *WordSpotting*, Verificación e Identificación, *Call Steering*, *Speech Analytics* y Transcripción de Voz a Texto.

La tecnología Verbio TTS convierte de forma automática un texto escrito en una locución del habla natural en distintos idiomas (castellano, catalán, valenciano, euskera, gallego, inglés, francés, portugués europeo, portugués brasileño, mexicano y argentino).

El *software* desarrollado por la empresa *Verbio* utiliza la concatenación de unidades como técnica para la síntesis de texto en habla, pues pues de esta manera es posible obtener alta naturalidad en el *output* del sistema a bajo costo computacional.

La técnica de concatenación de unidades utiliza un conjunto fijo de polífonos (segmentos sonoros formados por dos o más fonemas) que se encuentran almacenados en un diccionario sonoro compuesto por fragmentos de sonidos extraídos de grabaciones. El inventario de unidades acústicas está compuesto por fragmentos de distintos tamaños: desde semi-sílabas hasta secuencias de cinco fragmentos, totalizando 82.000 unidades.

El sistema fue programado en lenguaje C++ para plataforma Windows (NT, 2000, XP, 2003 y Vista) y Linux. La interfaz gráfica tambien fue desarrollada en lenguaje C++. Los requerimientos de memoria son los siguientes: motor de sintesis (*Vox Server*) de 10 MB; módulo de voz de 8 khz de aproximadamente 60 MB, módulo de voz de 16 khz de aproximadamente 120 MB, 30 MB en RAM como mínimo por

[17] Puede obtenerse más información sobre *Verbio Technologies SL* en http://www.*verbio*.com.

locutor. La memoria RAM tiene un consumo inicial elevado durante la puesta en marcha, pero posteriormente se mantiene estable en el tiempo.

La tasa de muestro puede ser de 8KHz o de 16KHz, 16 bits por ley A y ley MU. Las arquitecturas pueden ser por monopuesto (*desktop*) o por cliente-servidor. Los requisitos minimos necesarios del CPU son Pentium 4 (3 Ghz y 512 MB de memoria RAM). La carga de CPU es elevada al realizar una petición de síntesis, pero posteriormente se mantiene estable en el tiempo. Las interfaces son SAPI 4 y SAPI 5, MRCP v1, Verbio API (SDK) y VoiceXML.

Para el procesamiento de la señal del habla fue utilizada la técnica PSOLA (*Pitch Synchronous Overlap Add Method*). Según Barbosa (1999: 28-29), la técnica PSOLA opera sobre la señal del habla, modificando los valores de duración y frecuencia fundamental de la señal concatenada por medio de la multiplicación, reducción, compresión o expansión de los periodos realizados por la apertura y cierre de la glotis.

La figura 3. a continuación representa el digrama general del sistema de síntesis desarollado en *Verbio*:

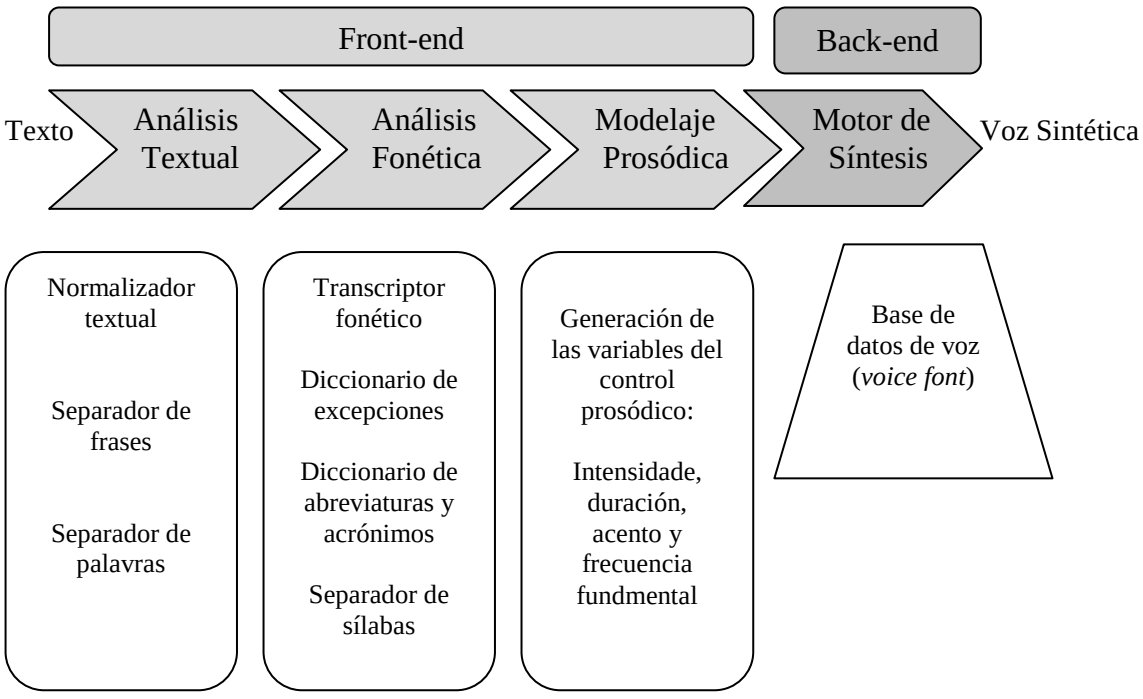

Figura 3. Digrama general del sistema de síntesis desarollado en *Verbio*

La arquitectura general del sintetizador de la empresa *Verbio* cuenta con tres módulos: *front-end*, *back-end* y *voice font*.

El pre-procesamiento textual (o *front-end*) del sintetizador de texto en habla de *Verbio* es compuesto por: (i) un sistema de normalización textual, (ii) el transcritor grafema a fonema y (iii) un módulo de procesamiento prosódico. Para sintetizar una frase, el *software* integra los segmentos y suprasegmentos del componente fónico.

El módulo *back end*, a su vez, está constituido por el motor de síntesis. Por último, la base de datos de voz (*voice font*) está vinculada al proceso de selección del locutor y de grabación en estudio.

El análisis textual se hace mediante el proceso de normalización, el separador de frases (información útil para la posterior generación de los modelos prosódicos) y el separador de palabra.

El módulo de análisis fonético es compuesto por el convertidor de grafema a fonema (dónde están los algoritmos basados en conocimientos fonéticos), los diccionarios de siglas y acrónimos diccionario y excepciones (incluyendo las palabras extranjeras agregadas al léxico del portugués brasileño). Primeramente, son procesados los diccionarios incorporados al sistema y después son procesadas las reglas del algoritmo de conversión de grafemas en fonemas.

Después del tratamiento y procesamiento textual, hay la última etapa del módulo *front-end*, el llamado módulo de generación prosódica. En esta etapa, ocurren los siguientes procesos: división silábica, marcador de sílaba tónica, para el modelado prosódico.

El sistema de normalización textual de *Verbio* es compuesto por dos diccionarios distintos: el diccionario de excepciones y el diccionario de abreviaturas y acrónimos. El diccionario de excepciones presenta las transcripciones de las palabras extranjeras que no siguen los mismos procesos fonológicos y prosódicos de la lengua en cuestión.[18] El segundo diccionario presenta las expansiones de las abreviaturas y acrónimos.

Em caso de dudas con relación a la ortogrfia de las palabras en portugués brasileño, fue consultado siempre que posible el VOLP[19] (Vocabulario Ortografico de la

[18] La falta de estudios más completos sobre los procesos fonológicos en el tratamiento de las palabras extranjeras provoca muchas dudas en la transcripción fonética.
[19] Disponible en: http://www.academia.org.br/abl/cgi/cgilua.exe/sys/start.htm?sid=19 [08/04/2013].

Lengua Portuguesa), preparado por la Academia Brasileña de Letras, en línea con el nuevo Acuerdo Ortográfico.

En el módulo de transcripción fonética están las reglas de conversión de grafema a fonema y las reglas de transformación de fonema a fonema, es decir, dónde se encuentran las reglas que describen los procesos de coarticulación entre palabras, como es el caso de los procesos de sándi externo (degeminación, elisión, sonorización de fricativa, *tapping*, haplología y diptongación).[20]

Además de los módulos que componen la arquitectura general del sistema, es necesario también elaborar un *corpus* que será utilizado como soporte de lectura para la posterior grabación en estudio. Según Parodi (2008:106), *corpus* puede ser definido como un conjunto amplio de textos digitales de naturaleza específica y que cuenta con una organización predeterminada en torno a categorías identificables para la descripción y análisis de una variedad de lengua. El autor enumera algunas características relevantes para la construcción de un corpus de entre ellas, se puede destacar: extensión, representatividad, diversificación y tamaño de las muestras.

El *corpus* seleccinado para este trabajo es compuestao por 1807 frases aisladas extraídas de libros, periódicos, portales de voz y otros textos con formatos digitales y disponibles de manera libre en la *web*. El *software* se basa, principalmente, en el uso de grandes corpus, lo que da lugar a mejoras en la calidad de los resultados.

La base de datos de voz (*voice font*) fue seleccionada de manera rigorosa y grabada en estudio profesional. La cualidad de la grabación también es un factor importante para las posteriores analices acústicas. Los datos acústicos fueron generados en una cámara anecoica preparada para grabaciones de voz de alta cualidad.

La base de datos genera los segmentos de voz, pero también los modelos de entonación y los modelos de duración de los segmentos. La calidad del sintetizador depende en gran medida de la calidad de la grabación y de la voz grabada.

En primer lugar, la prioridad es que la voz y el habla sean naturales, claras y articuladas. Es preciso que la lectura sea concatenada, es decir, realizada a un ritmo y velocidad contantes a lo largo de la grabación, de tal forma que en la lectura del listado de frases, la última se parezca al máximo a las primera y a las centrales.

La elocución debe ser pausada, aunque no debe llegar a ser lenta. El locutor tiene que entonar con naturalidad y sobre todo no debe haber entonacionales demasiado

[20] Para más informaciones sobre tales procesos sugerimos la lectura de la tesis de doctorado de Tenani (2002).

exageradas o teatralizadas. En la medida de lo posible, la entonación debe ser coherente y homogénea, es decir, conviene que la entonación y la duración sean similares.

El locutor debe leer las frases de manera a evitar la producción de foco estrecho, pues el énfasis puede interferir en los procesos entonacionales proyectando curvas de F0 con características peculiares. Las frases deben ser lidas como si fuera el titular de un reportaje, en el cual toda la información contenida en la oración es nueva y no hay énfasis o foco de un elemento determinado.

Es importante que sean realizadas pausas y descansos en la grabación, evitandose así la voz cansada o temblorosa del locutor. El tono de voz debe ser claro y constante en toda la grabación. Durante la grabación, son muy perjudiciales los ruidos, tales como movimientos de papeles o de brazaletes.

En este trabajo fueron utilizadas las grabaciones de una única persona, ya que las unidades de síntesis extraídas son superpuestas a los datos prosódicos de un solo hablante. Para la grabación fue escogido un hablante del sexo femenino denominada Julia. Los criterios de selección de la locutora fueron la capacidad de hacer largas locuciones y los buenos resultados en las pruebas de síntesis.

Las frases seleccionadas para la grabación son fonéticamente balanceadas, es decir, el diseño de listas fonéticamente balanceadas requiere la determinación de datos sobre la frecuencia de aparición de los fonemas y alófonos existentes en la variante lingüística investigada. Para esto, es necesario disponer datos procedentes de la lengua oral y datos que consideren los fonemas y los alófonos de estas variantes.[21]

La interfaz del sintetizador de habla permite una entrada de texto (*input*) desde el archivo en formato *.txt* o desde la ventana de texto puesto a disposición del usuario. El audio producido por el sintetizador (*output*) puede ser grabado en formato *.wav*. La frecuencia de muestreo del audio es de 16.000 Hz.

En la siguiente figura 4. se puede observar la interfaz del sistema TTS del *software*.

[21] Sobre la cuestión de la frecuencia de ocurrencia de fones puede ser citado el trabajo desarrollado por Alcaim *et al*. (1992). El referido trabajo tiene como objetivo la construcción de un conjunto de veinte listas de diez frases fonéticamente equilibradas para el portugués hablado en Río de Janeiro. Para el establecimiento de listas de frases fue necesaria la realización de un análisis previo de la frecuencia relativa de los fones en portugués hablado en Río de Janeiro. El valor de intervalo chi-cuadrado ($\chi 2$) entre la frecuencia de los fones en cada lista y su incidencia en la lengua fue utilizado como indicador del equilibrio de las frases. Una de las principales aplicaciones de las listas de frases balanceadas está en la evaluación subjetiva de la calidad de voz procesada por codificadores digitales.

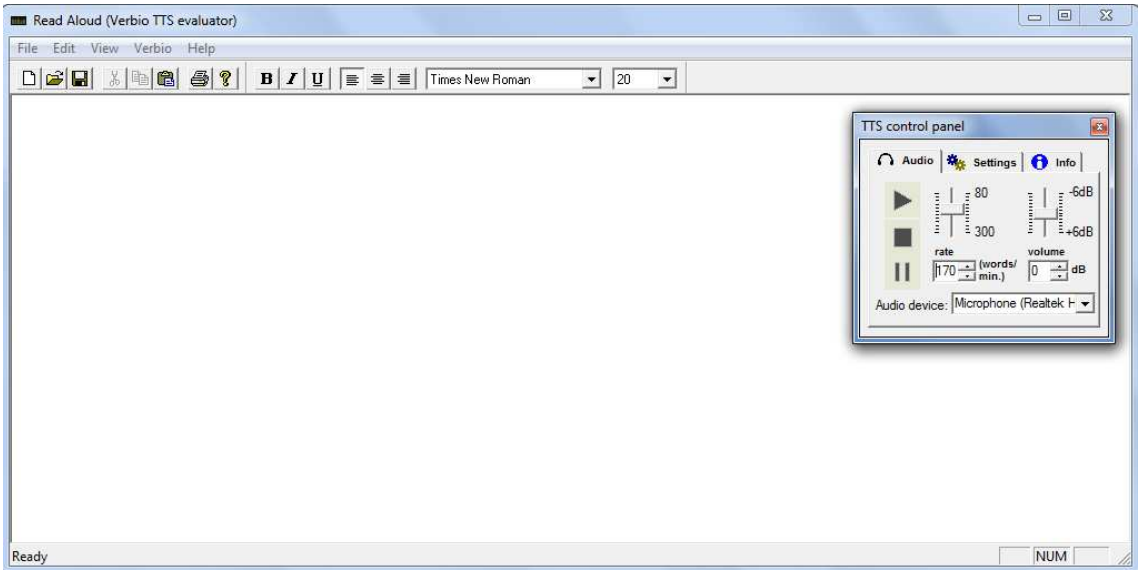

Figura 4. Interfaz del sintetizador de *Verbio*

3.2 El transcriptor fonético

El transcriptor fonético de Verbio fue desarrollado utilizándose dos sistemas distintos: (i) el sistema SAMPA y (ii) el sistema SEGRE. Estos sistemas serán descritos a continuación.

3.2.1 Sistema SAMPA

A nivel de notación fonética, el alfabeto SAMPA (*Speech Assessment Methods Phonetic Alphabet*) desarrollado por Wells en 1996 suele ser el más utilizado en la comunidad científica para la transcripción en tecnologías del habla (Braga *et al*., 2003: 1352). Así como en la mayoría de las investigaciones, en este trabajo será utilizado el transcriptor SAMPA (Wells, 1996) creado para el portugués de Portugal y adaptado al portugués de Brasil. La representación SAMPA es más difundida entre las tecnologías del habla, pues los caracteres están disponibles en el teclado del ordenador y esto facilita la transcripción fonética.

En el nivel de notación fonética, en este trabajo fue utilizado el alfabeto SAMPA, conforme la tabla 1., por ser el alfabeto más adecuada y ampliamente utilizado en los lenguajes de procesamiento computacional. Los símbolos utilizados en el alfabeto SAMPA son caracteres especiales que aparecen en los teclados. El SAMPA constituye la mejor base sólida de colaboración internacional para un estándar legible por máquina de codificación de notación fonética.

El SAMPA (*Speech Assessment Methods Phonetic Alphabet*) es un alfabeto fonético legible por máquina. Fue desarrollado originalmente en el proyecto ESPRIT 1541, SAM en 1987-89 por un grupo internacional de fonética, y se aplicó en primer lugar a las lenguas Comunidades Europeas, danesa, española, inglés, francés, alemán y italiano (1989), y más tarde a noruega y Suecia (1992) y, posteriormente, al portugués griego y español (1993).

En la tabla a continuación están los alófonos, los fonemas correspondientes, el tipo de fonema (sordo o sonoro), el punto de articulación y el modo de articulación de los segmentos utilizados en este trabajo:

Alófonos	Fonemas	Tipo	Punto	Modo
p	p	Sorda	Bilabial	Oclusiva
b	b	Sonora	Bilabial	Oclusiva
t	t	Sorda	Alveolar	Oclusiva
d	d	Sonora	Alveolar	Oclusiva
k	k	Sorda	Velar	Oclusiva
g	g	Sonora	Velar	Oclusiva
tS	t	Sorda	Palatoalveolar	Africada
dZ	d	Sonora	Palatoalveolar	Africada
f	f	Sorda	Labiodental	Fricativa
v	v	Sonora	Labiodental	Fricativa
s	s	Sorda	Alveolar	Fricativa
S	s	Sorda	Palatoalveolar	Fricativa
z	z	Sonora	Alveolar	Fricativa
S	S	Sorda	Palatoalveolar	Fricativa
Z	Z	Sonora	Palatoalveolar	Fricativa
m	m	Sonora	Bilabial	Nasal
n	n	Sonora	Dental	Nasal
J	J	Sonora	Palatal	Nasal
l	l	Sonora	Alveolar	Lateral
L	L	Sonora	Palatal	Lateral
r	r	Sonora	Alveolar	Tepe
R	r	Sonora	Alveolar	Vibrante
X	r	Sorda	Velar	Fricativa
w	u	Semi	Posvelar	Aproximante
w~	u	Glide Nasal	Posvelar	Aproximante
j	i	Semi	Palatal	Aproximante
j~	i	Glide Nasal	Palatal	Aproximante
i	i	Vocal	Anterior	Alta no redondeada
I	e	Vocal	Anterior	Media alta no redondeada
e	e	Vocal	Anterior	Media no redondeada
E	e	Vocal	Anterior	Media baja no redondeada
a	a	Vocal	Central	Baja no redondeada
6	a	Vocal	Central	Media no redondeada
o	o	Vocal	Posterior	Media redondeada
O	o	Vocal	Posterior	Media baja redondeada
u	u	Vocal	Posterior	Alta redondeada
U	o	Vocal	Posterior	Media alta redondeada
6~	a~	Vocal Nasal	Central	Media no redondeada
i~	i~	Vocal Nasal	Anterior	Alta no redondeada
e~	e~	Vocal Nasal	Anterior	Media no redondeada
o~	o~	Vocal Nasal	Posterior	Media redondeada
u~	u~	Vocal Nasal	Posterior	Alta redondeada

Tabla 1. Características fonéticas de cada segmento

La transcripción fonética es el módulo que trata de la conversión grafema a fonema. En el proceso de conversión de grafemas en fonos el sistema de transcripción debe ser unívoco.

Una transcripción fonética puede ser realizada de dos maneras distintas: detallada (restricta/ estrecha) o más general (amplia/ ancha). El tipo de transcripción depende, sobretodo, de la finalidad del trabajo, de los variantes seleccionados, del estilo o velocidad del habla. Según Silva (2002: 36), la transcripción fonética amplia considera solamente los aspectos que no están relacionados con determinados contextos o características específicas de la lengua o de un variante específico. En este trabajo serán realizadas transcripciones fonéticas restricta/estrecha, pues serán tratadas las propiedades de los segmentos de las variantes de São Paulo y Rio de Janeiro.

El no isomorfismo entre los fonemas y los grafemas se verifica para las vocales del PB. Según Silva (2002: 79-86) existen en la lengua portuguesa brasileña siete vocales tónicas orales /a, e, i ,o, u, ɛ, ɔ/, tres vocales pretónicas orales /ɛ, ɔ, ə/, tres vocales postónicas finales /I, ə, ʊ/ y cinco vocales nasales: /ĩ, ẽ, ã, õ, ũ/. Sin embargo, hay solamente cinco grafemas vocálicos: <a, e, i, o, u>.

Las vocales pasan a nasales cuando seguidas por una consonante nasal. Silva (2002: 121) en varios variantes de la región Sudeste, una vocal tónica es siempre nasalizada cuando seguida de consonante nasal. Sin embargo, cuando la vocal está en posición pretónica, la nasalidad es facultativa. En este trabajo, las vocales seguidas de consonante nasal en posición pretónica no pasan por el proceso de asimilación, por ejemplo: b[a]nana y no b[ã]nana.

Según Silva (2002: 121) cuando la consonante nasal es palatal, las vocales tónicas y pretónicas son nasalizadas en la gran parte de los variantes del portugués brasileño: m[ã]nha.

Las vocales reducidas representadas en IPA como [I], [ʊ], [ə] son, en realidad, alófonos de los fonemas /i/, /u/ y /a/ respectivamente. Las vocales reducidas son unidades de variantes posicionales que se relacionan con la manifestación fonética de los fonemas.

Las vocales reducidas suelen ocurrir en silabas pre-tónicas o pos-tónicas. Así, para el análisis de un fonema vocálico en PB es importante considerar la posición del segmento vocálico con relación al patrón acentual, es decir, si la vocal <u> aparece en

una sílaba átona, la representación del SAMPA-PB será [U]. Si la vocal está asociada a la sílaba tónica, la representación fonética será [u]. Este proceso ocurre también con la vocal <i>: en contexto de sílaba átona, la representación basada en el SAMPA-PB es [I] y en sílaba tónica es [i].

Proceso semejante también puede ser observado con la vocal <a>. La vocal reducida representada en IPA como [ə] es un alófono del fonema /a/ en contexto átono.

Según las representaciones del SAMPA-PB, en contexto tónico, esta vocal es representada como [a] y en contexto átono pasa a ser representada como [6]. Es importante considerar que la vocal nasal del IPA [ã] pasa a ser representada como [6~] en SAMPA-PB.

Además, las semi-vocales [j] e [w] también son representadas por los grafemas <i> e <u>, hecho que aumenta la ambigüedad gráfica. En esta investigación, los grafemas <i> y <u> son representados como [j] e [w], cuando estos son semi-vocales.

Según Silva (2002:155), existen en PB 18 fonemas consonánticos /p, b, t, d, k, g, f, v, s, z, ʃ, ʒ, r, m, n, ɲ, l, ʎ/. Sin embargo, hay 21 consonantes gráficas <b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, w, x, y, z>. El no isomorfismo 1:1 entre fonemas y grafemas se verifica una vez más.

Matte et al. (2006: 36) enseñan un ejemplo de fenómeno fonológico muy frecuente en portugués brasileño: la palatalización de los fonemas /d/ y /t/ antes de la vocal /i/. El fonema /d/, que es fonológicamente equivalente en cualquier variedad geográfica de Brasil, puede ser producida como una africada [dʒ] en determinadas regiones cuando es seguido por /i/ o como una oclusiva [d] en otras regiones del país. Lo mismo ocurre para el fonema /t/ que puede ser producido como una africada [tʃ] o como una oclusiva [t], dependiendo del contexto geográfico. Ese fenómeno fonológico puede ser verificado en palabras como “dia” (día), “dialeto” (variante), “diferença” (diferencia), “tia” (tía), “tímpano” (tímpano), “timbre” (timbre).

Otro fenómeno fonológico observado en PB es la epéntesis de la vocal anterior alta /i/ entre la secuencia de fonemas [p] e [s] como en la palabra “psicológico" [pɪ.sɪ.ko.’lɔ.ZɪkU].

Silva (2002: 77) considera que “las vocales tónicas son aquellas que presentan prominencia acentual cuando comparadas a las otras vocales de la palabra”. Las vocales tónicas o acentuadas están, por lo tanto, asociadas al acento más fuerte de la palabra o también llamado acento primario. Las vocales átonas o no acentuadas (las

pre-tónicas o pos-tónicas) pueden estar asociadas al acento secundario o no llevan acento (Silva, 2002: 77).

Es importante destacar que en el módulo fonético del TTS de *Verbio*, el acento primario es representado por el apóstrofo ['] y el acento secundario es representado por el símbolo gráfico [`] y se marca inmediatamente antes de la vocal acentuada. Palabras con 3 o más silabas pre-tónicas suelen presentar acento secundario en PB para la construcción del ritmo del habla.

Los ejemplos a continuación ilustran las explicaciones anteriores:

Vocales			
Símbolo	**Palabra**	**Transcripción**	**Descripción según el IPA**
Orales			
i	mico	m'i - KU	Alta anterior no redondeada
I	bote	b 'O - tS I	Alta anterior no redondeada
e	seco/pêssego	s'e - KU/ p'e - se - gU	Media alta anterior no redondeada
E	teto	t'E - tU	Media baja anterior no redondeada
a	tato	t'a - tU	Baja central no redondeada
6	mata	m'a - t6	Media baja central no redondeada
o	cômodo	c'o - mo - dU	Media alta posterior redondeada
O	chora	S'O - r6	Media baja posterior redondeada
u	suco	s'u - KU	Alta posterior redondeada
U	soco	s'o - KU	Alta posterior redondeada
Nasales			
6~	rã	r'6~	Media baja central no redondeada nasal
i~	rim	r'i~	Alta anterior no redondeada nasal
e~	abençoo	6 - b e~ - s 'o – U	Media alta anterior no redondeada nasal
o~	onda	'o~ - d6	Media alta posterior redondeada nasal
u~	rum	r'u~	Alta posterior redondeada nasal

Tabla 2. Vocales del alfabeto SAMPA y caracteristicas fonéticas

Consonantes			
Simbolo	**Palabra**	**Transcripción**	**Descripción según IPA**
Oclusivas			
p	pata	p'a.t6	Oclusiva bilabial sorda
b	bela	b'E.l6	Oclusiva bilabial sonora
t	toma	t'O.m6	Oclusiva alveolar sorda
d	dedo	d'e.du	Oclusiva alveolar sonora
k	cada	k'a.d6	Oclusiva velar sorda
g	gato	g'a.tU	Oclusiva velar sonora
Africadas			
tS	tia	tS'i6	Africada alveopalatal sorda
dZ	dia	dZ'i6	Africada alveopalatal sonora
Fricativas			
f	foca	f'O.k6	Fricativa labiodental sorda
v	vaca	v'a.k6	Fricativa labiodental sonora
s	sapo	s'a.pU	Fricativa alveolar sorda
z	zebra	z'e.br6	Fricativa alveolar sonora
S	chave	S'a.vI	Fricativa alveopalatal sorda
Z	jaca	Z'a.k6	Fricativa alveopalatal sonora
Nasales			
m	mata	m'a.t6	Nasal bilabial sonora
n	nata	n'a.t6	Nasal dental sonora
J	nhoque	J'O.kI	Nasal palatal sonora
Líquidas			
l	lata	l'a.t6	Lateral alveolar sonora
L	lama	L'a~.m6	Lateral palatal sonora
r	prato	pR'a.t6	Tepe alveolar sonora
X	mar	m'aX	Fricativa velar sorda
R	Rato/jarra	R'a.tU/ Z' a.R6	Vibrante alveolar sonoro
Glides			
w	mau/sol	m'aw/s'Ow	Aproximante labiovelar sonora
w~	mão	m'a~w~	Aproximante labiovelar sonora nasal
j	mais	m'a.js	Aproximante palatal sonora
j~	mãe	m'a~j~	Aproximante palatal sonora nasal

Tabla 3. Consonantes del alfabeto SAMPA y caracteristicas fonéticas

3.2.2 Sistema SEGRE

La metodología de esta investigación está basada fundamentalmente en el trabajo realizado por Pachès et al. (2000), que desarrollaron un transcriptor fonético automático llamado SEGRE para los cuatro variantes del catalán. La sintaxis del transcriptor fue elaborada para posibilitar la conversión del grafema en fonema de un fonema a otro, necesarios para el funcionamiento de un sistema de síntesis del habla.

Para la generación del habla sintética es necesaria que la información fonética sea derivada de un texto de entrada (el *input*). La herramienta Segre posibilita la transcripción fonética automática de un texto de entrada. Las reglas de conversión son procesadas por archivos de la tabla ASCII [22] que, a su vez, especifican reglas según una sintaxis determinada por los contextos adyacentes de cada fonema. La transcripción fonética automática toma como entrada cualquier texto escrito en portugués brasileño y genera como salida una cadena de carateres del tipo ASCII.

Para la transcripción fonética es necesario, en primer lugar, la segmentación de la palabra en segmentos es decir, si la palabra es compuesta y presenta un guión (como: guarda-chuva, criado-mudo, bem-vindo), si presenta un apóstrofo (como: d'água) o si empieza por un prefijo (como: desrespeitar, impaciente, submarino) se debe realizar la separación de cada una de las partes que forman la palabra. Es necesaria también, la división de las palabras en sílabas por intermedio de la identificación del núcleo silábico y de la distribución de los grafemas de silabas distintas.

Después de estos dos pasos iniciales, se procede a la atribución del acento. En PB, específicamente, las reglas de atribución de acento son fijas.[23] Son determinados también los acentos primarios y secundarios, además de las fronteras entre silabas.

Por fin, existen las reglas de redistribución silábica que son aplicadas cuando hay reestructuración de la sílaba, como es el caso de la epéntesis vocálica. La epéntesis genera reestruucturación silábica, porque se añade una vocal a la consonante en coda.

El sistema Segre obtiene la transcripción fonética por medio de la aplicación de los siguientes procedimientos a partir de un texto de entrada: identificación de símbolos de puntuación para la delimitación de las frases y, consecuentemente, determinación de los silencios. Cada frase del texto de entrada es procesada separadamente. A continuación, cada palabra es transcrita de modo aislado, como definido anteriormente.

[22] La tabla ASCII (*American Standard Coding for Information Interchange*) es compuesta por una lista de caracteres que pueden ser interpretados por el sistema informático.
[23] Estas reglas serán explicadas y descritas en el capítulo 4.11 sobre la acentuación.

En seguida, las reglas fonema a fonema son aplicadas, pues alófonos de las fronteras de palabras pueden sufrir los efectos de coarticulación. El último paso para obtener la transcripción fonética final es la aplicación de las reglas de redistribución silábica.

La herramienta SEGRE es definidos en catalán. A continuación están los comandos utilizados para empleo de las reglas grafema a fonema:

-nucli: núcleo

-posmot: posición del grafema en la palabra

-possil: posición del grafema en la sílaba

-avant: el grafema ocurre antes de

-darrere: el grafema ocurre después de

-átona: sílaba átona

-tónica: sílaba tónica

-propernucli: próximo núcleo

-fitxer_exc: ficheros de excepción

-defecte: *default*

Las reglas grafema a fonema siguen el siguiente formato:

<GRAFEMA> [FONE] - CONTEXTO IZQUIERDO -CONTEXTO DERECHO

Las reglas grafema a fonema son interpretadas a partir de los contextos izquierdo y derecho del grafema en cuestión. Estos contextos pueden ser: anteposición o posposición a determinados grafemas, la posición en sílaba átona o tónica y la posición inicial, media o final de la palabra. Es importante destacar que las reglas son organizadas en una lista jerarquizada, tomándose primeramente las reglas más específicas y dejandose al final la regla por defecto, es decir, la regla más general.

3.2.3 Nueva Ortografia de la Lengua Portuguesa

La nueva ortografía de la lengua portuguesa fue decretada en 01/01/2009 sob el numero 6.583/2008 y su vigencia sera obligatoriaa a partir del año 2013, su objetivo es la unificación de la ortografía de la lengua portuguesa que, actualmente, es el único idioma de occidente que presenta dos grafías oficiais en Brasil y en Portugal.

Además, el portugués es la lengua oficial de ocho países: Angola, Brasil, Cabo Verde, Guiné-Bissau, Mozambique, Santo Tomé y Príncipe, Portugal y Timor Leste, totalizando aproximadamente 230 millones del hablantes. La justificación de la nueva

ortografía es el establecimiento de una comunidad que constituya una unidad linguistica expresiva, ampliando la participación del portugués en los organismos internacionales.

La unificación ortográfica permitirá, por ejemplo, la circulación de materiales impresos (libros, periódicos) y documentos oficiales sin la necesidad de traducción de dichos materiales. Una vez unificado, el idioma podrá ser inserido como uno de los idiomas oficiales de la Organización de las Naciones Unidas (ONU).

En la nueva ortografía de la lengua portuguesa son abordados los siguientes temas:

- grafía de nombres extranjeros;
- el uso de h;
- los grafemas consonánticos;
- las secuencias consonánticas;
- las vocales átonas;
- las vocales nasales;
- los diptongos;
- los cambios en las reglas de tilde;
- la supresión de la diéresis;
- el uso del guión;
- el uso del apóstrofo;
- el uso de las letras mayúsculas y minúsculas y la división silábica.

Los cambios ortográficos se indican a continuación y están basados en el manual escrito por Tufano, (2008) y también según las “Bases do Novo Acordo Ortográfico da Língua Portuguesa” (2009). Además de las bases sobre los cambios en la nueva ortografía, fue utilizado también el Vocabulario Ortográfico de la Lengua Portuguesa[24] en caso de dudas sobre la ortografía de las palabras.

(a) La tilde gráfica de los diptongos abiertos <éi> y <ói> es suprimida:
idéia > ideia
assembléia > assembleia
jibóia > jiboia
asteróide > asteroide

[24] http://www.academia.org.br/abl/cgi/cgilua.exe/sys/start.htm?sid=23 [consulta: 08/06/2012].

(b) La diéresis también es eliminada en la nueva ortografía:

lingüiça > linguiça

tranqüilo > tranquilo

cinqüenta > cinquenta

(c) En las palabras agudas, la tilde permanece sin alteraciones. Es eliminada la tilde grafica del hiato <oo>:

vôo > voo

enjôo > enjoo

(d) No se utiliza la tilde gráfica en la terminación verbal <eem>:

Eles crêem > Eles creem

Elas relêem > Elas releem

Todos vêem > Todos veem

(e) Las tildes que diferencian palabras homógrafas fueron suprimidas:

Para (verbo) – Pára > Para

Polo (sustantivo) – Pólo > Polo

Pelo (sustantivo) – Pêlo > Pelo

Pera (sustantivo) – Pêra> Pera

(f) En palabras compuestas, cuando la primera palabra termina en vocal y la palabra siguiente empieza con la misma vocal hay un guion separándolas:

contra-ataque

micro-ondas

anti-inflamatório

(g) En palabras compuestas, cuando la primera palabra termina en vocal y la palabra siguiente empieza con una vocal diferente no hay guión:

anti + aéreo = antiaéreo

auto + escola = autoescola

agro + indústria = agroindústria

(h) En palabras compuestas, cuando la primera palabra termina en vocal y la segunda palabra empieza por consonante, no hay guión separándolas:
ante + projeto = anteprojeto
agro + negócio = agronegócio
semi + círculo = semicírculo

(i) En palabras compuestas, cuando la primera palabra termina en vocal y la segunda palabra empieza por las consonantes <r> o <s>, no hay guión separándolas:
anti + rábico = antirrábico
anti + social = antissocial
ultra + som = ultrassom
auto + retrato = autorretrato

(j) En palabras compuestas, cuando la segunda palabra empieza por el grafema <h> hay guion separándolas:
anti + higiênico = anti-higiênico
super + homem = super-homem
sub + humano = sub-humano

(k) En palabras compuestas, cuando la primera palabra termina con una consonante y la palabra siguiente empieza con la misma consonante hay un guion separándolas:
inter + regional = inter-regional
super + resistente = super-resistente
hiper + requintado = hiper-requintado

(l) En palabras compuestas, cuando la primera palabra termina con una consonante y la palabra siguiente empieza con una consonante distinta, no hay guion separándolas:
super + sônico = supersônico
inter + municipal = intermunicipal
sub + solo = subsolo
sub + tenente = subtenente

(m) El prefijo <sub> cuando seguido por una palabra que empieza por <r> hay un guión de separación entre ellas:
sub + raça = sub-raça
sub + região = sub-região

(n) Cuando la primera palabra termina en consonante y la palabra siguiente empieza por vocal, no hay guión de separación:
super + interessante = superinteressante
inter + estadual = interestadual

(o) Cuando la palabra <mal> es seguida por vocal o por <h> hay guión de separación:
mal + humorado = mal-humorado
mal + estar = mal-estar
mal + agradecido = mal-agradecido

(p) Siempre hay guión en palabras compuestas con los prefijos: ex-, vice-, recém-, bem-, pós-, pré- y pró-:
ex-aluno
vice-diretor
recém-nascido
bem-humorado
pós-operatório
pré-vestibular
pró-aborto

Un ejemplo característico en PB es la palabra “subentendido”, en que hay la epéntesis de la vocal /I/ después del prefijo sub-.

3.4 Evaluación del sistema

En este capítulo se presentan las modificaciones realizadas en las reglas de transcripción fonética en el sistema de síntesis desarrollado por *Verbio Tecnologies S.L* para el portugués brasileño.

Para el trabajo fue necesario la comparación de dos diccionarios distintos: el diccionario desarrollado por *Verbio Technologies SL* y otro diccionario desarrollado por el Centro de Investigación y Desarrollo en Telecomunicaciones CPqD. La empresa CPqD Telecom & IT Solutions está ubicada em Campinas (Brasil) y para este trabajo fue utilizada la versión 1.4 de mayo de 2003.

Para fines ilustrativos es presentada una regla del sistema SEGRE empleada en este trabajo:

```
# Brasil, Lisboa, PIB...
<b> [b] -defecte
```

La primera línea están algunos ejemplos de cada grafema. Los ejemplos enseñados anteriormente se puede observar que el grafema <b> puede ocurrir en en la posisicín inicial, medial y final de la palabra. Los contextos en que hay el grafema <b> al final de lapalabra tratan sobretodo de palabras extranjeras o abrevisturas. Según las reglas de programación, todo lo que aparece después del símbolo # no será procesado por el sistema, pues esta representación indica que todo lo que hay después son comentários sobre los comandos.

En la segunda línia se puede verificar el comando informático escrito según el sistema SEGRE. Este comando puede ser leído de la siguiente manera: el grafema <b> pasa al fonema /b/ por defecto, es decir el fonema /b/ no tiene ninguén alófono y su representación de salida será siempre el fonema /b/. La regla por defecto es la última regla aplicada, suele ser la regla más general y en el listado aparece en la última posición.

El sistema procesa toda la secuendia del token, caracter a caracter, y de acuerdo con los contextos derecho y izquierdo de cada caracter son aplicadas determinadas reglas especificadas en el fichero de reglas. A continuación un ejemplo del procesamiento del convertidor grafema en fonema para el nombre propio Amanda:

```
# Amanda
<a> [6] -davant G_CONS -darrere <m> <n> -possil F
<m> [m] -defecte
<a> [6~] -davant G_CONS -darrere <m> <n> -possil F -tonica
```

<n> [] -davant G_VOCAL -darrere G_CONS
<d> [d] -defecte
<a> [6] -defecte

La tabla siguiente presenta cada grafema utilizado en PB y los correspondientes fonemas, para las variantes de São Paulo y de Rio de Janeiro:

SP	RJ
<a> [a] [6]	<a> [a] [6]
<b> [b]	<b> [b]
<c> [k] [s]	<c> [k] [s]
<d> [d] [dZ]	<d> [d] [dZ]
<e> [e] [E] [I] [j]	<e> [e] [E] [I] [j]
<f> [f]	<f> [f]
<g> [g] [Z]	<g> [g] [Z]
<h> [ ]	<h> [ ]
<i> [i]	<i> [i]
<j> [Z]	<j> [Z]
<k> [k]	<k> [k]
<l> [l]	<l> [l]
<m> [m]	<m> [m]
<n> [n]	<n> [n]
<o> [o] [O] [U]	<o> [o] [O] [U]
<p> [p]	<p> [p]
<q> [k]	<q> [k]
<r> [r] [R]	<r> [r] [R] [X]
<s> [s] [z]	<s> [s] [z] [S]
<t> [t] [tS]	<t> [t] [tS]
<u> [u]	<u> [u]
<v> [v]	<v> [v]
<w> [u] [v]	<w> [u] [v]
<x> [S] [ks]	<x> [S] [ks]
<y> [i]	<y> [i]
<z> [z]	<z> [z]

Tabla 4. Grafemas y fonemas de SP y RJ

Para la variante carioca fueron añadidos dos alófonos distindos: /X/ y /S/. En contexto de coda silábica, la vibrante alveolar sonora [r] y la fricativa alveolar sorda /s/ pasan a fricativa velar deshablaeada /X/ y a fricativa albeopalatal sorda /S/, respectivamente (Silva, 2002).

El grafema <a> tiene dos realizaciones en portugués brasileño: cuando está asociada a la vocal tónica utilizamos la vocal baja no redondeada [a] y cuando está asociado a las vocales átonas utilizamos la central no redondeada [6].

El grafema <e> en portugués brasileño tiene cuatro realizaciones distintas: en posición tónica pueden ser realizados los fonemas [e] y [E]. En posición de sílabas atona el /e/ final puede sufrir el proceso de reducción vocálica y pasar al fonema [I]. En situación de diptongos crecientes el /e/ se realiza como [j].

El grafema <o> puede realizarse de tres maneras distintas: cuando asociada a la silaba tónica se realizan los fonemas [o] y [O], en posición de sílabas atona el /o/ hay el proceso de de reducción vocálica y pasa a [U].

Los grafemas: <b>, <f>, <i>, <k>, <l>, <m>, <n>, <p>, <v>, <u>, <z> siempre son realizados como los fonemas [b], [f], [i], [k], [l], [m], [n], [p], [u], [v], [z] en cualquier de los contextos posibles.

El grafema <h> no presenta ninguna producción acústica y, por este motivo, no será convertido a ningún fonema en las reglas de grafema a fonema.

La representación grafémica [rr] utilizada en el diccionario elaborado por CPqD fue remplazada por [R], por ser más práctico al teclear el ordenador. En el variante paulista, el grafema <r> pasa a ser representado por los fonemas [r] en contexto de coda silábica y delante de:
, <cr>, <dr>, <fr>, <gr>, <pr>, <tr>. Tal grafema <r> puede ser representado también por el fonema [R] en contexto de ataque silabico. En el variante carioca, el grafema <r> pasa a ser representado por los fonemas [X] en contexto de coda silábica, a [R] en ataque silabico y a [r] delante de:
, <cr>, <dr>, <fr>, <gr>, <pr>, <tr>.

En el variante paulista, el grafema <s> pasa a ser representado por dos fonemas distintos [s] y [z]. El grafema <s> pasa al fonema [s] en contexto de coda y ataque silábico. El grafema <s> es representado pro el fonema [z] delante de las vocales orales, pues ocurre el proceso de sonorización de fricativa. En el variante carioca, hay trés representaciones para el grafema <s>: [s], [z] y [S]. El grafema <s> pasa al fonema [s] en contexto de ataque silábico, pasa a lfonema [z] delante de las vocales orales (sonorización de fricativa) y al fonema [S] en coda silábica.

El grafema <c> pasa al fonema [k] en ataque silábico cuando seguido por las vocales <a>, <o> y <u> y pasa a [s] cuando seguido por <e>, <i>.

Los grafemas <j>, <q>, <y> pasan a los fonemas [Z], [k], [i], respectivamente.

El grafema <g> pasa al fonema [g] en ataque silábico cuando seguido por las vocales <a>, <o> y <u> y pasa a [Z] cuando seguido por <e>, <i>.

En PB, los fonemas /t/ y /d/, cuando seguidos de [i, I, i~] se palatalizan y vuelven las consonantes africadas [tS] y [dZ]. Según Silva (2002: 129), “los alófonos o variantes dialectales de un fonema son identificados por intermedio del método de distribucióm complementar”. La autora describe que dos segmentos en distribución complementar ocurren en ambientes exclusivos, es decir, cuando una de las variantes es realizada, la otra no será. Además, tal proceso debe pasar a todas las palabras de la lengua.

Alcaim et al (1992: 24) considera que la clasificación de un sonido como fonema o alófono presupone un sistema de análisis de fonológico especifico del idioma analizado. Así, por ejemplo, en portugués brasileño la oclusiva alveolar sorda [t] y la africada alveolapatal sorda [tS] son alófonos del fonema /t/. En italiano, por ejemplo, la oclusiva alveolar sorda y la africada alveolapatal sorda no son alófonos, pero fonemas distintos, ya que su alternancia corresponde a una oposición de significados.

De este modo, el grafema <t>, por lo tanto, pasa a [tS] cuando seguido por <i>, <I> y <i~>. El grafema <t> se mantiene como [t] cuando es seguido por <a>, <e>, <o> y <u>. El grafema <d> pasa a [dZ] cuando seguido por <i>, <I> y <i~>. El grafema <d> permanece como [d] cuando es seguido por <a>, <e>, <o> y <u>.

El grafema <x> puede pasar a los fonemas [S] en contexto de ataque silábico y pasa a [ks] en coda silábica.

El grafema <w> puede pasar a los fonemas [u] o [v] dependiendo de los contextos de ocurrencia. Tal grafema <w> es más utilizado en nombres propios, ciudades o marcas extranjeras. Por este motivo, las palabras escritas con <w> están en el diccionario de excepciones.

Capítulo 4

4. Resultados

Este capítulo está dividido en dos sessiones distintas la primera parte se persentaran las modificaiones introducidas en el sistema de reglas del transcriptor grafema a fonema según las variaciones en la Normativa. La segunda parte dedicase a los analises acusticos del modulo prosódico.

4.1 Modificaciones introducidas

Las modificaciones del transcritor fonético desarrollado por *Verbio* están basadas principalmente en los cambios en las reglas de acentuación, en las vocales y en las consonantes, además de las palabras homógrafas.

4.1.1 La acentuación

4.1.1.1 El acento circunflejo

a) No se utiliza el acento circunflejo del primer *o* en palabras llanas finalizadas en <oo>, seguidas o no del morfema marcados del plural <s>:

abençôo, vôo, enjôo => abençoo, voo, enjoo

Regla graffon aplicada: excepción.
6 - b e~ - s 'o – U

abençôo, vôo, enjôo => abençoo, voo, enjoo

#voo
Regla graffon aplicada: <v> [v] -defecte
Regla graffon aplicada: <o> [o] –tonica –darrere <o>
Regla graffon aplicada: <o> [U] -davant <o> -posmot F
vo - U

b) No se utiliza el acento circunflejo de las formas verbales de la tercera persona del plural terminadas por <eem>:

vêem, lêem, dêem => veem, deem, leem.
Regla graffon aplicada: <v> [v] -defecte
Regla graffon aplicada: <e> [ ] -darrere <em$>
Regla graffon aplicada: <e> [e~] -darrer <m$>
Regla graffon aplicada: <m> [ ] -davant <e> <é> <ê> -postmot F
v - e~

4.1.1.2 La tilde

a) Los diptongos éi, ói dejan de ser acentuados en las palabras llanas:
idéia, jibóia => ideia, jiboia

jiboia...
Regla graffon aplicada: <o> [O] -darrere <ia$> <iam$> <ico$> <ica$><ide$> <ie$> <iem$> <ios$> <ito$>

ideia...
Regla graffon aplicada: <e> [E] -darrere <ia$> <iço$> <ica$>

b) Desaparece la tilde de las vocales i y u tónicas precedidas de diptongo:

baiúca, feiúra => baiuca, feiura.

baiuca, feiura...
<u> [w] -davant G_VOCAL

4.1.1.3 La diéresis

Según el acuerdo, se suprime la diéresis. Es decir, desaparece el signo ortográfico < ¨ > que se pone sobre la vocal *u* de las sílabas *gue, gui, que,* para representar la pronúcia de esta letra. Fueron añadidas las siguientes reglas grafema a fonema del nuevo transcriptor:

agüenta, lingüística, conseqüência => aguenta, linguistica, consequencia.

```
# averigue, apazigue, arguem. aguenta, antiguidade...
<u> [w] -davant <g> -darrere <en> <em> <e> <é> <ê> <id>
# aluguel, guepardo, guerra, alguém...
<u> [ ] -davant <g> -darrere <el> <ep> <er> <ém$> <i> <í> <î>
```

```
# averigue, apazigue, arguem. aguenta, antiguidade...
<u> [w] -davant <g> -darrere <en> <em> <e> <é> <ê> <id>
```

Las palabras derivadas de nombres extranjeros mantienen este signo. Por ejemplo: Müller, Mülleriano (incorporadas al diccionario de extranjerismos).

Las palabras “lingüística” y “lingüísticas” fueron incorporadas al diccionario de excepciones:

```
linguística
l i~ - g w 'i s - tS I - k 6
```

```
linguísticas
l i~ - g w 'i s - tS I - k 6 s
```

Fueron elaboradas reglas para las palabras que no llevan diéresis y al diccionario de excepciones fueron incorporadas las palabras en que el sistema transcribía mal. Además fue añadido al fichero de reglas un conjunto de sufijos que aparecen antes del grafema <qu>:

```
CONS_QU = <encia> <ência> <enta> <entar> <ível> <entado> <estr> <ilidade> <idar> <estro>
```

4.1.2 Las vocales

4.1.2.1 Las vocales nasales

Las siguientes reglas grafema a fonema fueron inseridas en nuevo transcriptor para la implementación de las vocales nasales en el sistema:

acabam, canhão, canhoto, antes, anã, anos, britânica, amador, planejando, mão, anos, andando...

<a> [6] -darrere <nh> -posmot I
<a> [6~] -darrere <nh>
<a> [6~] -darrere <ns$>
<a> [6~] -darrere <n+G_CONS>
<a> [6~] -davant G_CONS -darrere <m> <n> -possil F -tonica
<a> [6] -davant G_CONS -darrere <m> <n> -possil F
<a> [6~] -davant G_CONS -darrere <m> <n>
<a> [6~] -darrere <mb> <mp> -posmot I
<a> [6~] -darrere <m> -posmot I -tonica
<a> [6] -darrere <m> -posmot I
<a> [6~] -darrere <m+G_CONS>
<a> [6~] -davant G_VOCAL -darrere <m+G_VOCAL> -tonica
<a> [6] -davant G_VOCAL -darrere <m+G_VOCAL>
<a> [6~] -darrere <m+G_VOCAL>
<a> [6~] -darrere <m$>
<a> [6~] -darrere <no$> <nos$> <na$> <nas$>

Estas reglas se aplican también a las vocales <e>, <i>, <o> y <u>.

4.1.2.2 Las vocales epentéticas

Collischonn (2005: 17) retoma las consideraciones propuestas por Câmara Jr. (1969) sobre sílaba como una unidad constituida por una subida (ataque silábico), un ápice y de un declive (coda silábica). La vocal constituye el núcleo, pues presenta el grado más elevado de sonoridad. La subida es constituida por una o dos consonantes y el declive, a su vez, es formado por una de las siguientes consonantes /S/, /r/ o /l/ o por las semi-vocales /j/ o /w/.

Conforme Collischonn (2005: 17) los patrones silábicos del PB son los siguientes:

Patrones silábicos	Ejemplos
V	é
VC	ar
VCC	ins.tante
CV	cá
CVC	lar
CVCC	mons.tro
CCV	tri
CCVC	três
CCVCC	trans.porte
VV	au.la
CVV	lei
CCVV	grau
CCVVC	claus.tro

Tabla 5. Patrones silábicos del PB

De entre los patrones verificados anteriormente lo más común es la secuencia C V, así que en portugués brasileño, es común el proceso de epéntesis vocálica realizada por la inserción de la vocal [I] para que la secuencia CV sea mantenida. En palabras dónde hay coda no rellenada, hay la inserción de esta vocal.

Cagliari (2007: 118-119), presenta una lista de palabras con ejemplos tipicos dónde el fenomeno de la epentesis vocalica puede ocurrir. La lista pude ser observada a continuación:

b+p subproduto
b+t obter
b+d subdito
b+k subconsciente
b+m submarino
b+n abnegado
b+s absoluto
b+z obséquio
b+r sub-reptício
b+Z objeto
b+v óbvio
b+l sub-locação
p+t captou
p+s psicose
d+m admirar
d+v advogado
d+Z adjetivo
t+m ritmo
k+t compacto
k+s fixe
k+n técnica
g+m pigmeu
g+n ignorância
m+n amnésia
a+f afta

Las siguientes reglas grafema a fonema fueron inseridas en fichero REC del nuevo transcriptor para que el proceso de epéntesis vocalica sea implementada por el sistema:

bdélio, abdicar, bjarebyita, transubstanciação, submeter...
<b> [b+I] -darrere <d> <j> <s> <m>

mib, PIB...
<b> [b+I] -posmot F

cnidário, ctonógrafo, csiclovaíta, czar, acne, octogenário, hacquécia, acmo...
<c> [k+I] -darrere <b> <s> <t> <z> <q> <m> <n> <p> <ç> <c>

Graac, Puc...
<c> [k+I] -posmot F

adjunto, adquirir, advérbio, adnominal, adclividade, adfalangina, adgenerar, administrar, adpresso, adscrever, adventista, adzâneni, iddingsita. addisoniano ...
<d> [dZ+I] -darrere <j> <q> <v> <n> <c> <f> <g> <m> <p> <s> <v> <z> <d>

riad, adad...
<d> [dZ+I] -posmot F

afta, afvilita, afzélia, afwillita, naftalina...
<f> [f+I] -darrere <t> <v> <z> <w>

paf, pluft...
<f> [f+I] -posmot F

agdéstis, ágmen, agnóstico...
<g> [g+I] -darrere <d> <m> <n>

blog, log...
<g> [g+I] -posmot F

hiato...
<h> [ ]- defecte

taj...
<j> [Z+I] -posmot F

rock, pack, click...
<k> [+I] -posmot F

amnésia, amniótico...

<m> [m+I] -davant G_VOCAL -darrere <n>

ndendo, ngana, njango, nkumbi...

<n> [i+n] -posmot I -darrere G_CONS

psicólogo, pterodátilo, pneumonia...

<p> [p+I] -darrere <s> <t> <n>

top, bip, pop , Unicamp...

<p> [p+I] -posmot F

atmosfera, ritmo, étnico...

<t> [tS+I] -darrere <m> <n>

nit, net, pet...

<t> [tS+I] -posmot F

ambev, mev, telaviv...

<v> [v+I] -posmot F

Existe la necesidad de revisión de las reglas de separación silábica cuando hay vocales epentéticas. Según Braga *et al*. (2003: 1356), la importancia de la definición de reglas de segmentación silábica y de identificación de silabas tónicas para un sistema de conversión de texto en habla parte de la necesidad de identificación de la alternancia entre silabas tónicas y átonas, proceso que marca el ritmo de las frases y que afecta el nivel prosódico, ya que en la silaba tónica puede darse un incremento de la intensidad y de la duración.

4.1.2.3 Encuentros vocalicos

Fue añadida una regla al nuevo transcriptor, en la cual el primer <o> desaparece cuando antecedido por el grafema <co>.

álcool, coordenação, cooperação, coobrigação, coobação, coocupado, cooficiabilidade, cookeíta, coomologia, coonestar, coorte, coossificar...
<o> [ ] –davant <co>

4.1.3 Las consonantes

En este apartado serán descritas las reglas modificadas para las consonantes dobles y para la consonante <m>.

4.1.3.1 Las consonates dobles

Fueron adicionadas las reglas que delimitan las consonantes dobles del tipo <tt>, <bb> etc:

abba...
<b> [ ] -darrere <b>

acca...
<c> [ ] -darrere <c>

adda, iddingsita. addisoniano ...
<d> [ ] -darrere <d>

officer, office, tiffani...
<f> [ ] -darrere <f>

agga, maggi...
<g> [ ] -darrere <g>

ajja...
<j> [ ] -darrere <j>

akka...
<k> [ ] -darrere <k>

ballet, bolléa, ellus...
<l> [ ] -darrere <l>

commodity, emmonita
<m> [ ] -darrere <m>

anna...
<n> [ ] -darrere <n>

appa...
<p> [ ] -darrere <p>

aqqa...
<q> [ ] -darrere <q>

arraia, arredio, arrimo, arranha...
<r> [ ] -darrere <r>

pássaro, osso
<s> [ ] -davant <s>

wattômetro...
<t> [] -darrere <t>

avva...
<v> [] -darrere <v>

axxo...
<x> [K] -darrere <x>

puzzle, Ozzy
<z> [] -darrere <z>
<z> [s] -davant <z> -darrere G_CONS
<z> [z] -davant <z>

4.1.3.2 Las consonates nasales

Adición de la regla:

<m> [ ] -davant G_VOCAL -darrere <b> <p>

4.1.4 Palabras homógrafas

Desaparece el acento ortográfico en palabras llanas homógrafas de palabras átonas:

[pára, para] => [para] <p 'a – r 6>

[péla, pélas, pela] => [pela] <p e - l 6>

[pélo, pêlo, pêlos, pelo] => [pelo] <p e - l U>

[pólo] => [polo] <p O - l U>

La homografía puede producir errores en la producción acústica final del sintetizador, pues no hay la distinción de las vocales abiertas y cerradas, como se observa en los ejemplos a continuación:

Ele pela (VERBO) a maçã (vocal media baja)
Traducción al español: Él pela la manzana.

Vá pela (PREP) direita (vocal media alta)
Traducción al español: Ve por la derecha.

Eu pelo (VERBO) o tomate (vocal media alta)
Traducción al español: Yo pelo el tomate.

O pelo (SUSTANTIVO) do gato pode provocar alergia (vocal media alta)
Traducción al español: El pelo del gato puede causar alergia.

Como consecuencia de esta no unicidad entre grafema y fonema surgen algunos desafíos que deben ser solucionados en proceso de conversión de texto en habla.

Ejemplos de palabras homógrafas con distintas categorías morfológicas:

<selo> (estampilla - sustantivo) – [s 'e.l U]
<selo> (primera persona singular del verbo “estampillar”) – [s 'E.l U]

eu jogo (ó) x o jogo (ô), eu gosto(ó) x o gosto (ô), eu olho (ó)x o olho (ô), eu selo (ó), o selo (ô). Este problema puede ser resuelto por POS.

El transcriptor solamente considera las vocales cerradas:

Jogo (“juego”) Z 'o - g U
Olho (“ojo”) 'o - L U

Gosto (“gusto”)	g 'o s - t U
Selo (“estampilla”)	s 'e - l U

aeroporto	6 - E - r o - p 'o r - t U (singular)
aeroportos	6 - E - r o - p 'O r - t U s (aeroportos)

El sintetizador solo toma una transcripción como correcta, sería interesante desarrollar un POS para la desambiguizacion de estas palabras. La resolución de los problemas citados anteriormente depende de la formulación de reglas de resolución de la ambigüedad por medio de un recurso a un diccionario (donde estarán las entradas del léxico de la lengua).

4.1.5 Palabras funcionales

Las palabras funcionales de clase cerrada (preposiciones, conjunciones y pronombres) fueron divididas en dos grupos distintos: los monosílabos átonos y los que contienen dos o más silabas y, por este hecho, son acentuados.

pra	p r 6
por	p o r
que	k e
lhe	L e
lhes	L e s
lho	L o
lo	l o
se	s e

para	p 'a - r 6
porque	p o r - k 'e
porquê	p o r - k 'e
quem	k 'e~
pelo	p 'e - L U
pelos	p 'e - L U s
pela	p 'e - L 6
pelas	p 'e - L 6 s
sobre	s 'o - b r I
sob	s 'o - b I

4.2 Comparación entre los sistemas

Además de la adecuación de las reglas a la nueva ortografía de la lengua portuguesa, fue utilizado también un diccionario con palabras transcritas desarrollado por CPqD. Este diccionario fue utilizado como punto de partida para la definición de los fonemas y para la posterior elaboración de las nuevas reglas.

El primero transcriptor de *Verbio* presentaba los siguientes datos con relación a las palabras correctas y fonemas adecuados:

Primero transcriptor:
Palabras correctas = 33,19%
[Correctas = 3.352, Substituídas = 6.747, Total= 10.099]
Fonemas Correctas = 86,20%
[Correctos= 60.872, Apagados= 136, Substituídos= 9.609, Inseridos= 207, Total= 70.617]

La tasa de acierto del primero transcriptor para las palabras correctas es de 33,19% y de 86,20% para los fonemas. El total de palabras era de 10.099, siendo que el sistema transcribió adecuadamente 3.352 palabras y un total de 6.747 palabras presentaron algún tipo de substitución en la transcripción final.

Con relación a los fonemas, el porcentaje de acierto fue 86,20%, siendo que del total de 70.617 fonemas, el número de fonemas correctamente transcrito fue 60.872. La cantidad de fonemas apagados fue de 136, de fonemas remplazados fue 9.609 y el número de fonemas inseridos en las transcripciones fue 207.

Según los resultados obtenidos, a fecha de 17/04/2012, con el transcriptor del que se han introducido las modificaciones propuestas como resultado de este trabajo:

Segundo transcriptor:
Palabras Correctas = 93,18%
[Correctos= 9.408, Substituídos= 689, Total= 10.097]
Fonemas Correctos= 98,90%
[Correctos= 69.827, Apagados= 58, Substituídos= 718, Inseridos= 72, Total= 70.603]

Los testes fueron aplicados en el módulo de transcripción fonética del *front-end* del sintetizador de *Verbio*. Los resultados de las pruebas revelan que las tasas de éxito

de la transcriptor fonético nuevo son de 93,18% a 98,90% para las palabras y fonemas, respectivamente.

El total de palabras era de 10.097, siendo que el sistema transcribió adecuadamente 9.408 palabras y un total de 689 palabras presentaron algún tipo de substitución en la transcripción final.

Con relación a los fonemas, el porcentaje de acierto es de 98,90%, siendo que del total de 70.603 fonemas, el número de fonemas correctamente transcrito fue 69.827. La cantidad de fonemas apagados fue de 58, de fonemas remplazados fue 718 y el número de fonemas inseridos en las transcripciones fue 72.

Además de las nuevas reglas añadidas al sistema, fueron agregadas nuevas palabras a los diccionarios de excepciones y de extranjerismos, principalmente con relación a las palabras extranjeras, que, en la mayoría de los casos no eran transcritas de manera satisfactoria o pertinente por el transcriptor. Por esto, los valores totales de fonemas y palabras del primero y segundo transcriptores es distintos, porque fueron agregadas y/o extraídas algunas palabras del diccionario de excepciones y el diccionario de abreviatura y acrónimos.

El gráfico siguiente recoge los resultados obtenidos tras la incorporación de las modificacions explicadas en el apartado anterior respecto al índice de acierto del sistema antiguo en comparación con el nuevo sistema:

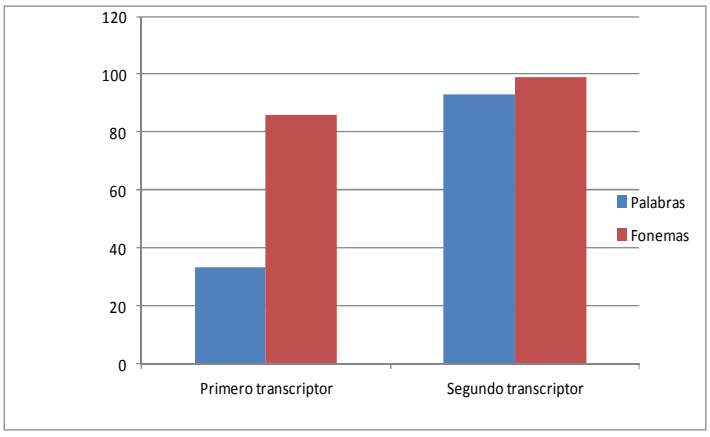

Gráfico 1. Comparación de los resultados entre el transcriptor antiguo y el nuevo

Como referencia de comparación de los resultados obtenidos fue escogido el convertidor grafema-fonema Ortofon del sintetizador AIURUETÊ desarrolladas por el grupo LAFAPE / IEL. Se trata de un sistema híbrido que usa un diccionario y las reglas lingüísticas para la integración en el sintetizador por concatenación. De todos los enfoques, los mejores resultados se exponen en Barbosa *et al*. (2003a), con una precisión del 98,43%. Según los autores, este módulo presenta una tasa de error de cerca de 4%, que puede ser tratado por medio del uso de diccionarios de excepciones.

Capítulo 5

5. Conclusiones y perspectivas futuras

El desarrollo de las tecnologías del habla depende del trabajo conjunto entre ingenieros de telecomunicación, programadores y lingüistas. La integración entre esas dos áreas se hace necesaria para la construcción de sistemas realmente eficientes, eficaces y aplicables.

Justificamos nuestro enfoque basado en reglas lingüísticas que nos apoyan bien en cuatro premisas: en primer lugar, el portugués es un idioma con muchas regularidades fonológicas, y en segundo lugar, un enfoque basado en normas es más económico en términos de memoria computacional de un enfoque de diccionario, tercero no existe un enfoque que ha utilizado métodos estadísticos o "aprendizaje automático" ha mostrado un rendimiento superior al 98% de los teléfonos correctamente transcritos, y por último, un enfoque basado en normas es siempre capaz de "leer" una nueva palabra.

Los cambios realizados en las reglas de transformación de grafema a fonema del módulo de transcripción fonética del sintetizador de *Verbio Technologies S.L* fueron fundamentales para la mejora de la producción acústica final del sistema.

La inclusión de las reglas mediante la nueva ortografía de la lengua portuguesa ha producido mejoras significativas en el resultado final y ha posibilitado la actualización del sistema.

El tratamiento de los procesos fonológicos y la descripción de las características fonéticas de las variantes tratadas en esta investigación también han producido resultados positivos en la salida del sintetizador, pues las modificaciones e inclusiones realizadas en las reglas grafema a fonema permiten una aproximación al habla natural.

Para trabajos futuros hay la posibilidad de estudios acerca del *part-of-speech* (POS) para el portugués brasileño y análisis más detallados de la expresión emocional en el habla sintetizada: un análisis acústico del fenómeno prosódico en los enunciados exclamativos.

6. Bibliografía

ALCAIM, A.; SOLEWICZ, J. A.; MORAES, J.A. de (1992) "Frequência de Ocorrência dos Fones e Listas de Frases Foneticamente Balanceadas no Português Falado no Rio de Janeiro". *Revista da Sociedade Brasileira de Telecomunicações*. Vol.7. nº 1, p.23-41.

BARBOSA. P. A. (1999). "Revelar a estrutura rítmica de uma língua construindo máquinas falantes: pela integração de ciência e tecnologia da fala". En: SCARPA, E. (org). *Estudos sobre prosódia*. Campinas: UNICAMP, p. 21-52.

BARBOSA, P. A. (2006). "Análise e modelamento dinâmicos da prosódia do português brasileiro". En: *IX CongressoNacional/ III Congresso Internacional de Fonética e Fonologia, (Belo Horizonte, 28 de noviembre de 2006),* [s.n.].

BARBOSA, P. A.; VIOLARO, F.; ALBANO, E.; SIMÕES, F.; AQUINO, P.; MADUREIRA, S.; FRANÇOSO, E. (1999). "Aiuruetê: A High-Quality Concatenative Text-to-Speech System for Brazilian Portuguese with Demisyllabic Analysis-Based Units and a Hierarchical Model of Rhythim Production". En: *Eurospeech'99 - 6th European Conference on Speech Communication and Technology, (Budapest 5-9 de septiembre de 1999), v.* V., [s.n.], p. 2059-2062.

BARROS, M. J. (2001). Estudo Comparativo e Técnicas de Geração de Sinal para a Síntese da Fala. Dissertação de Mestrado. Universidade do Porto.

BARROS, M. J.; MAIA, R.; TOKUDA, K.; RESENDE, F.; FREITAS, D. (2005). "HMM-based European Portuguese TTS System", *Proceedings of Interspeech 2005*. Lisboa, Portugal. pp. 2581-2584.

BENNETT, C.L. & BLACK, A.W. (2006). "The Blizzard Challenge 2006", *Blizzard Challenge Workshop, satellite event of Interspeech 2006 – ICSLP*. Pittsburgh, USA.

BLACK, A. & TOKUDA, K. (2005). "The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common databases", *Proceedings of Interspeech 2005*. Lisbon, Portugal. pp. 77-80.

BRAGA, D. (2008). “Algoritmos de Processamento da Linguagem Natural para Sistemas de Conversão Texto-Fala em Português”. Tesis Doctoral. Directora: Nieves Rodríguez Brisaboa. Universidad de la Coruña, Facultad de Filología.

BRAGA, D.; FREITAS, D.; FERREIRA, H. (2003). “Processamento Lingüístico Aplicado à Síntese da Fala”. En: *III Congresso Luso-Moçambicano de Engenharia – CMLE III, (Maputo 19-21 de agosto de 2003)*, [s.n.], p. 1349-1360.

BRAGA, D.; MARQUES, M. A. (2004). “The Pragmatics of the prosodic features in the political debate”, *Proceedings of the International Conference of Speech Prosody 2004*, 23-26 Março 2004, Nara, Japão. pp. 321-324.

BRAGA, D. & Dias, M. S. (2009). *Sistemas de Conversão Texto-Fala: estado da arte, aplicações, arquitectura e desafios.* Disponible en: <web.letras.up.pt/bhsmaia/EDV/apresentacoes/Braga_Texto_Fala.pdf> [Consultado el 30 de octubre de 2012].

BULUT, M.; NARAYANAN, S. S., & SYRDAL, A. K. (2002). “Expressive speech synthesis using a concatenative synthesizer”, *Proceedings of ICSLP 2002*, Denver, USA. pp. 79-84.

CABRAL, J. (2006). *Emotive Speech Synthesis*. Ms. Thesis. IST/INESC-ID, Lisboa.

CABRAL, J.; OLIVEIRA, L.C. (2006). “EmoVoice: a System to Generate Emotions in Speech”, *Proceedings of Interspeech 2006*. Pittsburgh, USA. pp. 1798-1801.

CAGLIARI, L. C. (2007). *Elementos de Fonética do Português Brasileiro*. São Paulo: Paulistana. 194 p.

CÂMARA JR, J. M. (1969). *Problemas de linguística descritiva*. Petrópolis: Hablaes. 71 p.

CAMPBELL, N. W. (1992). “Syllable-based segmental duration”. En: BAILLY, C.B. G.; SAWALLIS, T. R (eds). *Talking Machines: theories, models and designs*. Elsevier Science Publishers B. V.

CAHN, J. E. (1990). “The generation of affect in synthesized speech”, *Journal of the American Voice I/O Society*, Vol. 8, pp. 1–19.

CARVALHO, P.; TRANCOSO, I.; OLIVEIRA, L. C. (2003). “WFST based Unit Selection for Concatenative Speech Synthesis in European Portuguese”, *Proceedings of ICPhS'2003 - 15th International Congress of Phonetic Sciences*. Barcelona, Spain. Pp 2333-2336.

CASEIRO, D. A., Trancoso, I. (2002) “Grapheme-to-Phone Using Finite-State Transducers”, *Proceedings of 2002 IEEE Workshop on Speech Synthesis*. Santa Monica, USA.

CASEIRO, D. A.; Trancoso, I.; Viana, M. Céu; Barros, M. (2003). “A Comparative Description of GtoP modules for Portuguese and Mirandese using Finite State Transducers”, *Proceedings of ICPhS'2003 - 15th International Congress of Phonetic Sciences*, Barcelona, Spain. Pp 2605-2608.

Censo Demográfico 2010. En: Instituto Brasileiro de Geografia e Estatística [em línea]. *Sinopse do Censo Demográfico 2010*, abril de 2011. Disponible en Web: <http://www.ibge.gov.br/home/estatistica/populacao/censo2010/sinopse/default_sinopse.shtm> [Consultado el 01 de junio de 2012].

CHOTIMONGKOL, A. & BLACK, A. (2000). “Statistically trained orthographic to sound models for Thai”, *Proceedings of ICSLP2000*. Beijing, China. Volume 2, 551-554.

CONDADO, P. A. (2009). *Quebra de barreiras de comunicação para portadores de paralisia cerebral. Tesis de Doctorado*, Universidade do Algarve, Portugal.

COKER, CECIL H.; CHURCH, KENNETH W.; LIBERMAN, MARK Y. (1990). "Morphology and rhyming: Two powerful alternatives to letter-to-sound rules for speech synthesis", *Proceedings of the ESCA Workshop on Speech Synthesis*. Autrans, France. pp. 83-86.

COLLISCHONN, G. (2005). "A sílaba em português". En: BISOL, L. (org). *Introdução a estudos de fonologia do português brasileiro*. Porto Alegre: EDIPUCRS. p. 101-133.

EIDE, E.M., BAKIS, R., Hamza, W., & PITRELLI, J. F. (2004). "Towards synthesizing expressive speech", *Narayanan, S. S. and Alwan, A. (Eds.), Text to Speech Synthesis: New paradigms and Advances*. New Jersey: Prentice Hall.

FRASER, M.; KING, S. (2007). "The Blizzard Challenge 2007", *Sixth ISCA Workshop on Speech Synthesis*, Bonn, Germany.

HAMZA, W.; BAKIS, R.; EIDE, E.M.; PICHENY, M. A.; & PITRELLI, J. F. (2004). "The IBM expressive speech synthesis system", *Proceedings of ICSLP 2004*, Jeju, Korea. pp. 1099- 1108.

HENTZ, A. (2009). *Compressão de bancos de fala para sistemas de síntese concatenativa de alta qualidade.* Disertación de máster. Universidad Federal de Santa Catarina.

KAPLAN, R. M. & KAY, M. (1994). "Regular models of phonological rule systems", *Computational Linguistics 20(3)*, pp. 331-378.

KLATT, D. H. (1987). "Review of text-to-speech conversion for English". En: *Journal of the Acoustical Society of America, v.* 82, p.737-793.

LEE, S.; BRESCH, E.; ADAMS, J.; KAZEMZADEH, A.; & NARAYANAN, S. S. (2006). "A study of emotional speech articulation using a fast magnetic resonance imaging technique", *Proceedings of ICSLP 2006*, Pittsburgh, USA.

LLISTERRI, J.; AGUILAR, L.; GARRIDO, J. M.; MACHUCA, M. J.; MARÍN, R.; de la MOTA, C.; RIOS, A. (1999). "Fonética y tecnologías del habla". En: BLECUA, J.M.; CLAVERÍA, G.; SÁNCHEZ, C.; TORRUELLA, J. (eds), *Filología e informática. Nuevas tecnologías en los estudios filológicos*, Barcelona, Seminario de Filología e Informática, Departamento de Filología Española, Universidad Autónoma de Barcelona, Editorial Milenio, p. 449-479.

LLISTERRI, J. & MARTÍ ANTONÍN, M. A. (2002). *Tratamiento del Lenguaje Natural*. Barcelona: Edicions de la Universitat de Barcelona, S.L. Unipersonal, 208 p.

LLISTERRI, J.; CARBÓ, C.; MACHUCA, M. J.; de la MOTA, C.; RIERA, M.; RIOS, A. (2004). "La conversión de texto en habla: aspectos lingüísticos". En: MARTÍ, M.A. y LLISTERRI, J. (eds). *Tecnologías del texto y del habla.* Barcelona: Edicions de la Universitat de Barcelona – Fundación Duques de Soria (UB, 72), p. 145-186.

LUCASSEN, J. M. & MERCER, R. L. (1984). "Discovering Phonemic based forms automatically: an information theoretic approach", *IEEE International Conference on Acoustics, Speech and Signal Processing*. pp. 42.5.1-42.5.4.

MAIA, R. (2006). "Speech Synthesis and Phonetic Vocoding for Brazilian Portuguese Based on Parameter Generation form Hidden Markov Models". Tesis Doctoral. Departamento de Ciencias de la Computación e Ingeniería, Instituto de Nagoya Tecnología, Nagoya, Japón.

MAIA, R.; ZEN, H.; TOKUDA, K.; KITAMURA, T. & RESENDE Jr., F. G. V. (2003). "Towards the development of a Brazilian Portuguese text-to-speech system based on HMM". En: *Proceedings of the European Conference on Speech Communication and Technology EUROSPEECH 2003*, (Ginebra, 1-4 de septiembre de 2003). [s.n.], p. 2465–2468.

MAIA, R.; ZEN, H.; TOKUDA, K.; KITAMURA T. & RESENDE Jr., F. G. V. (2006). "An HMM-based Brazilian Portuguese speech synthesizer and its characteristics". *IEEE Journal of Communication and Information Systems*. Nº 2, *v.* 21, p. 58-71.

MARTINO, J. M. (2005). “Animação Facial Sincronizada com a Fala: Visemas Dependentes do Contexto Fonético para o Português do Brasil”. Tese de Doctorado. DCA/FEEC/UNICAMP, 2005.

MATTE, A.C.F.; MEIRELES, A. & FRÁGUAS, C. C. (2006). “SILWeb – analisador fonológico silábico-acentual de texto escrito”, *Revista de Estudos Lingüísticos*, v. 14, nº 1, p. 31-50.

MENG, H. M.; SENEFF, S. ; ZUE, V. (1994). “Phonological parsing for bi-directional letter-to-sound/sound-to-letter generation”, *ARPA Human Language Technology Workshop*. Princeton, USA.

NICODEM , M. V.; Seara, R.; Pacheco, F. S. (2005). “Reducing the Natural Click Effect within Database for High Quality Corpus-Based Speech Synthesis", *8th IEEE International Symposium on Signal Processing and its Applications*. Sydney, Austrália. pp. 607-610. Nicodem; M. V. , Kafka , S. G.; Seara Junior , R.; SEARA, R. (2007). "Refinamento da Segmentação Fonética em Aplicações de Síntese de Fala", *XXV Simpósio Brasileiro de Telecomunicações (SBrT 2007*). pp.1-6.

OLIVEIRA, C.; MOUTINHO, L.; TEIXEIRA, A. (2004). “Um novo sistema de conversão grafema-fone para PE baseado em transdutores”, *Actas do II Congresso Internacional de Fonética e Fonologia*, Maranhão, Brasil (no prelo).

OLIVEIRA, L. C., VIANA, M. C., TRANCOSO, I. M. (1991). “DIXI - Portuguese Text-to-Speech System”, *Proceedings of EUROSPEECH'91 - 2nd European Conference on Speech Communication and Technology*, Genoa, Italy. pp.1239-1242.

OLIVEIRA, L.; VIANA, M. C.; MATA, A. I. & TRANCOSO, I. (2001). Progress report of project dixi+: A portuguese text-to-speech synthesizer for alternative and augmentative communication. Technical report, FCT. Oliveira, L. C. 1996. *Síntese de Fala a Partir de Texto*. Dissertação de Doutoramento. Universidade Técnica de Lisboa.

OLIVEIRA, L. C. (1996). *Síntese de Fala a Partir de Texto*. Tesis de Doctorado. Universidad Técnica de Lisboa.

PACHÈS, P.; de la MOTA, C.; RIERA, M.; PEREA, M. P.; FEBRER,A.; ESTRUCH, M.; GARRIDO, J. M.; MACHUCA, M. J.; RIOS, A.; LLISTERRI, J.; ESQUERRA, I.; HERNANDO, J.; PADRELL J.; NADEU, C. (2000). "Segre: an automatic tool for grapheme-to-allophone transcription in catalan", en Ó Cróinín D. (ed.) *Proceedings of the Workshop on Developing Language Resources for Minority Languages: Reusability and Strategic Priorities (LREC-2000 Second* International Conference on Language Resources and Evaluation), Atenas (Grecia), 30 de mayo de 2000, pp.52-61.
Disponible en Web:
<http://liceu.uab.es/~joaquim/publicacions/Paches_et_al_00_SEGRE_Phonetic_Transcription_Catalan.pdf> [Consultado el 22 de abril de 2012].

PARODI, Giovanni (2008). Lingüística de corpus: una introducción al ámbito [en línea]. *RLA- Revista de Lingüística Teórica y Aplicada*, vol. 46, n.1, pp. 93-119 <http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-48832008000100006&lng=es&nrm=iso> [Consultado el 17 de octubre de 2011].

RAIMUNDO, G.; CABRAL, J. MELO, C.; OLIVEIRA, L.; PAIVA, A.; TRANCOSO, I. (2007). "Telling Stories with a Synthetic Character: Understanding Inter-modalities Relations", *COST Action 2102 International Workshop on Verbal and Nonverbal Communication Behaviours*. Heidelberg: Springer. pp. 310-323.

ROCHE E. & SCHABES, Y. (1995). *Exact Generalization of Finite-State Transductions: Application to Grapheme-to-Phoneme Transcription*. Technical Report TR-95-08, Mitsubishi Electric Research Laboratories. Cambridge, USA.

SANTOS, Diana (2001). "Processamento da linguagem natural: uma apresentação através das aplicações", Ranchhod (org.) 2001. Tratamento das Línguas por Computador. Uma introdução à Lingusitica Computacional e suas aplicações. Lisboa: Caminho.

SEJNOWSKI, T. J. & ROSENBERG, C. R. (1987). "Parallel networks that learn to pronounce English Text", *Complex Systems, 1*, pp. 145-168.

SILVA, Thaïs Cristófaro (2002). *Fonética e Fonologia do Português: roteiro de estudos e guia de exercícios*. São Paulo, SP: Contexto, 275 p.

SCHRODER, M. (2006). "Expressing degree of activation in synthetic speech". *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4), 1128–1136.

SEARA, I. C.; NICODEM, M. V.; Seara, R.; SEARA JUNIOR, R. (2007). "Classificação Sintagmática Focalizando a Síntese de Fala: Regras para o Português Brasileiro", *XXV Simpósio Brasileiro de Telecomunicações (SBrT 2007).* pp. 1-6.

SILVA, Solimar de Souza (2004). "Um estudo de modelos básicos de prosódia para o português brasileiro". Tesis de Maestría. Director: Sérgio Lima Neto. Universidad Federal del Rio de Janeiro, Facultad de Ingeniaría Eléctrica.

SIMÕES, Flávio Olmos, VIOLARO, Fábio; BARBOSA, Plínio Almeida y ALBANO, Eleonora Cavalcante (2000). "Um Sistema de Conversão Texto-Fala para o Português Falado no Brasil". *Revista da Sociedade Brasileira de Telecomunicações*. *v*. 15, nº 2, p. 70-77.

TAYLOR, Paul (2005). "Hidden Markov Models for Grapheme to Phoneme Conversion", *Proceedings of Interspeech 2005*, Lisboa, Portugal. pp .1973-1976.

TAYLOR, Paul (2007). *Text-to-Speech Synthesis*. Disponible en Web: <http://mi.eng.cam.ac.uk/~pat40/book.html > [Consultado el 22 de abril de 2012].

TEIXEIRA, A., OLIVEIRA, C., MOUTINHO, L. (2006a). "Machine Learning of European Portuguese Grapheme-To-Phone Conversion using a Richer Feature Set", *Revista do DETUA*, vol. I, nº 1, Aveiro.

TEIXEIRA, A., OLIVEIRA, C., MOUTINHO, L., (2006b). "On the Use of Machine Learning and Syllable Information in European Portuguese Grapheme- Phone Conversion", *R. Vieira, P. Quaresma, Maria G. V. Nunes, N. Mamede, C. Oliveira, M. C. Dias (Eds), Computacional Processing of the Portuguese Language.*

TEIXEIRA, J. P. (1995). *Modelização Paramétrica de Sinais Para Aplicação em Sistemas de Conversão Texto-Fala*. Tese de Mestrado. Faculdade de Engenharia da Universidade do Porto.

TEIXEIRA, J. P.; FREITAS, D. (1998). "MULTIVOX- Conversor Texto-Fala para Português", *Proceedings of PROPOR'98*, Porto Alegre, Brasil.

TEIXEIRA, J. P.; GOUVEIA, P., FREITAS, D. (2000). "Divisão silábica automática do texto escrito e falado", *Proceedings of PROPOR'2000*. Atibaia, SP. Brasil.

TEIXEIRA, , J. P.; BARROS, M. J. y FREITAS, D. (2003). "Sistemas de Conversão Texto-Fala". En: *III Congresso Luso-Moçambicano de Engenharia – CMLE III, (Maputo 19-21 de agosto de 2003).*

TEIXEIRA, , J. P.; BARROS, M. J. & FREITAS, D. (2003) "Sistemas de Conversão Texto-Fala". En: *III Congresso Luso-Moçambicano de Engenharia – CMLE III, (Maputo 19-21 de agosto de 2003).*

TEIXEIRA, J. P. (2004). *A Prosody Model to TTS Systems*. PhD Thesis, Faculdade de Engenharia da Universidade do Porto.

TENANI, L. E. (2002). "Domínios prosódicos no português do Brasil: implicações para a prosódia e para a aplicação de processos fonológicos". Tesis Doctoral. Directora: Maria Bernadete Marques Abaurre. Universidad Estatal de Campinas, Instituto de los Estudios en Lenguaje.

TOKUDA, K. (2004). "An HMM-Based Approach to Multilingual Speech Synthesis," Shrikanth Narayanan, Abeer Alwan (eds*). Text-to-Speech Synthesis: New Paradigms and Advances*. New Jersey: Prentice Hall.

TOKUDA, K.; MASUKO, T; YAMADA, T.; KOBAYASHI, T.; e IMAI, S. (1995). "An algorithm for speech parameter from continuous mixture HMM with

dynamic features". En: *Proceedings of Eurospeech 95, (Madrid 18-21 de septiembre de 1995)*. [s.n.], p. 757-760.

TRANCOSO, I.; VIANA, M. C.; SILVA, F.; MARQUES, G. & OLIVEIRA, L.(1994). "Rule-based vs. neural network based approaches to letter-to-phone conversion for Portuguese common and proper names", *Proceedings of ICSLP'94*, Yokohama, Japan. pp. 1767-1770.

TRILLA, A. (2009). "Natural language processing techniques in text-to-speech synthesis and automatic speech recognition". In:*Working Paper, Engenieria y Arquitectura La Salle*, Universitat Ramon Llull, Barcelona.

TUFANO, D. (2008). Guia Prático da Nova Ortografia: saiba o que mudou na ortografia brasileira. São Paulo: Editora Melhoramentos.

VIOLARO, F. & BÖEFFARD, O. (1994). "Using a Hybrid Model in a Text-To-Sppech System to Enlarge Prosodic Modifications", *Proceedings of ICSLP 94*. Yokohama, Japan. pp.727-730.

VIOLARO, F. & BÖEFFARD, O. (1998). "A Hybrid Model for Text-to-Speech Synthesis". *IEEE Transactions on Speech and Audio Processing*, *v*. 6, nº 5, p. 426-434.

VOLP (Vocabulario Ortografico de la Lengua Portuguesa). Disponible en: <http://www.academia.org.br/abl/cgi/cgilua.exe/sys/start.htm?sid=23> [Consultado el 22 de abril de 2012].

WEISS, L. C.; OLIVEIRA, L.; PAULO, S.; MENDES, C.; FIGUEIRA, L.; VALA, M.; SEQUEIRA, P.; PAIVA, A.; VOGT, T.; ANDRE, E. (2007). "ECIRCUS: Building Voices for Autonomous Speaking Agents", *6th Speech Synthesis Workshop*, ISCA.

WELLS, J. (1996). *SAMPA for Portuguese* [Página web]. London: Division of Psychology and Language Sciences, University College London. (Primera edición: 20/09/1995) Disponible en: <http://www.phon.ucl.ac.uk/home/sampa/portug.htm>. [Consultado el 15/12/2011].

ZÁGARI, M.R.L. (2005). “Os falares mineiros: esboço de um atlas lingüístico de Minas Gerais”. En: Aguilera V. (ed.). A geolinguística no Brasil: trilhas seguidas, caminhos a percorrer. Londrina: Editora da Universidade Estadual de Londrina, vol. 1, p. 45-72.

7. Anejos

7.1 Diccinario

alexander	6 - l e k - s 6~ - d 'e r
ambev	6~ - b 'E - v i
american	6 - m 'E - r i - k 6~
amro	6~ - R 'U
amsterdã	6~ s - t e r - d 6~
antárctica	6~ - t 'a r - tS I - k 6
ballet	b6 - l 'E
bank	b 'e~ - k I
bankboston	b e~ - k - b 'O s - t o~
banrisul	b 6~ - r I - z 'u w
bauxi	b 6 w - S 'i
bayer	b 'a j - e r
belfast	b E w - f '6 s - tS I
belgrado	b E w - g r '6 - d U
belmonte	b E w - m 'o~ - tS I
beliche	b e - l 'i - S I
beltrão	b e w - t r '6~ w
bicbanco	b `I - k I - b '6~ - k U
bilbao	b I w - b 'a w
birigui	b I - r I - g u 'j
blues	b l 'u s
bob's	b 'O - b I s
bobs	b 'O - b I s
bosch	b 'O - S I
bosque	b 'O s - k I
boston	b 'O s - t o~
braun	b r '6 u~
byte	b 'a j - tS I
bytes	b 'a j - tS I s
carrefour	k 6 - R e - f 'u r
cassiane	k 6 - s j '6~ - n I
chique	S 'i - k I
chiques	S 'i - k I s
chofer	S o - f 'E r
citibank	s i - tS I - b 'e~ - k I
commodity	k o - m 'O - dZ i - tS I
comsat	k o~ - s 'a - tS I
d'água	d 'a - g w 6
dell	d 'E w
dinners	d 'a j - n e r s
dow	d 'a w
drummond	d r u~ - m 'o~
dumont	d u - m 'o~
duratex	d U - r 6 - t 'E k s
eaton	'i - t o~
einstein	6 i~ s - t 6 'i~
electric	l 'E - t r I - k I
electrolux	e - l E - t r o - l 'u - k s
express	'E k s - p r E s
extra	'E s - t r 6
faxes	f 'a - k i - s I s
george	Z e - 'O r - Z I
glaxo	g l 'a k - s U
google	g 'u - g o w
gols	g 'o w s
grow	g r 'o w
habib	R a - b 'i - b I
habibs	R a - b 'i - b I s
habib's	R a - b 'i - b I s
hackers	R 'a - k e r s
haicai	R '6 j - k 'a j
haicais	R '6 j - k 'a j s
hannah	R 'a~ - n 6
hashi	R 6 - S 'i
honda	R 'o~ - d 6
hostal	R 'O s - t a w
hostals	R 'O s - t a w s
hummus	R 'u - m u s
hyundai	R i w~ - d 'a j
jeans	dZ 'i~ s
kramer	k r '6~ - m e r
lee	l 'i
lexotan	l E - k i - s 'o - t '6~
light	l 'a j - tS I
lights	l 'a j - tS I s
madrid	m 6 - d r 'i
magneti	m 6 - g I - n 'E - tS i
mahal	m 6 - R 'a w
mahatma	m 6 - R 'a - tS i - m 6
mahayana	m 6 - R 6 - j 'a~ - n6
mam	m '6~
marketing	m 'a r - k e - tS i~
mastercard	m 6 s - t e r - k 'a r
mcDonald's	m E k - d 'o - n a w - dZ I
mercedes	m e r - s 'e - dZ I s
mercedes-benz	m e r - s `e - dZ I z - b 'e~ s
mercury	m 'E r - k U - r I
miami	m a j - '6~ - m I
microsoft	m 6 j - k r o - s 'O - f i - tS i
mitsubishi	m `I t - s U - b 'I - S i

mitsui	m I t - s 'u j
miyashiro	m I - j 6 - S 'i - r U
mobile m	'O - b I - l I
motors	m 'O - t o r s
munique	m U - n 'i - k I
nagasaki	n 6 - g 6 - z 'a - k i
net	n 'E - tS i
new	n 'i w
newtonn	I w - t 'o~
nhoque	J 'O - k I
nhoques	J 'O - k I
oxford	'O k s - f O r
pajero	p 6 - Z 'e - r U
pet	p 'E - tS I
rousseff	R u - s 'E - f i
sabesp	s a - b 'E s - p I
samsung	s 6~ - s 'u~ - g I
sexy	s 'E - k s I
shampoo	S 6~ - p 'u
shell	S 'E w
shopping	S 'O - p i~
show	S 'o w
shows	S 'o w s
showmício	S o w - m 'i - s i w
showmícios	S o w - m 'i - s i w s
siemens	s 'i j - m e~ s
software	s 'O f - tS i - w E r
softwares	s 'O f - tS i - w E r s
spaghetti	I s - p 6 - g 'E - tS i
taj	t 'a - Z I
tam	t '6~
tchau	tS 'a w
tchecoslováquia	tS E - k o s - l o - v 'a - k j 6
telesc	t e - l 'E s - k I
telesp	t e - l 'E s - p I
telaviv	t e - l 6 - v 'i - v I
TV	t e - v 'e
texas	t 'E k - s 6 s
twitter	t w 'i - t e r
uísque	w 'i s - k I
varig	v 'a - r i - g I
vasp	v 'a s - p I
volksvagen	v `o w k s - v 'a - g e~
wanderdey	v 6~ - d e r - l 'e j
warner	w O r - n 'e r
washington	w O - S i~ - t 'o~
watt	v 'a - t S I
watts	v 'a - t S I s
web	w 'E - b I
webs	w 'E - b I s
wet'n'wild	w `E - t a~ - w 'a j l d
white	w 'a j t
williams	w 'i - l j 6~ s
yakult	j 6 - k 'u w - tS I
yamaha	j 6 - m 'a - R 6
yin yang	i~ - j 'a~ - g I
zeneca	z e - n 'E - k 6
zogbi	z 'o - g I - b i
zuleica	z U - l 'e j - k 6
zurique	z U - r 'i - k I