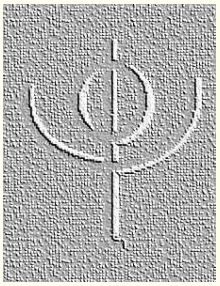
ISSN: 0325-2043



Laboratorio de Investigaciones Sensoriales (LIS)

Informe XLVI–2013

I N I G E M

CONICET

U B A

Instituto de Inmunología, Genética y Metabolismo
Córdoba 2351, Piso 9, (1121), Buenos Aires
Tel/Fax: 5950-9024
lis@fmed.uba.ar — http://www.lis.secyt.gov.ar

Índice

1. Introducción **1**

2. Personal **1**

3. Proyectos de Investigación **2**
3.1. CONICET PIP Nro. 5897/06: Análisis de las sensaciones de dulce, agrio y amargo en soluciones puras y mezcladas en medio acuoso y alcohólico 2
3.2. Proyecto de Cooperación MINCYT-BMBF (AL/12/06) Utilización de Prosodia para el Reconocimiento Automático de Habla 2

4. Proyectos de I+D **4**
4.1. PID 094/2007 - Desarrollo de un Sistema de Conversión de Texto a Habla . . 4
4.2. PID 35891 - Desarrollo de las técnicas de reconocimiento del hablante para su aplicación a nivel forense 4

5. Docencia **4**
5.1. Cursos de posgrado 4
5.2. Seminarios en el laboratorio 5

6. Intercambio Científico **5**
6.1. Visita de Investigadores al LIS 5
6.2. Estadías en el Exterior de Investigadores del LIS 5

7. Tesis **6**
7.1. Doctorales en curso 6

8. Actividades de Divulgación **9**

9. Publicaciones **9**
9.1. Revistas 9
9.2. Informes Técnicos 9

Apéndice **11**

A. Resúmenes de Trabajos **11**
A.1. Psychophysical Assessments of Sourness in Citric Acid-Ethanol Mixtures. *Guirao, M. et al.* 11

B. Informes Técnicos **11**
B.1. Guía para la segmentación y transcripción fonética para las tecnologías de habla. *Gurlekian J.A. y Torres H.M.* 12
B.2. Definición de las funciones de costo para la selección de unidades en AROMO *Torres H.M.* 31
B.3. Segmentación de audio de radio y televisión. *Torres H.M.* 43
B.4. Internal vs. external standards for the evaluation of F0 perturbation. *Gurlekian J.A., Torres, H.M. y Vaccari, M.E.* 46

B.5. Medición preliminar de la velocidad de conversión del sistema TTS AROMO. *Torres H.M.* 51

1. Introducción

Desde su creación en al año 1968, el LIS publica un informe anual en donde se consignan las publicaciones realizadas, los trabajos en curso, la actividad docente y el intercambio científico.

Los Informes LIS están registrados bajo ISSN 0325-2043 (International Standard Serial Number), a través de Latindex[1], reconocido internacionalmente para la identificación de las publicaciones seriadas. La serie comienza con el Informe I-1968, Laboratorio de Investigaciones Sensoriales, CONICET.

En los informes aparecen siglas que referencian las sedes del LIS, primero en el Hospital Escuela (HE), luego en la Facultad de Medicina (FM) y, actualmente, en el Hospital de Clínicas (HC) de la Universidad de Buenos Aires.

Desde el año 1997, los informes también están disponibles a través del sitio web del laboratorio: http://www.lis.secyt.gov.ar/.

2. Personal

Investigadores

- EVIN Diego, Bioingeniero, Dr. en Ciencias de la Computación.
- GUIRAO Miguelina, Prof. Filosofía, Dra. en Psicología Experimental.
- GURLEKIAN Jorge A., Ing. Electrónico, Dr. en Medicina. Responsable del LIS.
- TORRES Humberto, BioIngeniero, Dr. en Ingeniería.

Investigadores que participan en proyectos que se desarrollan en el LIS:

- CALVIÑO Amalia M., Farmacéutica, Dra. en Bioquímica.
- GRAVANO Agustín, Licenciado y Dr. en Ciencias de la Computación.
- VACCARI María Elena, Lic. en Fonoaudiología.

Becarios

- COSSIO MERCADO Christian, Ing. en Informática, Becario FONCYT. Tesista de Doctorado UBA.
- MARTINEZ SOLER Miguel, Ing. en Informática, Becario CONICET. Tesista de Doctorado UBA.

Tesistas

- TRIPODI Mónica, Lingüista. Tesista de Doctorado UBA.
- UNIVASO Pedro, Ing. Electrónico. Tesista de Doctorado UBA.

[1] Sistema Regional de Información en Línea para Revistas Científicas de América Latina, el Caribe, España y Portugal. Sitio: http://www.latindex.unam.mx

3. Proyectos de Investigación

3.1. CONICET PIP Nro. 5897/06: Análisis de las sensaciones de dulce, agrio y amargo en soluciones puras y mezcladas en medio acuoso y alcohólico

Direccion: Miguelina Guirao
Codirección: Amalia Mirta Calviño

Efecto del etanol en el gusto con y sin atributos trigeminales

Hasta el momento los estudios sobre la influencia del etanol en el sabor se habían dirigido a bebidas alcohólicas en las que se mezclan diferentes sustancias gustativas. En cambio no se tenía un conocimiento acabado acerca de la interacción etanol-gusto en sustancias puras. Los pocos intentos que se han realizado discrepan en los resultados. Algunos autores han encontrado que el efecto del etanol no es significativo y otros que los resultados dependen del método. En nuestro caso hemos investigado los cambios que se producen en el gusto cuando se le agrega etanol a tres sustancias puras: el dulce de la sacarosa, el agrio del ácido cítrico y el amargo de la cafeína. Para ese fin hemos experimentado con dos dimensiones de la sensación la intensidad del gusto y la duración o persistencia del gusto en la cavidad oral. Además para descartar una posible influencia del procedimiento elegido aplicamos tres métodos psicofísicos diferentes.

En general los resultados revelan que el efecto del etanol depende en gran medida de la concentración de la sustancia y de la graduación del alcohol. Sobre la base de los datos obtenidos se comparara el efecto que tiene el etanol en cada uno de los tres gustos y la posible influencia de los atributos trigeminales en la modificacion del gusto.

3.2. Proyecto de Cooperación MINCYT-BMBF (AL/12/06) Utilización de Prosodia para el Reconocimiento Automático de Habla

El Reconocimiento Automático del Habla (RAH) es un componente esencial en la construcción de interfaces hombre–máquina naturales y eficientes. De acuerdo a los planteos psicolingüísticos actuales la información prosódica es un componente que actua con simultaneidad —en paralelo— con la identificación de fonemas y palabras. Este conocimiento aún no se aprovecha de manera eficiente en el proceso de RAH. Si bien el empleo de información prosódica ha recibido un notable interés por parte de la comunidad científica, aún no se ha podido integrar en los sistemas de reconocimiento actuales mas exitosos como lo son los modelos probabilisticos de Markov.

Los objetivos de este proyecto son:

- La determinación automática de prominencias acentuales dentro de frases del habla que permitan identificar silabas y palabras remarcadas por cada locutor en su discurso.
- La determinación automática de los acentos de frontera emitidos al final de las frases que permitan segmentar e identificar la modalidad de cada frase.

- La comparación de los resultados contrastando la manifestación acústica de sílabas prominentes y de la modalidad en alemán y español.
- La incorporación de esta información prosódica, en paralelo con los sistemas de RAH estándar basados en modelos ocultos de Markov (HMM) actuando como activadores o inhibidores en la etapa de busqueda de las palabras mas probables, tal que resulten en una mejora del reconocimiento sin empeorar el tiempo de procesamiento. Este es el objetivo central del proyecto.
- El desarrollo informático de la incorporacion de la prosodia en la etapa de busqueda de palabras candidatas como complemento del modelo de lenguaje utilizado.
- La exploración y desarrollo de la incorporación de información prosódica en la etapa de preprocesameinto y de selección de hipotesis como complemento del objetivo central relativas al sistema de referencia.

La metodología a utilizar en el desarrollo del proyecto será:

- Se diseñará un sistema de referencia para el reconocimiento del habla basado en modelos ocultos de Markov.
- Se evaluarán diferentes arquitecturas de Redes Neuronales Artificiales, clasificadores fuzzy y/o Máquinas de Soporte Vectorial para la caracterización automática de curvas de entonación, segmentación de frases y sílabas, y clasificación de silabas prominentes.
- Se analizará la contribución de los sistemas propuestos para el empleo de información prosódica en el RAH respecto a los resultados de los sistemas de referencia empleando validación cruzada.

En resumen, el plan de trabajo será el siguiente:

- Los doctorandos de Alemania se entrenarán en Buenos Aires en el desarrollo de los sistemas de referencia de reconocimiento automatico dado que el LIS tiene una larga experiencia en la utilizacion de los HMM.
- Los doctorandos de Argentina se entrenaran en Berlín en la incorporacion a nivel algorítmico de la prosodia en los sistemas HMM y en la evaluacion de mejores estimadores de F0 en la etapa de preprocesamiento.
- Los investigadores principales estarán a cargo de la dirección de los estudiantes y de las modificaciones que serían necesarias en el proyecto. Discutirán las plataformas para la evaluacion de los desarrollos. Los directores del proyecto se entrevistaran con los profesionales de la universidad de Dresden con experiencia probada en la realizacion de patentes tecnologicas asociadas al tema, y se reunirán con investigadores de Deutsche-Telekom para coordinar tareas de evaluación de calidad de los productos desarrollados.

Entre las tareas especificas a desarrollar por los doctorandos se encuentran:

- Implementar y evaluar algoritmos para la segmentación automática en sílabas empleando anotaciones manuales durante el entrenamiento de los mismos

- Implementar algoritmos para la estimación de silabas promienentes a partir de las curvas de frecuencia fundamental y energía
- Evaluar los algoritmos desarrollados para extracción automática de rasgos prosódicos empleando el corpus oral disponible
- Proponer y evaluar usos alternativos de la información prosódica en los modelos acústicos, de lenguaje, y/o en el diccionario de pronunciaciones de sistemas de RAH basados en modelos ocultos de Markov
- Ciclos de propuesta, diseño, experimentación, análisis de resultados y pruebas de validación cruzada.

4. Proyectos de I+D

4.1. PID 094/2007 - Desarrollo de un Sistema de Conversión de Texto a Habla

PAE Nro: 37122, PID 2007. Nro. 094.FONCYT. Desarrollo de un sistema de conversión de Texto a Habla
Director: Jorge A. Gurlekian
Período: 2009-2013
Unidad de Ejecución: Laboratorio de Investigaciones Sensoriales
Institución de la que depende la Unidad de Ejecución: CONICET
Entidad Acreditadora y/o Financiadora: FONCYT
Financiamiento obtenido: 258.200 pesos. Costo total 780.200 pesos

4.2. PID 35891 - Desarrollo de las técnicas de reconocimiento del hablante para su aplicación a nivel forense

Director: Jorge A. Gurlekian
Período: 2009-2013
Secretaria de Ciencia y Técnica. Proyectos de Investigación y Desarrollo PID
Entidad adoptante: Policía Científica, Gendarmería Nacional Argentina.

5. Docencia

5.1. Cursos de posgrado

Cursos para la carrera de Especialistas en ORL

Cursos brindados para la Carrera de Especialistas en ORL, Facultad de Medicina, UBA.
Docente: Dra. Miguelina Guirao

- Codificación neurosensorial en el sistema gustativo (2 de Mayo 2013). En Sede de la FASO (Federación Argentina de Sociedades de ORL).
- Funciones neurosensoriales del sistema gustativo (6 de Agosto 2013). Asociación Médica Argentina.

- Mecanismos neurosensoriales del sistema olfatorio (12 de Noviembre 2013). Sede de la FASO (Federación Argentina de Sociedades de ORL).

Reconocimiento automático de habla e identificación de locutores

Docentes: Dr. Diego Evin, Dr. Jorge Gurlekian, Dr. Humberto Torres, Ing. Miguel Martínez-Soler, Ing. Christian Cossio Mercado. Lugar: Escuela de Ciencias Informáticas 2013, DC, FCEyN, UBA
Fecha: 22 a 27 de julio de 2013
Duración: 15 horas (teórico-práctico)
Descripción: El objetivo de este curso fue el de brindar los conocimientos básicos para el desarrollo de sistemas de reconocimiento automático del habla (RAH) y de locutores (RAL).

5.2. Seminarios en el laboratorio

- Prof. Dr. Hansjörg Mixdorff (University of Applied Sciences, Berlin). "Alignment of intonational events". 20 de marzo de 2013.
- Dr. Hartmut R. Pfitzinger (Institute of Phonetics, CAU, Kiel). "Software for speech modification". 20 de marzo de 2013.

6. Intercambio Científico

6.1. Visita de Investigadores al LIS

Dr. Hansjörg Mixdorff y Dr. Hartmut Pfizinger, 2–28 de Marzo 2013

Dentro del Programa de investigación conjunta entre MINCyT y el BMBF de Alemania, se trasladaron al LIS el Prof. Dr.-Ing. habil. Hansjörg Mixdorff de la Beuth-Hochschule für Technik Berlin (University of Applied Sciences) y el Dr. Hartmut R. Pfitzinger, del Inst. of Phonetics, CAU, Kiel. Los investigadores visitantes trabajaron durante un mes en el Proyecto y dictaron Seminarios.

6.2. Estadías en el Exterior de Investigadores del LIS

Dr. Jorge Gurlekian, Berlín (Alemania), 2–17 de Septiembre de 2013

Se realizaron reuniones de trabajo, tareas de coordinación y trabajos relacionados con el Proyecto de Cooperación MINCYT-BMBF (AL/12/06)

Dr. Diego Evin, Berlín (Alemania), 29 de agosto al 13 de Septiembre de 2013

Presentó el trabajo conjunto en la Conferencia internacional AVSP 2013 en Francia.

En Berlín realizó tareas dentro del Proyecto de Cooperación MINCYT-BMBF (AL/12/06).

7. Tesis

7.1. Doctorales en curso

Evaluación Automática de Calidad del Habla Artificial

Tesista: Christian Cossio Mercado
Director: Dr. José Castaño (FCEyN-UBA)
Consejero de Estudios: Dr. Agustín Gravano (FCEyN-UBA)
Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales
Resumen:

Un sistema de conversión de texto a habla (o sistema TTS, por el acrónimo del inglés *Text-To-Speech*) se encarga de transformar un texto de entrada en una secuencia de sonidos equivalente a la que produciría una persona al leerlo en voz alta. Es decir, dadas las palabras escritas, debe encargarse de pronunciarlas de manera: *inteligible*, que se comprendan todas las palabras dichas; *natural*, que sea similar a como lo diría un ser humano; y *expresiva*, que acompañe, refuerce y sea consistente con el mensaje comunicado.

En la mayoría de los sistemas actuales se alcanza una inteligibilidad próxima a la del habla humana. Sin embargo, un problema hasta ahora no resuelto satisfactoriamente es el de la naturalidad y expresividad del habla sintetizada. Para que un usuario preste atención al habla artificial durante un tiempo prolongado, es necesario que esta sea lo más natural posible dado que, de lo contrario, perderá la concentración, se sentirá cansado y perderá gran parte de la información, aun cuando la voz sea inteligible. Por otra parte, especialmente en sistemas de diálogo, la expresividad será necesaria por cuanto se desea transmitir un mensaje con una intención específica, demostrar un estado de ánimo determinado o indicar un tipo de relación con el oyente (e.g., de amistad o distante).

La falta de naturalidad y expresividad en los sistemas de diálogo automático ha restringido su empleo a ámbitos específicos como, por ejemplo, lectura de mensajes escritos, acceso por voz a información de servicios, y solicitud de reservas, aunque, por caso, no es aceptable para lectura de libros o diálogos extensos.

A continuación se listan las principales hipótesis de trabajo de la tesis:

- Es posible mejorar los métodos de evaluación automática de habla incorporando criterios utilizados por los seres humanos en el procesamiento del habla natural.
- Al obtener atributos propios del habla natural que están relacionados con percepciones de calidad buenas y malas, estos se pueden usar como referencia para evaluar habla sintetizada y predecir su calidad.
- Es posible mejorar la predicción de la evaluación de sistemas TTS al integrar en un único modelo atributos objetivos del habla, así como otros originados en el análisis de textos y acústicos, y el estudio de concordancia entre los resultados de estos dos análisis.
- El posible desarrollar métricas objetivas asociadas a la agradabilidad de un sistema de TTS y utilizarlas para mejorar la predicción de la calidad de un sistema de conversión de texto a habla.

Esta tesis en desarrollo busca realizar un **modelo integral para evaluación automática de la calidad del habla artificial**, aprovechando las ventajas de los modelos multidimensionales, en el cual se incluirán características originadas en el estudio de la percepción

de habla natural y del habla sintetizada , técnicas de reconocimiento automático de habla y descripciones de las Neurociencias.

El trabajo tiene como aporte principal la integración en un mismo modelo de métricas originadas en la percepción humana con otras basadas en atributos físicos acústicos del habla. Adicionalmente, se tendrá especial consideración para realizar una evaluación que permita su utilización en un esquema de Aprendizaje Automático.

Este trabajo está basado en evaluaciones perceptuales sobre habla artificial, obtenida a través de sistemas TTS conocidos (con voces masculinas y femeninas, comerciales y de uso académico, y que utilizan diferentes técnicas de síntesis), las que sirven de referencia para determinar un conjunto mínimo de atributos medidos en forma objetiva. Estos atributos permitirán predecir la evaluación subjetiva que recibirá un segmento de habla.

Se realizará un desarrollo iterativo e incremental del modelo, primero con la inclusión de atributos originados en evaluaciones objetivas, para luego incorporar atributos acústicos del habla, del texto sintetizado y de la concordancia entre ambos grupos de atributos.

Tanto los atributos perceptuales del habla, obtenidos experimentalmente, como cada conjunto de atributos medidos en forma automática en la señal serán procesados con algoritmos de selección de características y análisis multidimensional (e.g., SVD y MDS). De esta forma, sólo quedarán las características relevantes para explicar la variación en las evaluaciones registradas.

Una vez que se tenga un conjunto de atributos que permitan predecir la evaluación perceptual, se obtendrá un conjunto independiente de evaluaciones (i.e., con nuevos sujetos) para compararlas con las predichas utilizando diferentes clasificadores (e.g., RRNN y SVM) con las características seleccionadas. Adicionalmente, y en forma complementaria a la validación, se realizará estudios sobre la percepción del habla y su correlación con mediciones electrofisiológicas (ERP), de forma de determinar la sensibilidad a la variación de los atributos principales elegidos.

Reconocimiento forense de hablantes mediante el uso de información de alto nivel y metadatos

Tesista: Miguel Martínez Soler
Directores: Dr. Jorge A. Gurlekian (CONICET) y Dr. Agustín Gravano (FCEyN-UBA)
Consejero de Estudios: Dr. Diego Garbervetsky (FCEyN-UBA)
Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales
Resumen:

En la tarea de reconocimiento forense de hablantes se intenta determinar si un individuo conocido ha producido la voz registrada en una grabación, realizando comparaciones a diferentes niveles (acústico, lingüístico, prosódico, etc.). Las soluciones propuestas para resolver este problema incluyen la evaluación perceptual realizada por un panel de oyentes experimentados, el tratamiento y modelado de señales por sistemas totalmente automáticos y el uso de sistemas semi-automáticos, normalmente basados en la extracción de rasgos acústicos bajo la dirección del usuario.

En la última década el uso de enfoques bayesianos cobró popularidad en el área. Estos enfoques permiten calcular una probabilidad a posteriori de una hipótesis (dada la evidencia) tomando en cuenta su probabilidad a priori y una evaluación de la evidencia (dada la hipótesis). La principal ventaja que tiene esta forma de encarar el problema es que los sistemas que

utilizan este enfoque emiten resultados fácilmente interpretables por humanos.

Las fuentes de información disponibles en el habla se pueden organizar en diferentes niveles de abstracción que van desde el nivel más bajo, que contiene la información acústica, hasta el nivel más alto, que describe la forma en que se usa el lenguaje. En medio de estos extremos encontramos también la información fonológica y prosódica, entre otras. Las diferencias dialectales pueden encontrarse en cualquiera de estos niveles, siendo en nuestro país muy notorias las diferencias que se encuentran a nivel fonológico, prosódico y léxico.

Además de la información contenida en la señal de habla en sí, se ha demostrado en trabajos anteriores que, utilizando un enfoque bayesiano, se puede mejorar la con[U+FB01]anza del reconocimiento de hablantes mediante el uso de metadatos, como por ej. las duraciones de los segmentos de audio en estudio, o el tipo de canal (directo, telefónico convencional, telefónico celular, etc.).

En un trabajo de 2008, Ferrer et. al. [27] obtuvieron una mejora notable en una tarea de reconocimiento automático de hablantes mediante una estimación del grado de certeza de que el hablante sea nativo. Este resultado sugiere que el uso de información dialectal podría aportar mejoras a la tarea de reconocimiento de hablantes. Dado que se encuentran a disposición dos corpus de datos de habla segmentados por regiones de la República Argentina, se propone explotar el uso de información de alto nivel en la señal de habla para la tarea de reconocimiento forense de hablantes, tomando en cuenta las diferencias que se puedan encontrar en las diferentes regiones de nuestro país. Se estudiará también el uso de metadatos y su impacto en la confianza de las estimaciones. La evaluación del trabajo será realizada mediante la comparación con sistemas de reconocimiento de hablantes automáticos y semiautomáticos implementados mediante técnicas del estado del arte.

Reconocimiento automático de hablantes empleando información de largo plazo

Tesista: Pedro Univaso
Director: Dr. Jorge A. Gurlekian
Universidad de Buenos Aires, Facultad de Ingeniería.

Diagnostico diferencial de pacientes con movimientos anormales laringeos, complementacion entre el diagnostico neurológico y los resultados que brinda el abordaje otorrino-fonoaudiologico

Tesista: Liliana Sigal
Universidad de Buenos Aires, Facultad de Medicina
Director : Dr. Jorge A. Gurlekian.
Consejero : Profesor Dr Federico Micheli

El presente trabajo se desarrollará en el Laboratorio de Investigaciones Sensoriales, Conicet, la División Otorrinolaringología, el Departamento de Movimientos Anormales y el Departamento de Neurofisiología del Hospital de Clínicas José de San Martín. Tiene por objeto el estudio de la distonía laríngea, alteración que afecta el movimiento de las cuerdas vocales. Los síntomas son generalmente graduables desde una inestabilidad moderada a cortes incontrolables en la voz y un creciente esfuerzo que repercute en la inteligibilidad del habla.

8. Actividades de Divulgación

- Torres, Humberto. Conversión de texto en habla, Semana de Biomédica, Dto. de Biomédica, FI-UBA. 28 de Nov. 2013. Charla introductoria a los sistemas de conversión de texto en habla. Presentación y demostración de Aromo, el primer sistema de conversión de texto en habla de la Argentina.

- Cossio-Mercado, Christian. Presentación del tema de tesis, Programa "Yo Quiero", Tecnópolis TV. Septiembre 2013 (grabación). Entrevista para corto de divulgación en el cual se presentó el tema de investigación doctoral.

- Cossio-Mercado, Christian. Descubriendo el lado oculto de nuestra voz, Semana de la Computación, FCEyN, UBA. Agosto 2013. Dictado de charla sobre análisis automático de habla y su utilización en el desarrollo de sistemas de diálogo orientados a predecir el estado físico y emocional de los usuarios, y aprovechar esa información para mejorar la interacción con ellos.

- Cossio-Mercado, Christian. ¿Me escuchás bien? ¡Fuerte y claro! Cómo lograr que las computadoras hablen mejor, Ciclo CdB, DC, FCEyN, UBA. Junio 2013. Dictado de charla sobre percepción de habla y evaluación automática de la calidad del habla, y sus implicancias para el desarrollo de sistemas de diálogo.

- Entrevista al Dr. Jorge Gurlekian, incluída en el artículo "Reconocimiento de voz. Buen día compu, ¿qué hay de nuevo?". Revista Exactamente, Año 20, No. 52, pp. 10–13, Facultad de Ciencias Exactas y Naturales, UBA. ISSN 1514-920X. Disponible online en http://noticias.exactas.uba.ar/reconocimiento-voz-habla

- Entrevista al Dr. Jorge Gurlekian, incluída en el artículo "Ciencia de ficción. El Himno Nacional archivado en una bacteria, computadoras que hablan y robots rescatistas: los últimos avances en la Argentina que parecen inspirados del cine y la TV futurista". Revista 7 Días, No XXX, pp. 34–40, 2013.

9. Publicaciones

9.1. Revistas

- TORRES, H.M., GURLEKIAN, J.A., MIXDORFF, H., and PFITZINGER, H.: "Linguistically motivated parameter estimation methods for a superpositional intonation model". En EURASIP: Journal on Audio, Speech, and Music Processing, 2013.

- GUIRAO, M., GRECO DRIANÓ, E., EVIN, D.A. and Calviño, A.: "Psychophysical assessments of sourness in citric acid- ethanol mixtures". Perception and Motor Skills, n. 117, pp. 868–880, 2013.

9.2. Informes Técnicos

- GURLEKIAN, J.A. y TORRES, H.M.: "Guía para la segmentación y transcripción fonética para las tecnologías de habla". Ver contenido completo en los apéndices.

- TORRES, H.M.: “Definición de las funciones de costo para la selección de unidades en AROMO”, 2013.
- TORRES, H.M.: “Segmentación de audio de radio y televisión”, 2013.
- GURLEKIAN, J.A., TORRES, H.M. y VACCARI, M.E.: “Internal vs. external standards for the evaluation of F0 perturbation”, 2013.
- TORRES, H.M.: “Medición preliminar de la velocidad de conversión del sistema TTS AROMO”, 2013.

Apéndice

A. Resúmenes de Trabajos

A.1. Psychophysical Assessments of Sourness in Citric Acid-Ethanol Mixtures. *Guirao, M. et al.*

Miguelina Guirao, Ezequiel J. Greco Driano, Diego Evin y Amalia Calviño

Abstract

The effect of ethanol in modulating the intensity and duration of the perceived sourness induced by citric acid was studied. Magnitude Estimation-Converging Limits method was applied to rate the sourness of seven solutions(3–70 mM) of citric acid in aqueous solution presented alone and mixed with 8 % V/V or 15 % V/V ethanol. Dynamic sourness ratings of 5, 15, and 45 mM citric acid alone and mixed with the same two ethanol levels were assessed by the Time Intensity Method (TI). Results were consistent with both methods. Sourness changed with citric acid concentration and ethanol levels. From TI measurements, a similarinteractive pattern was obtained for parameters as duration, area under the curve, peak and average intensity.

B. Informes Técnicos

A continuación se incluye el texto completo de los informes técnicos realizados para el año 2013.

B.1. Guía para la segmentación y transcripción fonética para las tecnologías de habla. *Gurlekian J.A. y Torres H.M.*

Informe Técnico: Guía para la segmentación y transcripción fonética para las tecnologías de habla

Autores: Dr. Jorge Gurlekian; Dr. Humberto Torres

Lugar: Laboratorio de Investigaciones Sensoriales. INIGEM CONICET-UBA. Hospital de Clínicas Av. Córdoba 2351. 9 Piso Sala 2.

Fecha: 20 de Febrero de 2013

Resumen

En este trabajo se define una guía para la segmentación fonética y su transcripción mediante el alfabeto SAMPA (Speech Assessment Methods: Phonetic Alphabet). La transcripción fonética con SAMPA es de uso creciente en las tecnologías de habla en las tareas de preparación de datos acústicos para ser utilizados en el entrenamiento de sistemas de reconocimiento de habla/hablante y en la preparación de unidades de concatenación en sistema de conversión texto a habla. Se presentan ejemplos de cada realización fonética, los criterios de segmentación y su el etiquetado. Las ondas están disponibles para su verificación.

Palabras claves: habla, transcripción, segmentación acústica, tecnologías del habla

Abstract

A normalized guide is proposed here, for both phonetic segmentation and labeling using SAMPA, (Speech Assessment Methods: Phonetic Alphabet) based on acoustic productions of each speaker. SAMPA labeling is a standardized technique for speech pre-processing both for automatic speech recognition systems and text to speech concatenation systems. Speech samples with segmentation and, labeling are shown. The corresponding waveform files are also available for verification.

Keywords: speech, labeling, acoustic segmentation, speech technologies.

Introducción

La segmentación fonética consiste en indicar el inicio y fin de eventos acústico-fonéticos con marcas en el eje de tiempos. Esta tarea es de interés en las disciplinas que deben asociar segmentos acústicos de distinta longitud con etiquetas que representen unidades lingüísticas y no lingüísticas de duración segmental y supra-segmental de una manera confiable. Una de las aplicaciones tecnológicas más solicitadas en la actualidad es en el área de la comunicación verbal. Las tareas de análisis, síntesis y reconocimiento automático requieren de bases de datos

etiquetadas por fonoaudiólogos y lingüistas con entrenamiento musical para realizar la tarea de transcripción.

El proceso de transcripción fonética automática se inicia con la segmentación y etiquetado manual de una porción de los datos de entrenamiento. Luego se procede con la transcripción ortográfica de las palabras y pausas llenas del resto de los datos de entrenamiento. Con esta información se realiza la alineación forzada de los fonos empleando herramientas como el HTK que emplean modelos ocultos de Markov (HMM). Finalmente la alineación forzada puede corregirse para aumentar el nivel de desempeño del transcriptor. A continuación se comentan algunos ejemplos de síntesis, reconocimiento automático de habla y hablantes.

La conversión de texto a sonidos de habla de alta calidad mediante la concatenación de unidades presupone la corrección manual de unidades de habla -típicamente difonos- que son segmentos de habla que van desde un instante de mayor regularidad o estabilidad en las frecuencias de los formantes en un fono hasta otro instante semejante en el próximo fono. Si el habla es rápida, este método puede extenderse a varios fonos conformando unidades de longitud variable llamadas polifonos. Aun cuando la tarea de segmentación se inicia en forma automática, la mayor precisión se alcanza con la corrección manual por parte del experto en fonética. En estos sistemas, la alta naturalidad se obtiene además con la descripción fonética de la entonación, el ritmo y el acento.

En los sistemas de conversión de sonido a texto, las técnicas de reconocimiento automático utilizan algoritmos de aprendizaje o entrenamiento que requieren una segmentación inicial de unidades. El conocimiento a priori de la información lingüística se refleja en los diccionarios de pronunciaciones, en los modelos de lenguaje y en las bases etiquetadas en unidades fonéticas.

Finalmente los sistemas híbridos de identificación de hablantes para aplicaciones forenses han mejorado su desempeño al combinar la segmentación fonética y grafémica precisa con los métodos probabilísticos que utilizan representaciones espectrales de las unidades para obtener medidas de verosimilitud en forma automática.

Como resultado del entrenamiento realizado a profesionales de fonoaudiología surge este trabajo que compendia los apuntes y recomendaciones ofrecidas recientemente.

Los sistemas de etiquetado automático requieren de reglas de transcripción grafémico-fonética que se aplican de acuerdo al contexto en que aparece cada fonema. La tarea del experto consiste en corregir las marcas temporales de segmentación y decidir si la variante fonética que propone la regla de la convención es la que se observa. Por lo tanto estas reglas son el marco de referencia inicial para el etiquetador automático. La tarea del etiquetador manual es privilegiar lo observado en la emisión acústica. Las modificaciones a las reglas se reflejan luego mediante una estadística que permite adaptar las reglas para ese locutor. Con respecto a la marcación temporal existe un número de convenciones que serán presentadas para cada grupo de sonidos. Los casos que presentan ambigüedad deben consensuarse por los expertos que determinan la marca final.

La transcripción fonética

Recientemente se ha definido el alfabeto fonético SAMPA (Speech Assesment Methods: Phonetic Alphabet) [1] para la Argentina [2] fundamentado en las realizaciones fonéticas del español hablado en esta vasta región dialectal. Este alfabeto se ha definido para varias lenguas y es de aplicación en las tecnologías de habla.

Las ventajas principales de este alfabeto son: 1. Permitir el diseño de córporas con cobertura fonética completa, 2. Permitir el uso de un conjunto de caracteres ASCII estándar que se mapea con todos los símbolos del IPA. El mapeo asigna las diferentes variantes fonéticas regionales a un mismo símbolo SAMPA. De este modo los córpora grabados mediante textos SAMPA pueden garantizar la riqueza fonética de cada región.

En el trabajo mencionado [2] se han presentado los 22 fonemas del español de Argentina y Latinoamérica (17 consonantes y 5 vocales). Los dos fonemas que no se utilizan en Latinoamérica y nos diferencian del español peninsular son /z/ y /v/. Se han indicado además los ocho alófonos más frecuentes utilizados en la Argentina: [h, B, D, G, N, C, w, j) identificándose un total de 30 fonos. Ver síntesis en Tabla I y denominaciones en Tabla II.

Tabla 1: Relaciones fonémicas a la fonética SAMPA para el español de la Argentina.

Vocales
/a, o, e/ se transcriben a SAMPA con sus símbolos fonéticos [a, o, e]
/u, i/ se transcriben como [u, i,] siempre que no formen diptongo dentro o fuera de la palabra en cuyo caso se transcriben como [w] y [j]. Grafemas a, e, o, u, i, y.
Consonantes
Ruidos
/s/ se transcribe como [s] y cuando precede a una consonante puede transcribirse como [h]. Si el fonema es omitido se indica como [0]. Grafemas: s/c/z.
/x/ se transcribe como [x] antes de [a, o, u, w] y como [C] antes de [e, i/j]. Grafemas j, g.
/f / se transcribe como [f]. Grafema: f
/Z, $\widehat{dZ}$, ʃ, ɟ / como [Z]. Grafemas: ll, y
Silencio + Ruido
/p,t,k/ se transcriben como [p, t, k]. Grafemas p, t, k/c/q.
/ $\widehat{dZ}$, / se transcribe como [H]. Grafema: ch.
Nasales
/n/ se transcribe como [n], si precede a un sonido velar o palatal, puede transcribirse como [N].

o labial o bilabial, puede transcribirse como [m]. Grafema n

/m, ɲ/ se transcriben como [m, J]. Grafemas m, ñ.

Interrumpidos

/P, r/ se transcriben como [r, R]. Grafemas: r y r/rr.

Continuos

/b, d, g/ se transcriben como [B,D,G] salvo que sigan a una nasal o sean inicio de frase después de pausa, en cuyo caso se transcriben [b, d, g]. Grafemas b/v, d, g.

/l/ se transcribe igual.

Denominación de los sonidos de habla

Las realizaciones acústico-fonéticas utilizan términos extraídos de los distintos campos de la fonética Tabla I.

Tabla 2: Denominaciones derivadas de la fonética articulatoria, acústica y perceptual de los sonidos de habla.
[1]La clasificación perceptual es una adaptación de Guirao [3].

SAMPA	Articulatoria	Acústica	Perceptual[1]
p, t ,k (3)	Oclusivos sordos	Silencio + ruido corto, medio y largo	Pulsos fuertes no continuos compactos agudos
b, d, g (3)	Oclusivos sonoros	Un primer formante de baja frecuencia	Pulsos suaves no continuos difusos graves
B, D, G (3)	Aproximantes (no fricativos) [4]	Formantes transicionales de baja amplitud	(no evaluados) Transicionales periódicos
s, f, x h, C (5)	Fricativos sordos	Bandas de ruido (h,f) en todo el espectro	Ruidos en bandas Continuos
H (1)	Africada sorda	Silencio + ruido	Pulso de Ruido Fuerte
r (1)	Liquida vibrante simple	Una interrupción entre dos vocales	Periódico+ruido Cortos Suave No continuo
R (1)	Liquida vibrante múltiple	Múltiples interrupciones	Periódico+ruido largos continuos
l	Liquida Lateral	Formantes de amplitud	Periódico Continuos

SAMPA	Articulatoria	Acústica	Perceptual[1]
(1)	Pre-palatal lateral	media	de baja amplitud
Z **(1)**	Fricativa sonora Palatal y pre-palatal Africada sonora	Periodicidad + banda de ruido de alta frecuencia.	Periódico+ruido Largo Agudo
i, e, a, o, u **(5)**	Vocales	Periódicos de amplitud alta con formantes de alta int.	Periódicos Continuos de alta amplitud
w, j **(2)**	Semivocales	Formantes transicionales de amplitud media	(no evaluados) Idem vocales
m, n, N, J **(4)**	Nasales	Formantes fijos de baja frecuencia y amplitud	Periódicos Continuos de baja amplitud

Representación acústica de los sonidos de habla

La manifestación acústica clásica es la transducción eléctrica de la variación de presión en función del tiempo –forma de onda- y la del espectrograma. Las ondas de vocales y otros sonidos con vibración glótica se presentan como ciclos de presión cuasi-periódicos en el tiempo que se distinguen claramente de los sonidos ruidosos y de los sonidos que contienen silencio más ruido. En la Figura 1 se distinguen claramente los sonidos ruidosos, periódicos y oclusivos. Sin embargo cuando se emiten dos vocales seguidas o en combinación con consonantes con vibración cuasi-periódica, la diferenciación se logra con el espectrograma. La coarticulación se manifiesta mediante las transiciones de formantes. Para ver esquemas de las transiciones VCV del español ver [4].

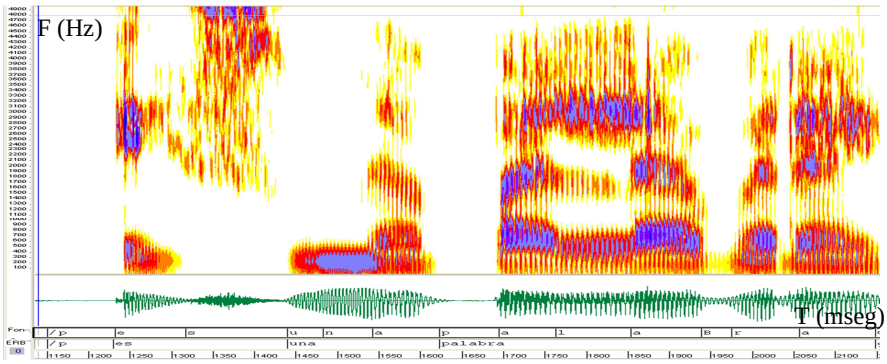


Figura 1: En la forma de onda se distinguen los sonidos periódicos [e,u,a,l] de los ruidos [s] y silencios de un sonido pulsante [p]. En el espectrograma se distinguen la [u] y la [n] de una [a]. También se distingue la [l] de las dos [a] adyacentes.

Sonidos Vocálicos

Los fonemas vocálicos /i,e,a,o,u/ se corresponden con los fonos [i,e,a,o,u] Figura 2 y los alófonos [w, j] Figura 3. A medida que el habla es más descuidada o tiene mayor velocidad los valores prototípicos de los formantes de las vocales aisladas se modifican al punto de desaparecer quedando como indicación de la vocal una representación acústica aproximada o gruesa que permite su identificación.

Como rasgos de rescate vocálico se observa para la vocal /a/ un 1er formante más alto que el resto de los formantes de todas las vocales presentes, con un 2do formante lo más próximo a este 1er formante. Para la vocal /i/ los rasgos distintivos gruesos son un 1er formante más bajo que el resto de los formantes y un 2do formante lo mas separado posible de este 1ro. La vocal /u/ tendrá el 1er formante más bajo (como la /i/) y un 2do formante lo más cercano posible. Con este nivel de descripción podrán diferenciarse las secuencias continuas de vocales. Las vocales /e/ y /o/ se diferenciaran de la /i/ y la /u/ al considerar su configuración de formantes como más cercano al patrón de una /a/.

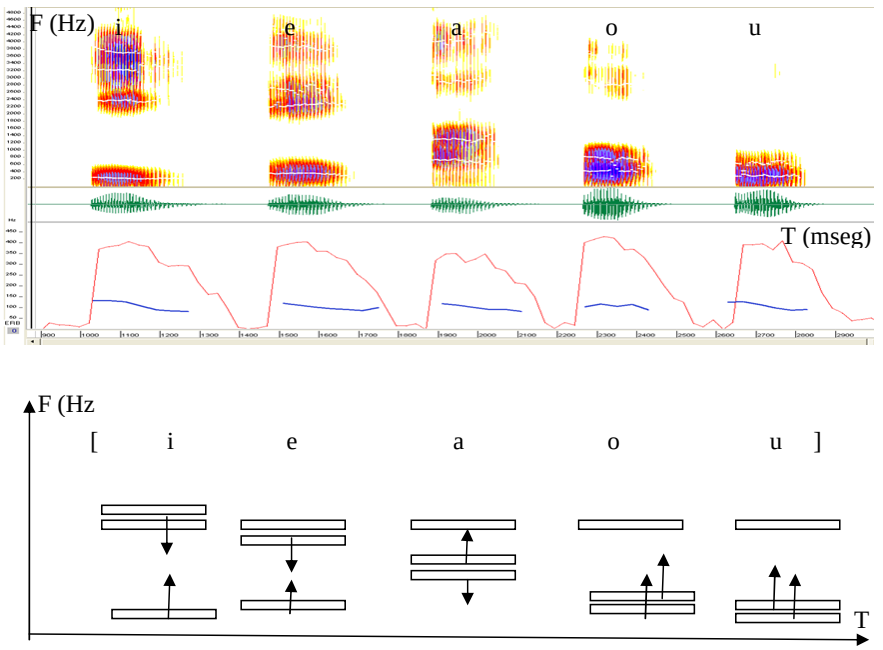


Figura 2: Representación de las vocales aisladas. Espectrograma, forma de onda y contornos de Energía Total y Frecuencia Fundamental (arriba) emitidas por un locutor masculino y de un esquema idealizado de las mismas vocales. (abajo).

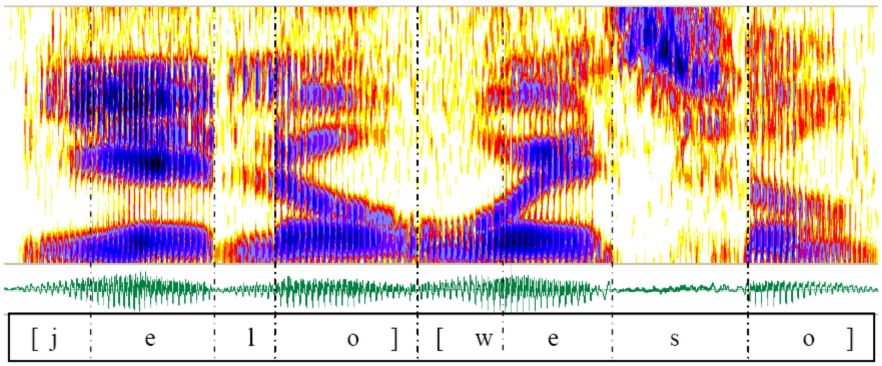


Figura 3: Representación espectrográfica de [j,w] en hielo, hueso.

Segmentación

En la segmentación en secuencias VV, el criterio es dividir la transición vocálica en dos partes iguales y asignar cada parte a un fono. Ver figura 4. En la segmentación de secuencias CV o VC el criterio es asignar toda la transición vocálica a la vocal. Si la vibración glótica de baja amplitud de la vocal continúa breve ó totalmente en el segmento que le corresponde al silencio o en el segmento correspondiente de ruido por falta de oclusión completa de la vocal previa, se asignará igualmente ese segmento a la consonante.

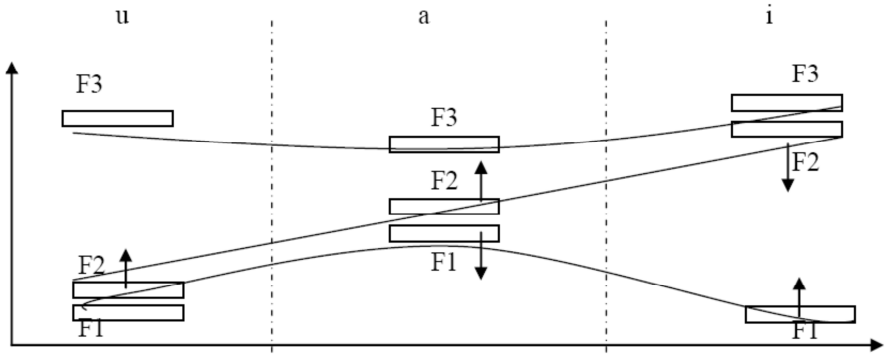


Figura 4: Esquema de tres vocales donde se indica la trayectoria de los tres primeros formantes. Los formantes 1 y 2 de la vocal [u] tienden a elevarse, en la [a] a separarse y en la [i] a acercarse. Las transiciones se dividen por la mitad y se asignan a cada vocal

Sonidos [p, t, k, H]

Los fonemas /p,t,k, $\widehat{t\int}$/ se corresponden con los sonidos con silencio y ruido tales como los fonos oclusivos sordos [p,t,k] y la africada [H]. Estos presentan un segmento de silencio típicamente de 30 a 60 mseg, resultado de la oclusión al paso del aire para finalizar con un ruido de explosión de 1-5 mseg para [p] de 5 a 15 mseg para [t], de 15 a 30 mseg para [k] y de 30 a 60 mseg para [H]. La terminación del ruido de explosión es el instante que debe ser marcada como límite final de estas consonantes. Como se mencionó en el inicio el segmento de silencio puede presentar ciclos amortiguados en amplitud del sonido sonoro anterior que son visibles por no existir una oclusión completa. Cuando estos sonidos están al inicio de una oración o de una frase melódica delimitada por silencio debe marcarse el inicio dejando

aproximadamente 60 mseg de silencio previo al ruido de explosión. El sonido [H] se considera como la unión de dos sonidos: la [t] y [S].

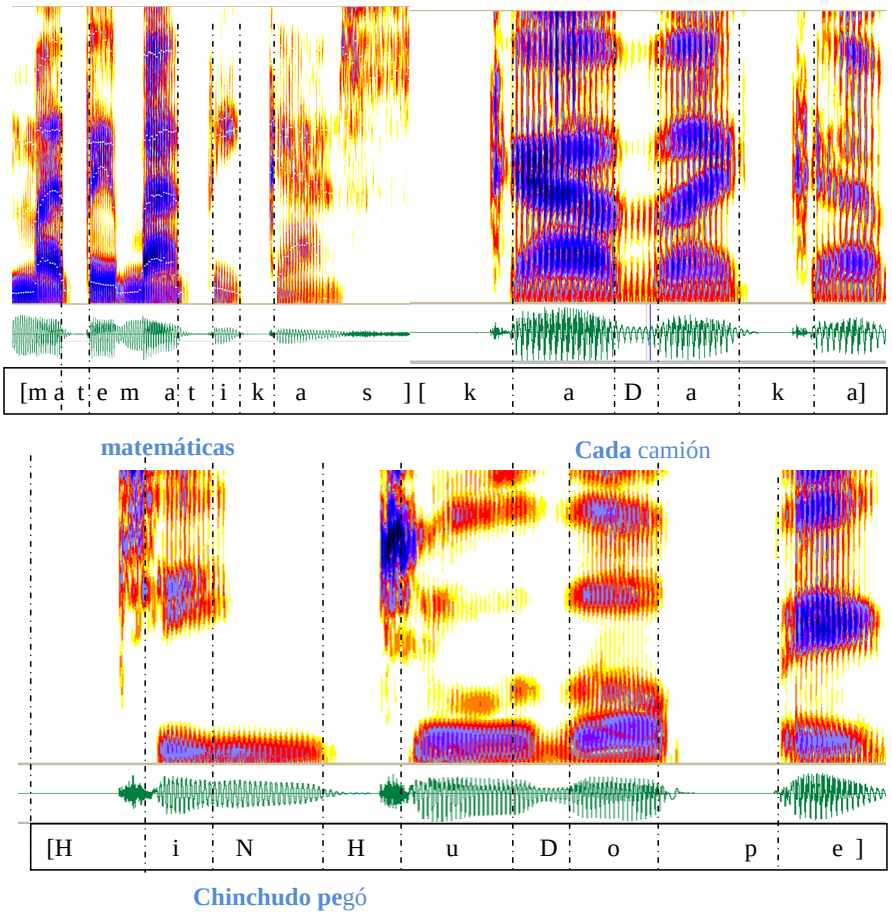


Figura 5

Son sonidos compuestos solo por ruido. Se distinguen claramente de los sonidos periódicos y de los pulsos. El sonido [s] tiene una duración típica de 100 a 200 mseg con unas concentración de alta energía en la banda de 4000 a 5000 Hz. Cuando se produce en secuencias pre-consonánticas se produce la variante aspirada [h] [5], donde la duración disminuye notablemente a 30-60 mseg y su energía se distribuye en todo el rango de frecuencias. El sonido [f] es de una duración similar a la [s] pero de baja intensidad a lo largo de todo el espectro de frecuencias. El sonido [x] y su alófono [C] presentan una concentración de alta energía en la banda de frecuencias media, entre los 2000 y 3500Hz.

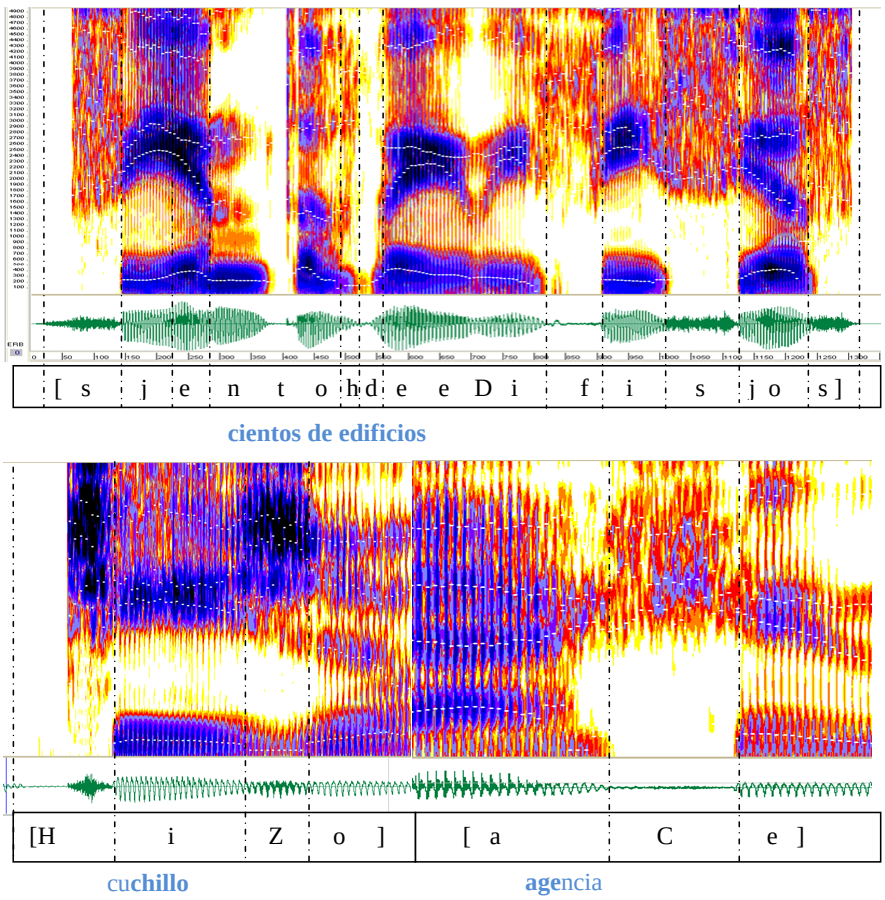
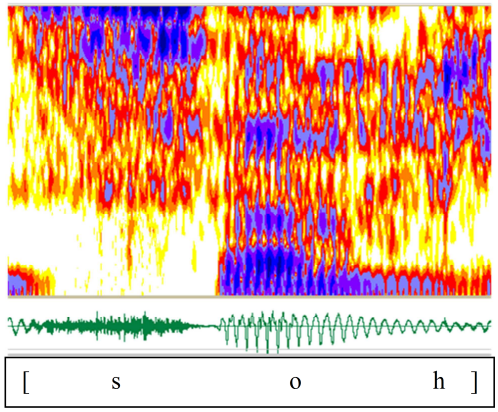


Figura 6

Alófonos [h] y [C]

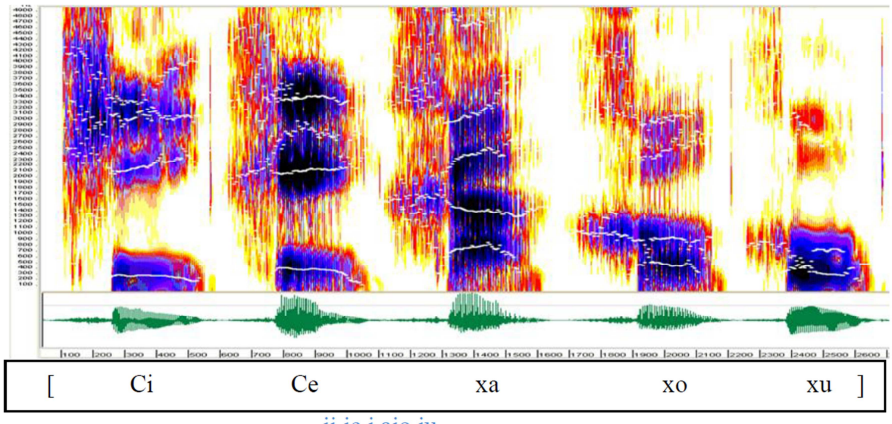
El sonido [h] presenta un espectro distribuido en frecuencias y puede aparecer superpuesto a una amortiguación periódica, en cambio [s] posee una concentración en 4500 Hz con mayor duración y sin vibración periódica.



sos

Figura 7

El fonema /x/ presenta una banda de ruido cuya posición es muy dependiente de la vocal. Esta dependencia vocálica se ha dividido las realizaciones en [C y x].



ji je j ajo ju

Figura 8

Sonido [r]

El sonido [r] puede presentarse como una breve interrupción de 10 a 20 mseg entre dos sonidos vocálicos. En secuencias de grupos consonánticos CC donde una de las C es [r] se visualiza una vocal epentética que se asignará a la consonante [r]. Esta vocal no audible aparece entre la consonante [r] y la consonante con la que forma la secuencia. Ejemplo: [tres] → [teres], [arte] →[arate]. En posición final de [r] seguida de silencio tambien aparcera una vocal epentética. Ejemplo: [ver]→[vere], [mar]→[mara].

Opcionalmente otras consonantes pueden presentar una vocal epentética como finalización o inicio. En estos casos también se asignarán a la consonante que produce la epéntesis. La vocal epentética posee una duración menor a la de las

vocales con una estructura de formantes y energía equivalentes a la vocal más cercana. De acuerdo a estos contextos hay tres realizaciones de [r]: 1. sin vocal epentética, 2. con vocal epentética a la derecha de la interrupción y 3. con vocal epentética a la izquierda de la interrupción.

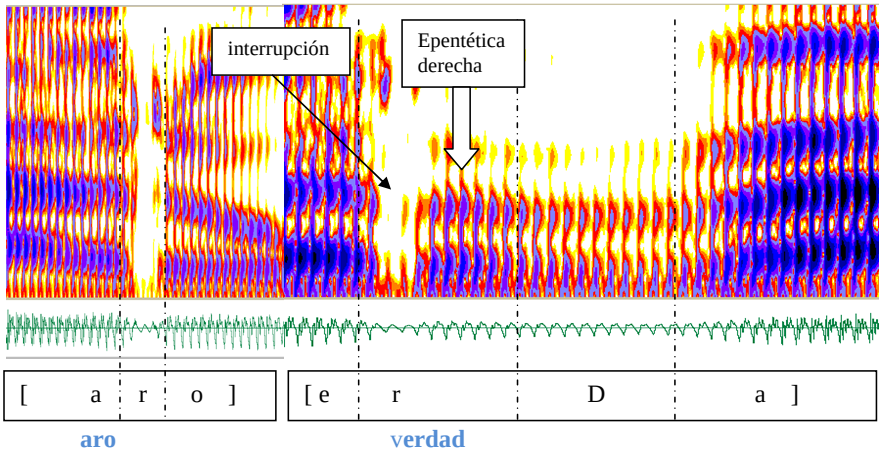


Figura 9: Realización del sonido /r/. En posición intervocálica y en grupo r¬+Consonante.

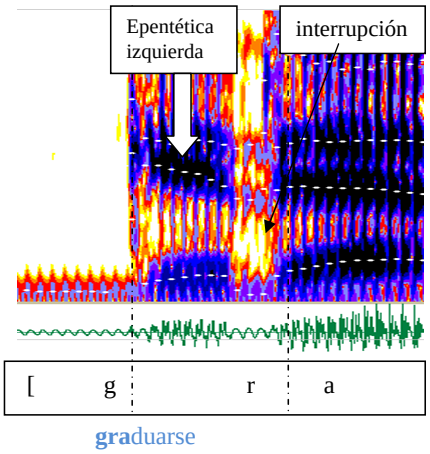
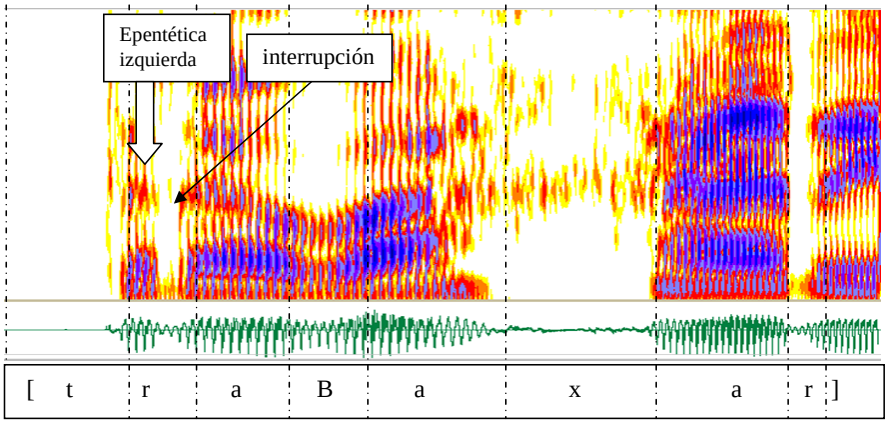


Figura 10: Realización del sonido /r/. En grupos consonante + r .

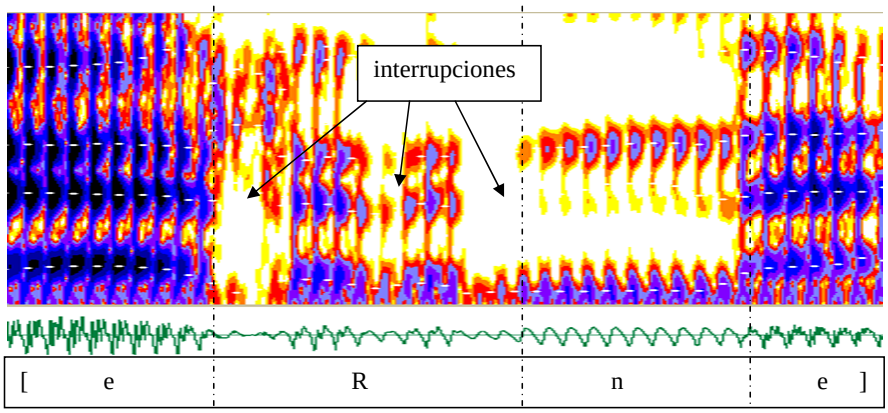


trabajar

Figura 11: Realización del sonido /r/. En grupos consonante + r y en posición final.

Sonido [R]

Este sonido presenta más de una interrupción rápida de 10 a 20 mseg en la secuencia de formantes vocálicos. Posee distintas variantes: ruidosa, aproximante y combinadas según la región dialectal. [6]

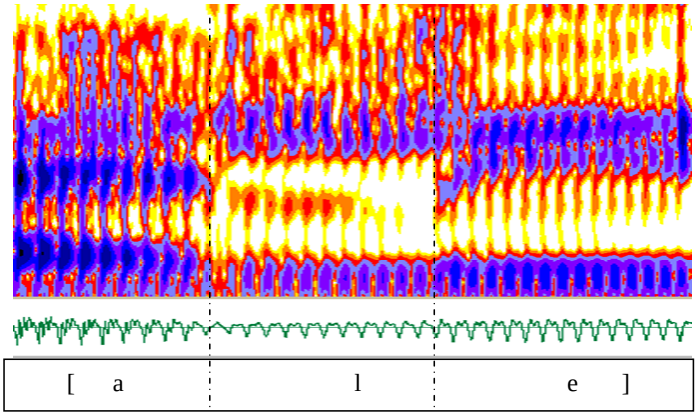


cab**ernet**

Figura 12

Sonido [l]

Es un sonido con una estructura de formantes bien definida similar a una vocal /e/.

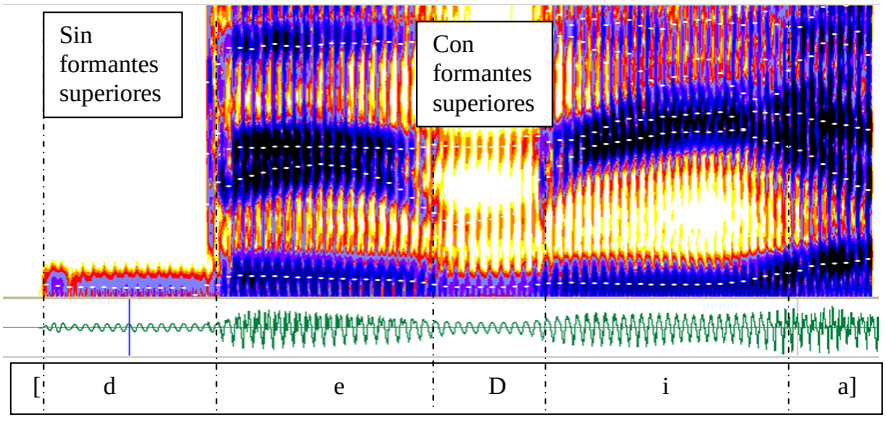


reales

Figura 13

Sonidos [b,d,g] y [B, D,G]

Los fonemas /b,d,g/ se realizan como los fonos [b,d,g] y sus alófonos [B,D,G]. Son sonidos con vibración glótica y una oclusión incompleta por lo que presentan ciclos de muy baja energía. En [b, d, g] el espectro se limita a una resonancia de muy baja frecuencia y un ruido de explosión breve muy débil. En [B,D,G] se visualiza energía en los formantes superiores -como mínimo en el 2do formante- sin ruido. Las transiciones a la vocal son lentas 30-40 mseg. La segmentación debe asignar las transiciones de los formantes de mayor amplitud a la vocal y las transiciones atenuadas la consonante.



de dia

Figura 14

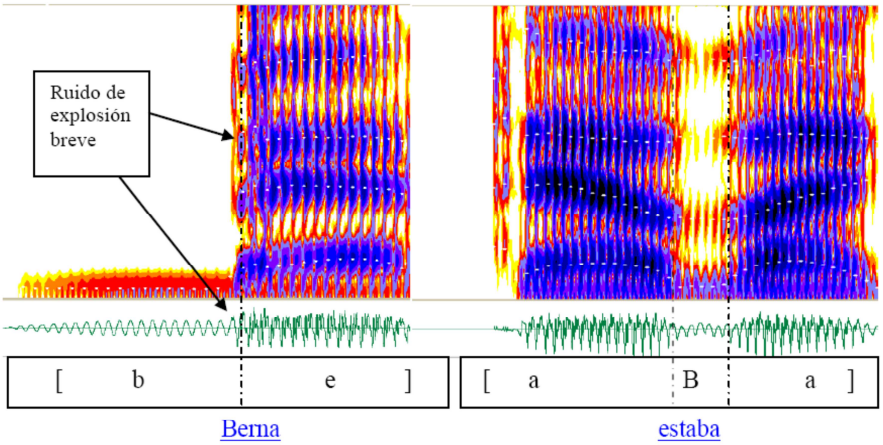


Figura 15

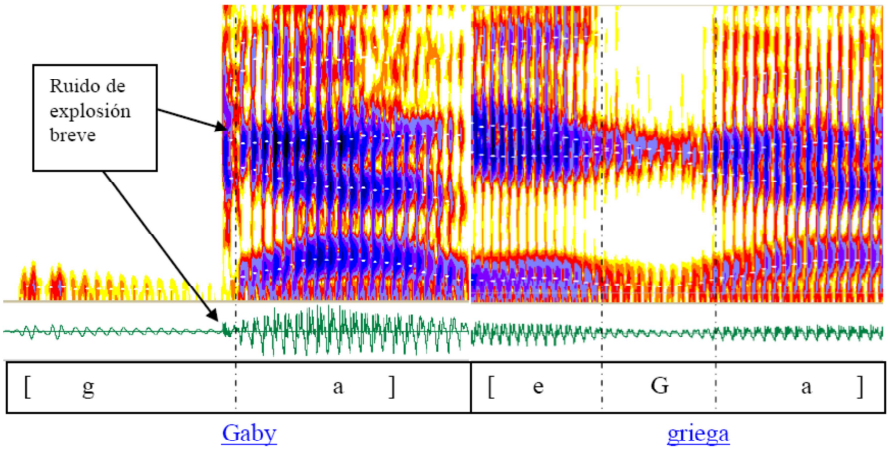


Figura 16

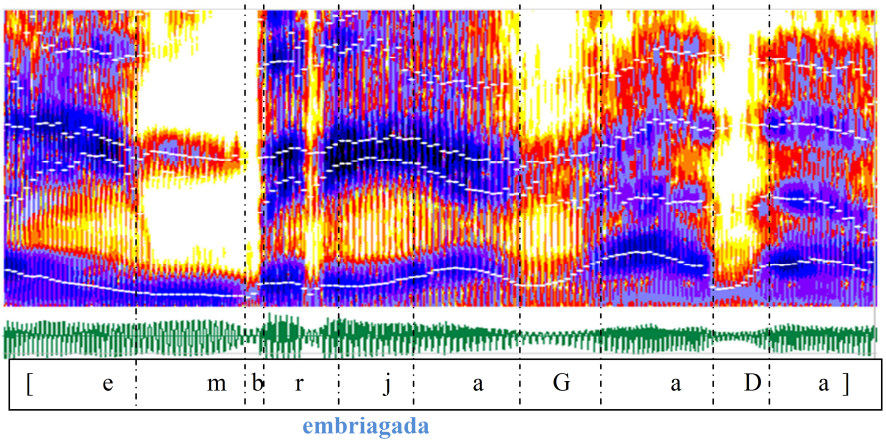


Figura 17

Sonidos [Z] y [L]

[Z] es un sonido con ruido de alta intensidad en la banda de 3000-4000 Hz puede aparecer superpuesto a ciclos de vibración glótica cuasi-periódica. El segmento con ruido se asigna completamente a la consonante. [L] es un sonido con vibración glótica y formantes sin ruido. Las transiciones de menor amplitud corresponden a la consonante.

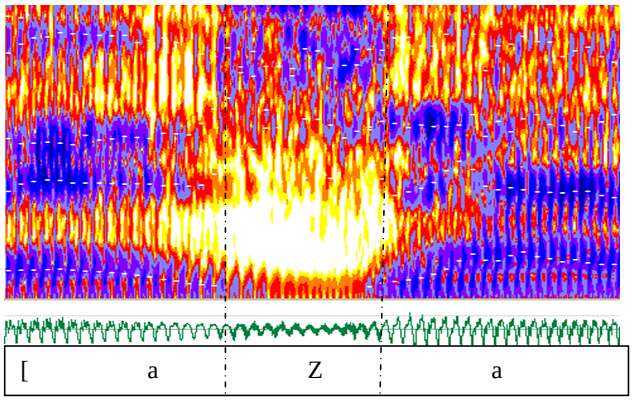


Figura 18

Sonidos [m, n, N, J]

Se caracterizan por presentar dos formantes nasales de menor energía que las vocales entre 250 y 550 Hz para FN1 850 y 1250 para FN2. Los formantes nasales poseen una rápida transición con los sonidos orales adyacentes. Este cambio abrupto y la atenuación o ausencia de continuidad de los formantes orales son un buen indicador del inicio y fin del sonido nasal. El sonido n presenta el alófono [N] que aparece antes de una consonante velar [g, x, C, k, h]. Este sonido se transforma en [m] antes de [p,b,f]. Figura.

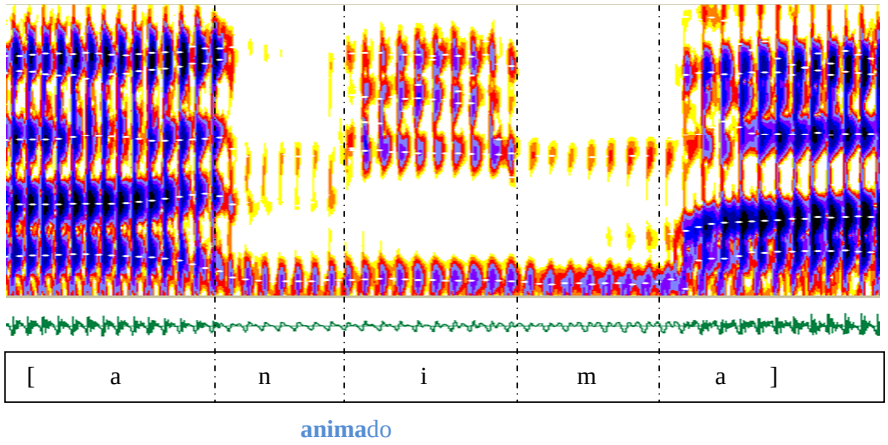


Figura 19

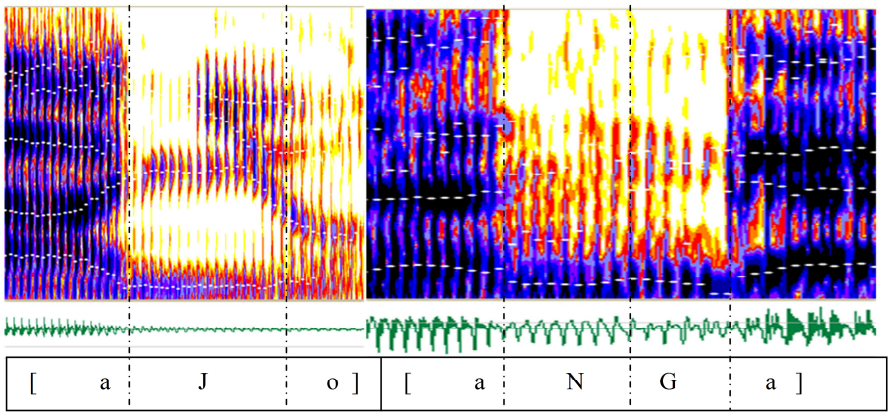


Figura 20

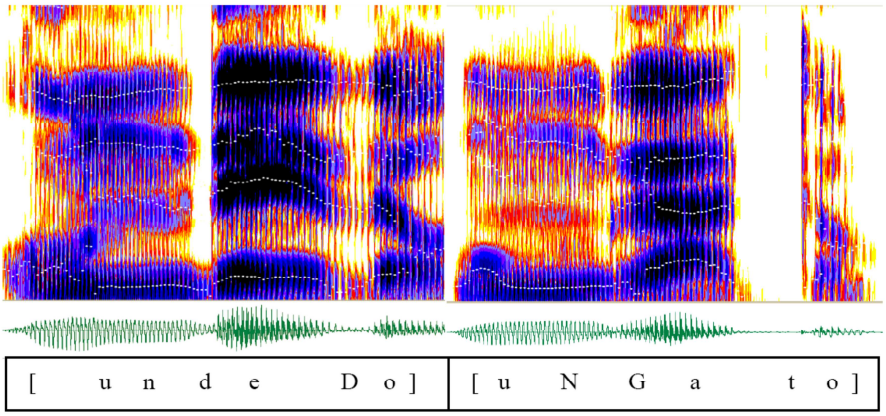


Figura 21

Segmentación Prosódica

La segmentación prosódica manual o su corrección se requiere para el diseño de los modelos de entonación y ritmo. Se indican con las marcas de pausa de frases y grado de separación entre palabras, los acentos de frases y tonos de juntura y acentos tonales.

Pausas

Las frases se delimitan por silencios o cambios tonales abruptos, en ambos casos la sensación producida es la de una pausa. La notación para estos casos es con un grado numérico 3 ó 4. Dentro de las frases se indica el grado de pausa entre palabras con 0, 1, 2 realizado por cada locutor. Ref. Valores típicos pero no obligatorios son: grado 0 para V1V1 y C1C1, grado 1 para CV, V1V2 y VC y grado 3

para C1C2 donde V es una vocal y C una consonante. Los números 1 y 2 representan distintas vocales.

Acentos de Frase

Se observan al final de la frase entonativa en el contorno de F0. Se indican como H- (alto), M- (medio) y L- (bajo).

Acentos tonales

Se perciben en la sílaba alrededor de cada acento léxico de la palabra de contenido. Si la palabra recibe foco estrecho, el máximo de F0 estará en correspondencia temporal con el segmento acústico de la vocal de la silaba acentuada lexicalmente. El acento de frase no cambiará esta asociación. Se indica con un acento mono-tonal H* o bitonales L+H, H*+L o tritonales L+H*+L.

Si la palabra recibe foco amplio, el máximo de F0 se observa en las cercanías de la sílaba con acento léxico o puede no existir. La posición de la palabra, el tipo de palabra, el acento de frase y las variaciones estilísticas del hablante influyen en esta localización. Por ejemplo: si el acento de frase es alto H-, el acento tonal tenderá a desplazarse hacia la derecha, acercándose al acento de frase en las palabras finales con acento nuclear como L*+ H en las palabras esdrújulas y graves, ó fundiéndose con él como L+H* en las palabras finales agudas. Este desplazamiento también se observa con claridad en palabras no finales con acentos pre-nucleares.

Si el acento de frase es bajo L-, los acentos tonales pueden desplazarse hacia la izquierda, alejándose de acento L- o más aun desapareciendo en la palabra final con acento nuclear. El alejamiento es mayor en palabras finales con acento lexical agudo H+L* y menor o sin corrimiento en los otros tipos de acentos.

Tonos de Juntura

Se observan en el extremo del contorno de F0 en la última frase de la oración. Pueden ser H%, L% ó M%. Hay notaciones en la literatura donde el tono de Juntura puede contener al acento de frase en una única secuencia. Se indica como H-H% ó HH%.

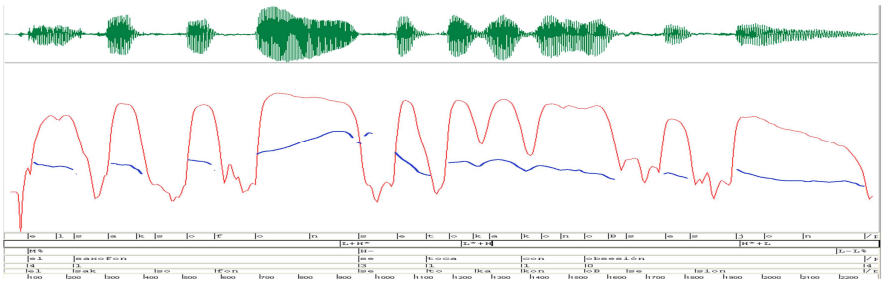


Figura 22: Declarativa.

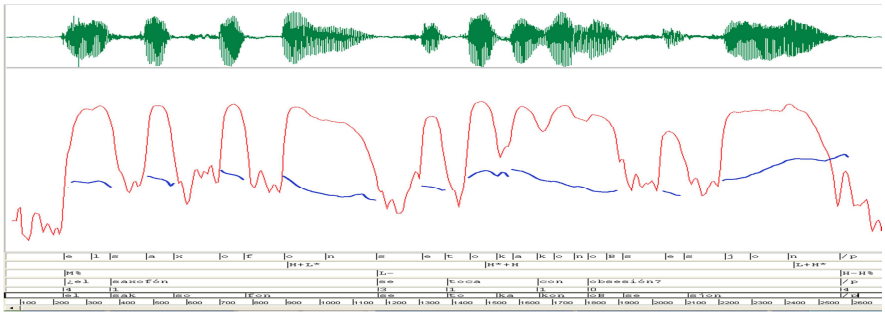


Figura 23: Interrogativa.

Conclusiones

Se han propuesto los criterios para realizar la segmentación acústico-fonética y el etiquetado fonético al nivel segmental y suprasegmental requeridos para caracterizar las bases de datos con aplicación a las tecnologías de habla. Los casos que contradicen las reglas deben indicarse para estimar estadísticamente sus frecuencias de aparición y asociarlos eventualmente como características de un locutor particular. La forma de onda permite delimitar los segmentos acústicos con mayor precisión. En segundo término se empleará el espectrograma y la evaluación perceptual de sílabas o palabras.

Referencias

[1] Wells, J.C., 1997. 'SAMPA computer readable phonetic alphabet'. Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter.

[2] Gurlekian J. A., Colantoni, L. y Torres, H., 2001 "El alfabeto fonético SAMPA y el diseño de córpora fonéticamente balanceados". Fonoaudiológica, Ed. Asalfa, 47, 3, 58-69.

[3] Guirao, M., 1980. Categorías psicoacústicas de señales de habla, Acústica 1 (2):20-49.

[4] Santagada, M.A. and Gurlekian J.A., 1993. "Spanish Voiced Stops in VCV contexts: Are they fricative variants or aproximants?", Revue de Phonetique Apliquee, Nº 90-91-92, pp. 90-100.

[5] Borzone de Manrique , A.M., 1980. La /s/ pre- consonántica en el español de Buenos Aires. Fonoaudiológica, Ed. Asalfa, 26, 1, 245-250.

[6] Martinez Soler M., Univaso P., Gurlekian J. A., "A Dissimilatory Process on the Phoneme /r/ in the Province of Corrientes, Argentina", en preparación.

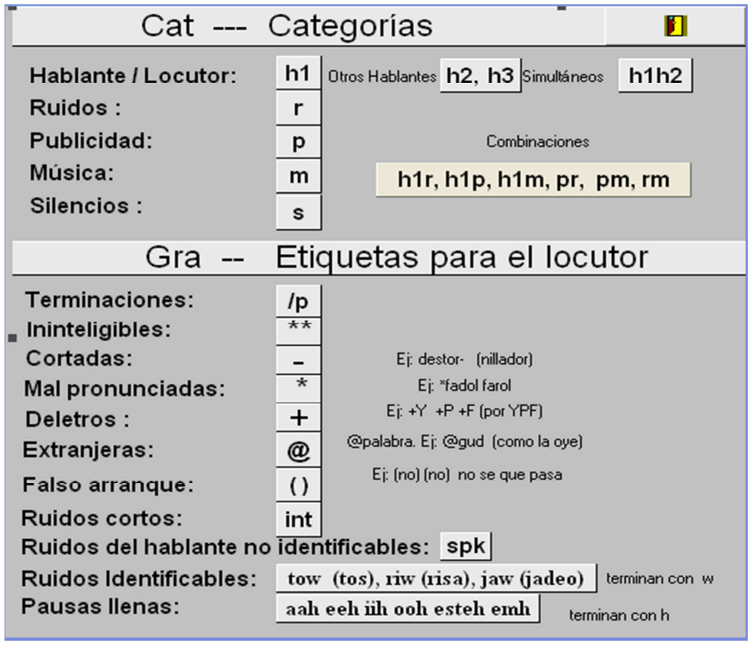
[7] Gurlekian J. A., Torres, H. y Colantoni, L., 2004. "Evaluación de las descripciones analítica y perceptual de la entonación de una base de datos de oraciones declarativas de foco amplio para el español hablado en Buenos Aires. Estudios de Fonética Experimental, Nro: XIII, pp. 275-302.

[8] Gurlekian JA y Facal, M. A., 1995. "Modelo de Informe Fonoaudiológico para el Análisis Acústico de Patologías de Habla", Fonoaudiológica, Ed. Asalfa, 41, 1, 54-68.

[9] Beckman y Gayle A. E., 1997. "The guidelines for ToBI labeling", The Ohio State University Research Foundation.

Anexo

Símbolos utilizados para la marcación de eventos multimedia.



Nivel Cat: representa las categorías gruesas que se asignan a segmentos acústicos: tales como el habla de locutores, música, ruido, publicidad, silencios. Los casos de superposición de hablantes se indican pero no se transcriben. Debe indicarse el inicio y fin de cada segmento. En la marca inicial se coloca la etiqueta de la categoría correspondiente y la marca final corresponderá al inicio del próximo segmento. La marca de silencio en este nivel es una “s”.

Nivel Gra: representa las emisiones de habla de cada locutor. Se transcriben las palabras correctas y con alteraciones que se escriben de acuerdo a la convención indicada en la figura. Tanto los ruidos identificables como las pausas llenas no descriptos en la figura pueden agregarse respetando el último carácter, respectivamente w y h. Los silencios correspondientes a la terminación de palabras y frases o pausas entre palabras se indican con /p.

B.2. Definición de las funciones de costo para la selección de unidades en AROMO *Torres H.M.*

Informe de técnico: Definición de las funciones de costo para la selección de unidades en AROMO

Autores: Dr. Bioing. Humberto Torres

Lugar: Laboratorio de Investigaciones Sensoriales, UBA-CONICET.

Fecha: 2 de Diciembre de 2013

Declaración: Las tareas realizadas en el marco de este proyecto, así como los resultados obtenidos, son de carácter confidencial. Por lo cual, el presente informe hace un listado resumido y una breve descripción de las tareas realizadas en el período indicado. Para mayor información y/o detalle, o permisos de divulgación, contactarse con el autor.

Objetivos

Definir los atributos que se emplearán para crear las funciones de costos en el sistema AROMO.

Definir las funciones de costo y sintonizar sus parámetros para la voz EMILIA

Implementar y evaluar las funciones de costo.

Resumen

La síntesis por selección de unidades consiste en elegir, de entre todas las secuencias posibles de unidades que sinteticen el texto de entrada, aquella que en forma conjunta se acerque más a la secuencia predicha y la que tenga menos ruidos por la concatenación. En este período hemos implementado un algoritmo de programación dinámica que permite encontrar la mejor secuencia posible de síntesis. El método de selección de unidades se basa en evaluar cada unidad con dos métricas: el costo de concatenación y el costo objetivo. El costo de concatenación mide el ruido generado al unir dos unidades. El costo objetivo mide que tanto se parece la unidad bajo análisis a la unidad deseada. Ambos costos luego se combinan para dar una métrica de la bondad de cada unidad posible. En este informe presentamos los resultados de nuestra evaluación de distintos atributos de las unidades para la construcción de los costos. También presentamos los resultados de la definición de las funciones de costo. Finalmente describiremos la implementación realizada para AROMO, el sistema de conversión de texto a habla del LIS.

Introducción

Los sistemas actuales de conversión de texto a voz utilizan como método de síntesis la concatenación de unidades acústicas previamente grabadas. Al momento de generar una frase, el sistema se encarga de buscar las unidades necesarias, las alinea y las concatena. Las unidades más utilizadas son los difonos, los cuales se definen como el segmento acústico entre dos puntos medios estables entre dos fonos consecutivos [1].

En los primeros sistemas de este tipo, se utilizaba una sola realización acústica de cada una de las unidades, y durante la síntesis, una vez alineadas se las posprocesaba para otorgarle las características de prosodia requeridas, utilizando métodos tales como los basados en el solapado sincrónico con la frecuencia fundamental y suma (PSOLA, del inglés (Pitch Synchronous Overlap Add Method) [12].

Si los cambios que se realizan son relativamente grandes, se genera una pérdida elevada de la calidad del sistema. Para tratar de disminuir esta degradación de la voz sintetizada se propuso tener varias realizaciones de cada una de las unidades, donde cada realización posea distintas características acústicas, de tal forma que ante un requerimiento dado, el sistema pueda seleccionar aquella unidad que más se acerque a la unidad óptima o deseada[1]. A la búsqueda de las unidades más cercanas a las deseadas, se la conoce como Selección de Unidades.

El problema de la selección de unidades puede plantearse de la siguiente forma: dada una secuencia de entrada U = {u0, u1, uN−1}, a la cual se la denomina deseada, se debe buscar en la base de datos de unidades una secuencia de unidades posibles Û = {û0, û1, ûN−1} que más se acerque a la deseada, y que se puede escribir como:

$$\hat{U}_{opt} = \text{argmin}_{\hat{U}_{opt}} C(U, \hat{U})$$

donde C es el costo de elegir la secuencia de unidades Û. En general, el costo C se descompone en dos subcostos: uno relacionado con el parecido de la unidad posible con respecto a la unidad deseada, denominado Costo objetivo $CO(U, \hat{U})$ y un costo que mide la distorsión generada al unir dos unidades posibles y al cual se lo conoce con el nombre de Costo de concatenación, $CC(\hat{U})$. Ambos subcostos se combinan de una forma lineal para dar lugar al costo total, lo cual se puede escribir como:

$$J(U, \hat{U}) = w_o CO(U, \hat{U}) + w_c CC(\hat{U})$$

donde los w_i es el peso que se le asigna a cada uno de los costos.

El costo objetivo se obtiene midiendo distintas características entre los fonos deseados y los fonos posibles. Las más utilizadas son la frecuencia fundamental, energía, contexto del cual fueron extraídos, y duración. Estos rasgos se combinan para obtener el costo objetivo, lo cual puede escribirse como:

$$CO(U, \hat{U}) = \sum_{i=0}^{N-1} CO_i(U_i, \hat{U}_i)$$

donde $CO_i(U_i, \hat{U}_i)$ es una función que mide la discrepancias del rasgo i, entre la unidad deseada U_i y la unidad posible $\hat{U}_i$. En la mayoría de los casos, la función CO sólo es la distancia entre la unidad posible y la deseada, debidamente valoradas. Uno de los grandes inconvenientes con esta aproximación es encontrar los pesos adecuados de la combinación lineal. Algunos autores proponen realizar una búsqueda exhaustiva mediante un entrenamiento [1], y otros relativizan su eficacia [13], dado que se necesitan grandes cantidades de datos para poder encontrar un conjunto de pesos que brinde un desempeño aceptable, y en su lugar fijan los pesos en forma empírica. En un trabajo anterior, utilizamos los índices de mínima diferencia perceptible (JND, del inglés Just Noticeable Difference) para medir las distancias entre las frecuencias fundamentales, las energías y las duraciones [19][10][14][19]. Estos índices miden en términos perceptuales los

[1] En inglés se utiliza el término target, que se puede traducir como objetivo. Aquí hemos rehuido a utilizar éste vocablo para evitar que se confunda con su acepción relacionada con objetivo/subjetivo.

cambios en estas magnitudes, y por lo cual no es necesario una posterior normalización de estas cantidades al momento de combinarlas en una sola medida. Un inconveniente con esta aproximación es que sólo hay JND en condiciones ideales o en ciertos contextos.

Para estimar el costo de concatenar dos unidades posibles, se procedió de forma similar a cuando se calcula el costo objetivo. Los rasgos que se tuvieron en cuenta fueron la frecuencia fundamental y la energía Además, en el caso de los sonidos sonoros se tuvieron en cuenta la la energía de los primeros 14 armónicos, y si el sonido era sordo, los valores de 14 coeficientes de mel cepstrum.

Analizar todas las posibles secuencias tiene un costo extremadamente elevado, y por lo cual es común en estos casos emplear el método de búsqueda de Viterbi. En la Figura 1 se presenta un esquema del algoritmo empleado para la búsqueda de la secuencia de unidades con menor costo total.

El cálculo de los distintos pesos es uno de los puntos críticos del método de selección de unidades [9]. Se han propuestos distintos métodos automáticos para hallar los pesos que mejor se adapten, entre los cuales se puede mencionar: Búsqueda en el Espacio de Pesos y Entrenamiento por Regresión. Ambos tienen un alto costo computacional, y sólo se obtienen resultados aceptables si se posee la suficiente cantidad de datos de entrenamiento. En otros casos se opta por fijar el valor de los pesos empíricamente, mediante una búsqueda manual de prueba y error [13]. En este trabajo se fijaron los pesos manualmente, con la consigna de minimizar los artefactos generados por la concatenación.

Atributos para creación de la función del costo objetivo

A continuación detallaremos los atributos empleados para definir la función de costo objetivo.

Frecuencia fundamental

El contorno de la frecuencia fundamental (F0) es uno de los rasgos u atributos más importantes en la naturalidad del habla. Revela la modalidad oracional (exclamación, pregunta, afirmación) [7], estructura las oraciones en frases o sintagmas entonativos, permite destacar determinados segmentos del discurso, permite desambiguar palabras, entre otros.

Duración segmental

La duración de los fonos es el parámetro físico asociado con el concepto perceptual del ritmo, y uno de los tres atributos del habla que más influyen en su naturalidad.

Energía

La energía de la señal de habla es otro de los parámetros responsable de la naturalidad del habla. Para que una frase se perciba como natural, la energía de los fonos componentes deben guardar cierta relación.

Sílaba final

Los rasgos de las sílabas de final de oración se distinguen del resto de las sílabas, tanto en su contenido frecuencial como en su duración [6][18].

Sílaba antes de una pausa

Al igual que la sílaba final, las sílabas antes de una pausa tienen rasgos propios, que hacen la naturalidad del habla [6]. La duración de los fonos también se ve afectada por este atributo [18]

Acento léxico

Las sílabas con acento léxico de palabras de contenido tienen, en la mayoría de los casos, asociado un movimiento tonal [11]. La duración de los fonos también se ve afectada por este atributo [18].

Contexto fonético

En habla continua los fonemas se realizan como fonos que se articulan con el fono que le precede y el fono que le sigue. Algunos fonos se ven tan afectados que se realizan como alófonos [5].

Clase de palabras

No todas las palabras tienen la misma importancia dentro de la estructura de una frase. La clase de palabra influye directamente sobre todos los rasgos prosódicos de una frase [6][7][15][16][18].

Monosílabo

Las palabras monosilábicas tienen características acústicas que en general se diferencian del resto de las palabras. Por ejemplo, su duración y energía promedio son menores que las palabras de más de un sílaba.

Definición de la función de costo objetivo

La función de costo objetivo se define como la suma de subcostos, como:

$$\mathrm{CO} = \mathrm{CO}_{\mathrm{F0}} + \mathrm{CO}_{\mathrm{dur}} + \mathrm{CO}_{\mathrm{final}} + \mathrm{CO}_{\mathrm{acento}} + \mathrm{CO}_{\mathrm{pausa}} + \mathrm{CO}_{\mathrm{contexto}} + \mathrm{CO}_{\mathrm{monosílabo}}$$

A continuación definiremos las funciones de subcostos.

Frecuencia fundamental

El sistema AROMO tiene un sistema de predicción de F0 [15][16], a la cual llamaremos frecuencia fundamental deseada ($\mathrm{F0_d}$). La frecuencia fundamental de cada unidad de síntesis ($\mathrm{F0_u}$) es estimada en tres puntos: inicio, medio y final.

El subcosto de frecuencia fundamental ($\mathrm{CO_{F0}}$) se define como la suma de las diferencias absolutas, en ERBs (del inglés Equivalent Rectangular Bandwidth Scale)

$$\mathrm{CO_{F0}} = |\mathrm{F0_{d\ inicio.}} - \mathrm{F0_{u\ inicio}}| + |\mathrm{F0_{d\ medio}} - \mathrm{F0_{u\ medio}}| + |\mathrm{F0_{d\ fin}} - \mathrm{F0_{u\ fin}}|$$

Duración segmental

Aromo predice la duración de cada fono ($\mathrm{Dur_d}$) a partir de rasgos extraídos del texto de entrada. La duración de cada una de las unidades posibles ($\mathrm{Dur_u}$) está almacenada en el corpus de datos.

El subcosto de duración ($\mathrm{CO_{dur}}$) se define como:

$$\mathrm{CO_{dur}} = \begin{cases} \alpha_{\mathrm{dur}}|\mathrm{Dur_d} - \mathrm{Dur_u}| & \text{si } |\mathrm{Dur_d} - \mathrm{Dur_u}| < \mathrm{Umbral_{dur}} \\ \beta_{\mathrm{dur}}|\mathrm{Dur_d} - \mathrm{Dur_u}| & \text{en otro caso} \end{cases}$$

Con $\alpha_{\mathrm{dur}} \ll \beta_{\mathrm{dur}}$, y que permite rechazar unidades cuya duración se alejen mas allá del valor de $\mathrm{Umbral_{dur}}$.

Energía

La función de subcosto de energía tiene por objetivo

Sílaba final

Cada unidad de síntesis en el corpus tiene un atributo que indica si pertenece o no a una sílaba en final de la oración.

El subcosto de sílaba final (CO_{final}) se define como como un valor fijo (F), dependiendo si el atributo del la unidad de síntesis coincide con el deseado o no. Esto puede escribirse como:

$$CO_{final} = \begin{pmatrix} \text{¿ Unidad deseada pertenece a una sílaba final?} \\ \text{XOR} \\ \text{¿ Unidad de síntesis pertenece a una sílaba final?} \end{pmatrix} * F$$

Acento léxico

Cada unidad de síntesis en el corpus tiene un atributo que indica si pertenece o no a una sílaba con acento léxico. En AROMO, a partir del texto de entrada se etiqueta a las unidades deseadas si poseen acento léxico, con y sin acento ortográfico. Además, a cada palabra se le asigna un atributo de ser o no una palabra de contenido [17].

El subcosto asociado al acento léxico (CO_{acento}) se define como como un valor fijo (A), dependiendo si el atributo del la unidad de síntesis coincide con el deseado o no. Esto puede escribirse como:

$$CO_{acento} = \begin{pmatrix} \text{¿ Unidad deseada pertenece a una sílaba con acento léxico?} \\ \text{XOR} \\ \text{¿ Unidad de síntesis pertence a una sílaba con acento lexico?} \end{pmatrix} * A$$

Sílaba antes de una pausa

En AROMO, las pausas tienen tres orígenes: marcadas por signos ortográficos, las forzadas por comando, y las predichas por el sistema. Cada unidad de síntesis en el corpus tiene un atributo que indica si pertenece o no a una sílaba antes de una pausa.

El subcosto de sílaba antes de una pausa (CO_{pausa}) se define como como un valor fijo (P), dependiendo si el atributo del la unidad de síntesis coincide con el deseado o no. Esto puede escribirse como:

$$CO_{pausa} = \begin{pmatrix} \text{¿ Unidad deseada pertenece a una sílaba antes de una pausa?} \\ \text{XOR} \\ \text{¿ Unidad de síntesis pertence a una sílaba antes de una pausa?} \end{pmatrix} * P$$

Contexto fonético

El subcosto de contexto fonético ($CO_{contexto}$) puede tomar dos valores, dependiendo si hay un epéntesis (C_{ce}) , o no (C_{se}), dependiendo si el atributo del la unidad de síntesis coincide con el deseado o no. Esto puede escribirse como:

$$CO_{contexto} = \text{¿ Unidad deseada tiene el mismo contexto fonético que la unidad de síntesis?} *$$
$$\{ (\text{¿ alófono con epéntesis?} * (C_{ce})) \text{XOR } (\text{¿ alófono sin epéntesis?} * (C_{se})) \}$$

Clase de palabras

El subcosto de clases de palabras (CO_{pos}) se define como como un valor fijo (P), dependiendo si el atributo del la unidad de síntesis coincide con el deseado o no. Esto puede escribirse como:

$$CO_{pos} = \text{¿ Unidad deseada y la unidad de síntesis pertencen a palabras de la misma clase ?} * P$$

Monosílabo

El subcosto de monosílabo ($CO_{monosílabo}$) se define como como un valor fijo (M), dependiendo si el atributo del la unidad de síntesis coincide con el deseado o no. Esto puede escribirse como:

$$\mathrm{CO}_{\text{monosílabo}} = \begin{pmatrix} \text{¿ Unidad deseada pertenece a un monosílabo?} \\ \text{XOR} \\ \text{¿ Unidad de síntesis pertence a un mpnosílabo?} \end{pmatrix} * \mathrm{M}$$

Atributos para creación de la función del costo concatenación

Los atributos empleados para construir la función de costo de concatenación son similares a los utilizados al definir la función de costo objetivo.

Definición de la función de costo de concatenación

La función de costo objetivo se define como la suma de subcostos, como:

$$\mathrm{CC} = \mathrm{CC}_{\mathrm{F0}} + \mathrm{CC}_{\mathrm{E}} + \mathrm{CC}_{\mathrm{F}} + \mathrm{CC}_{\text{acento}} + \mathrm{CC}_{\text{pausa}} + \mathrm{CC}_{\text{contexto}} + \mathrm{CC}_{\text{monosílabo}}$$

Frecuencia fundamental

Para el subcosto de frecuencia fundamental ($\mathrm{CC}_{\mathrm{F0}}$) se intentaron varias aproximaciones. La primera basada en JND, obteniendo una función de costo de la forma:

$$\mathrm{CC}_{\mathrm{F0}} = \mathrm{JND}_{\mathrm{F0}} * \left|\mathrm{F0}_{\text{izq}} - \mathrm{F0}_{\text{der}}\right|$$

Donde $\mathrm{JND}_{\mathrm{F0}}$ es el valor de JND de frecuencias para tonos puros, $\mathrm{F0}_{\text{izq}}$ y $\mathrm{F0}_{\text{der}}$ son los valores de frecuencia fundamental de las unidades de síntesis a en el punto de unión.

Como comentamos en la introducción, solo están definidos JND para tonos puros, y los resultados obtenidos no fueron satisfactorios. En una segunda aproximación se busco un valor de $\mathrm{JND}_{\mathrm{F0}}$, en forma manual. Los resultados fueron mejores que los obtenidos con el método anterior, pero aun insatisfactorios. En base al análisis de los resultados obtenidos con esta aproximación, supusimos que la falencia estaba en suponer una función lineal. En [4] se propone utilizar funciones de costo lineales por tramos, donde las funciones deben definirse a priori. Esta aproximación ofreció mejores resultados que los anteriores, pero la sintonización de las funciones es manual, y para lograr buenos resultados el trabajo de ajuste de las curvas es muy laborioso, y en general se cae en mínimos locales inestables, de forma tal que si las condiciones de contorno cambian, los resultados ya no son satisfactorios.

Finalmente propusimos una nueva aproximación a la resolución del problema. Planteamos el problema bajo la hipótesis de que la función de costo debe medir la bondad de la unión, teniendo en cuenta todas las uniones posibles. Con esto en mente, estimamos la distribución acumulada de las diferencias absolutas en F0 de todas las uniones posibles de unidades. En la Figura 1 se presentan las distribuciones acumuladas estimadas. En rojo están las distribuciones discriminadas por fonemas, en verde la distribución promedio, y en azul la función de costo propuesta, que estimada con la Ec. 1.

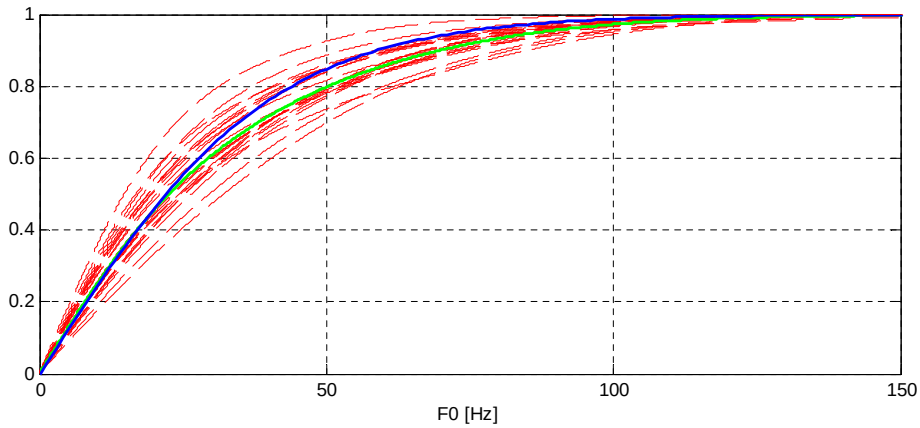


Figura 1: Distribuciones acumuladas estimadas de diferencias de F0 en las uniones de unidades. En rojo están las distribuciones discriminadas por fonemas, en verde la distribución promedio, y en azul la función de costo propuesta.

$$CC_{F0} = \begin{cases} \alpha_{F0} * \left(\frac{2}{1+e^{-\beta_{F0} * |F0_{izq} - F0_{der}|}} - 1 \right) & \text{si } |F0_{izq} - F0_{der}| < Umbral_{F0} \\ \gamma_{F0} * \left(\frac{2}{1+e^{-\beta_{F0} * |F0_{izq} - F0_{der}|}} - 1 \right) & \text{en otro caso} \end{cases} \quad (1)$$

Con $\alpha_{F0} \ll \gamma_{F0}$, y que permite rechazar unidades cuya duración se alejen mas allá del valor de $Umbral_{F0}$.

Una aproximación similar se utilizó para definir el subcosto asociado al coeficiente cero de Mel Cepstrum (CC_{C0}). E la Figura 2 se presentan las curvas estimadas, y en la Ec. 2 la función de costo propuesta.

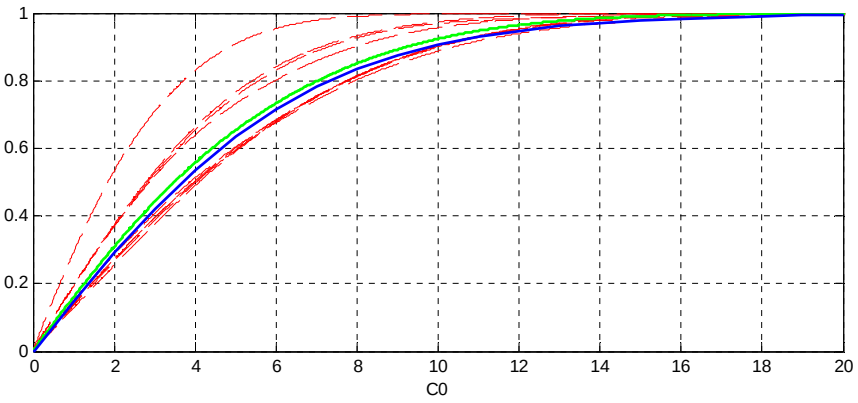


Figura 2: Distribuciones acumuladas estimadas de diferencias de C0 en las uniones de unidades. En rojo están las distribuciones discriminadas por fonemas, en verde la distribución promedio, y en azul la función de costo propuesta.

$$CC_{C0} = \begin{cases} \alpha_{C0} * \left(\frac{2}{1+e^{-\beta_{C0} * |C0_{izq} - C0_{der}|}} - 1 \right) & \text{si } |C0_{izq} - C0_{der}| < Umbral_{C0} \\ \gamma_{C0} * \left(\frac{2}{1+e^{-\beta_{C0} * |C0_{izq} - C0_{der}|}} - 1 \right) & \text{en otro caso} \end{cases} \quad (2)$$

Energía

Para la energía se plantearon aproximaciones similares a las propuestas para la frecuencia fundamental.

En la Figura 3 se presentan las curvas estimadas, y en la Ec. 3 la función de subcostos propuesta.

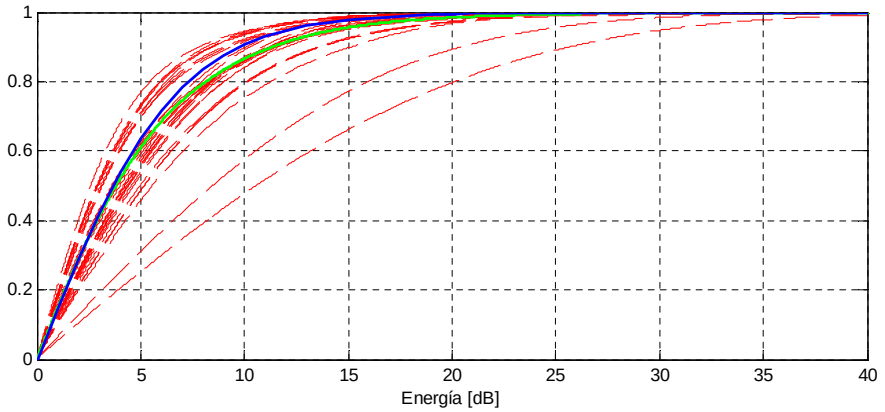


Figura 3: Distribuciones acumuladas estimadas de diferencias de energía en las uniones de unidades. En rojo están las distribuciones discriminadas por fonemas, en verde la distribución promedio, y en azul la función de costo propuesta.

$$CC_E = \begin{cases} \alpha_E * \left(\frac{2}{1+e^{-\beta_E * \left| E_{izq} - E_{der} \right|}} - 1 \right) & \text{si } \left| E_{izq} - E_{der} \right| < Umbral_E \\ \gamma_E * \left(\frac{2}{1+e^{-\beta_E * \left| E_{izq} - E_{der} \right|}} - 1 \right) & \text{en otro caso} \end{cases} \quad (3)$$

Distribución de la energía en frecuencia

Para la distribución de la energía en frecuencia se utilizó quince coeficientes mel cepstrum. Para definir la función de subcostos se plantearon aproximaciones similares a las propuestas para la frecuencia fundamental. Al estimar las distribuciones acumuladas se observaron diferencias apreciables para los diferentes coeficientes. Por lo cual, se hizo tres agrupamientos de coeficientes. En las Figuras 4, 5 y 6 se presentan las curvas estimadas, y en las Ec. 4,5 y 6 las funciones de subcostos propuestas.

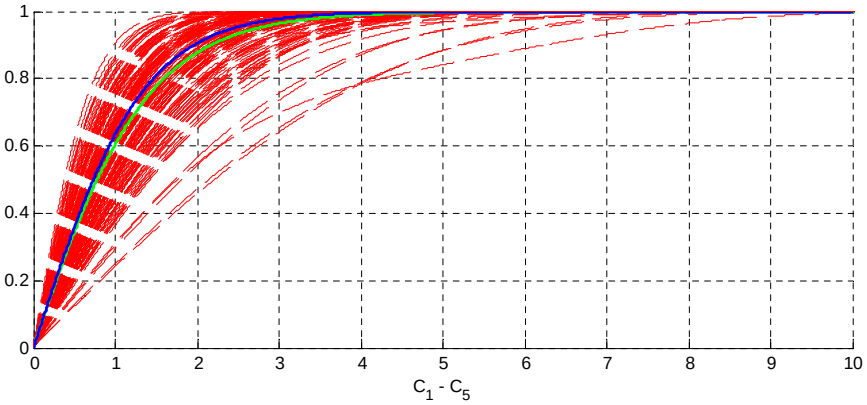


Figura 4: Distribuciones acumuladas estimadas de diferencias de los coeficientes 1 a 5 de mel cepstrum, en las uniones de unidades. En rojo están las distribuciones discriminadas por fonemas, en verde la distribución promedio, y en azul la función de costo propuesta.

$$CC_{C1_5} = \begin{cases} \alpha_{C1_5} * \left(\frac{2}{1+e^{-\beta_{C1_5} * \left| C1_5_{izq} - C1_5_{der} \right|}} - 1 \right) & \text{si } \left| C1_5_{izq} - C1_5_{der} \right| < Umbral_{C1_5} \\ \gamma_{C1_5} * \left(\frac{2}{1+e^{-\beta_{C1_5} * \left| C1_5_{izq} - C1_5_{der} \right|}} - 1 \right) & \text{en otro caso} \end{cases} \quad (4)$$

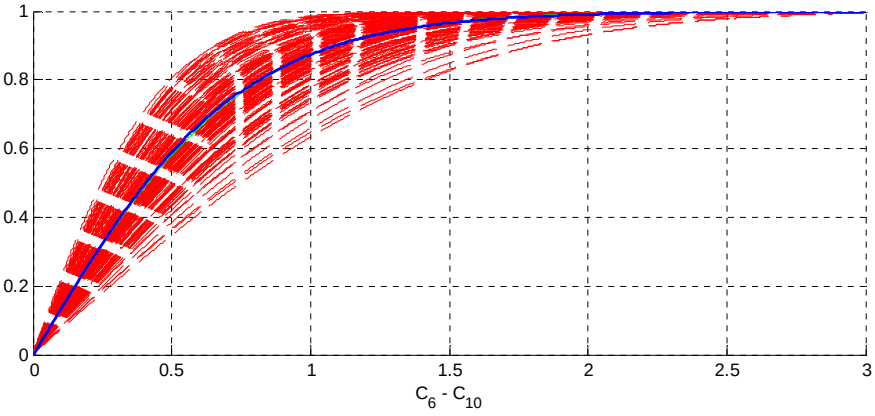


Figura 5: Distribuciones acumuladas estimadas de diferencias de los coeficientes 6 a 10 de mel cepstrum, en las uniones de unidades. En rojo están las distribuciones discriminadas por fonemas, en verde la distribución promedio, y en azul la función de costo propuesta.

$$CC_{C6_10} = \begin{cases} \alpha_{C6_10} * \left(\frac{2}{1+e^{-\beta_{C6_10} * \left| C6_10_{izq} - C6_10_{der} \right|}} - 1 \right) & \text{si } \left| C6_10_{izq} - C6_10_{der} \right| < Umbral_{C6_10} \\ \gamma_{C6_10} * \left(\frac{2}{1+e^{-\beta_{C6_10} * \left| C6_10_{izq} - C6_10_{der} \right|}} - 1 \right) & \text{en otro caso} \end{cases} \quad (5)$$

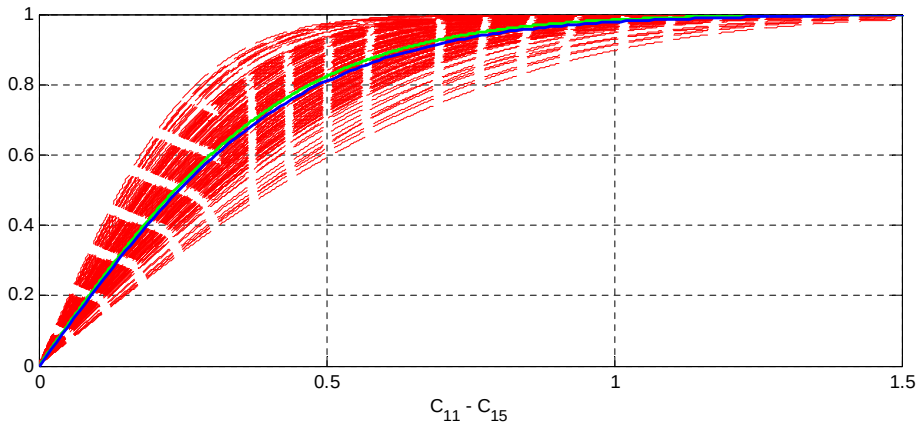


Figura 6: Distribuciones acumuladas estimadas de diferencias de los coeficientes 11 a 15 de mel cepstrum, en las uniones de unidades. En rojo están las distribuciones discriminadas por fonemas, en verde la distribución promedio, y en azul la función de costo propuesta.

$$CC_{C11_15} = \begin{cases} \alpha_{C11_15} * \left(\frac{2}{1+e^{-\beta_{C11_15} * \left| C11_15_{izq} - C11_15_{der} \right|}} - 1 \right) & \text{si } \left| C11_15_{izq} - C11_15_{der} \right| < Umbral_{C11_15} \\ \gamma_{C11_15} * \left(\frac{2}{1+e^{-\beta_{C11_15} * \left| CC11_15_{izq} - CC11_15_{der} \right|}} - 1 \right) & \text{en otro caso} \end{cases} \quad (5)$$

Acento léxico

El subcosto asociado al acento léxico (CC_{acento}) se define como como un valor fijo (A), dependiendo si los atributos de las unidades a concatenar coinciden o no. Esto puede escribirse como:

$$CC_{acento} = \begin{pmatrix} \text{¿ La unidad de síntesis a la izquierda de la unión pertenece a una sílaba con acento léxico?} \\ \text{XOR} \\ \text{¿ La unidad de síntesis a la derecha de la unión pertence a una sílaba con acento lexico?} \end{pmatrix} * A$$

Contexto fonético

El subcosto de contexto fonético ($CC_{contexto}$) se define como como un valor fijo (C), dependiendo si los atributos de las unidades a concatenar coinciden o no. Esto puede escribirse como:

$$CO_{contexto} = (\text{¿ Las unidades a concatenar tiene el mismo contexto fonético?}) * C$$

Monosílabo

El subcosto de monosílabo ($CC_{monosílabo}$) se define como como un valor fijo (M), dependiendo si los atributos de las unidades a concatenar coinciden o no. Esto puede escribirse como:

$$CC_{monosílabo} = \begin{pmatrix} \text{¿ La unidad de síntesis a la izquierda de la unión pertenece a un monosílabo?} \\ \text{XOR} \\ \text{¿ La unidad de síntesis a la derecha de la unión pertence a un mpnosílabo?} \end{pmatrix} * M$$

Clase de palabras

El subcosto de clases de palabras (CC_{pos}) se define como como un valor fijo (P), dependiendo si los atributos de las unidades a concatenar coinciden o no. Esto puede escribirse como:

$$CC_{pos} = \text{¿ Las unidades a concatenar pertencen a palabras de la misma clase ?} * P$$

Estimación de los parámetros de las funciones de costo

Las funciones de costo presentadas arriba tienen uno o varios parámetros que se deben ajustar para lograr un desempeño aceptable. Se han presentado algunos trabajos relativos a estimar en forma automática los valores de estos parámetros [1][3][9]. Una desventaja de estas propuestas es que se necesitan grandes cantidades de datos etiquetadas, lo que limita fuertemente su implementación. Actualmente, la técnica más empleada es la de fijar los valores de parámetros de las funciones en forma manual, mediante algún tipo de búsqueda sistemática o aleatoria. Esta es la aproximación que hemos implementado para AROMO.

Evaluaciones perceptuales preliminares han demostrado el buen desempeño de las funciones de costo propuestas. Resta realizar una evaluación perceptual sistemática del sistema.

Referencias

[1] Bellegarda, J., A novel approach to cost weighting in unit selection TTS. In Proc. of 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009), pp. 744-747. Brighton, United Kingdom, September 2009.

[2] Black, A. and Campbell, N., Optimizing selection of units from speech databases for concatenative synthesis. In Proc. of 4th European Conference on Speech Communication and Technology (EUROSPEECH'95), pp. 581–584, Madrid, Spain, September 1995.

[3] Campbell W. and Black, A., Prosody and the Selection of Source Units for Concatenative Synthesis. In Progress in Speech Synthesis, J. van Santen, R. Sproat, J. Hirschberg, and J. Olive, Eds., New York, NY: Springer–Verlag, pp. 279–292, 1997.

[4] Coorman, G.; Fackrell, J.; Rutten, P. & Coile, B. V., Segment selection in the L&h Realspeak laboratory TTS system. In Proc. of the Int. Conf. on Spoken Language Processing, pp. 395-398, Beijin, China, 2000.

[5] Gurlekian, J., Colantoni, L. y Torres, H., El alfabeto fonético SAMPA y el diseño de corpora fonéticamente balanceados, Fonoaudiológica Nº 3, pp. 58-69. Diciembre de 2001.

[6] Gurlekian, J., Mixdorff, H., Evin, D., Torres, H., Pfitzinger, H., Alignment of F0 model parameters with final and non-final accents in Argentinean Spanish. In Proc. of 5th International Conference on Speech Prosody 2010, Chicago, Illinios, USA. May 11-14, 2010.

[7] Gurlekian, J., Evin, D., Mixdorff, H., Torres, H., Pfitzinger, H., Accent command model parameter alignment in Argentine Spanish absolute interrogatives. In Proc. of the 21 Konferenz Elektronische Sprachsignalverarbeitung (ESSV 2010), pp. 77-83, TUDpress Verlag der Wissenschaften, Berlin, Germany, September 8-10, 2010.

[8] Huggins, A. W. F.: Just Noticeable Differences for Segment Duration in Natural Speech. J. Acoust. Soc. Am., 51(4):1270–1278, 1972.

[9] Hunt, A.J. and Black, A.: Unit selection in a concatenative speech synthesis system using a large speech database. In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96), Vol. 1, pp. 373–376, Atlanta, Georgia, May 1996.

[10] Klatt, D.: Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. J. Acoust. Soc. Am., 53:8–16, 1973.

[11] Ladd, R., Intonational phonology. Cambridge: Cambridge University Press.1996.

[12] Moulines, E. y F. Charpentier: Pitch-SynchronousWaveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. Speech Communication, 9:453–467, 1990.

[13] Prudon, R. y d'Alessandro, C.: A selection /concatenation text to speech synthesis system: databases development, system design, comparative evaluation. In Proc. of the 4th Speech Synthesis Workshop (SSW4-2001), Pitlochvy, Scotland, August 2001. paper 138.

[14] Quené, Hugo: On the just-noticeable difference for tempo in speech. J. Acoust. Soc. Am., 15(5):2607–2607, 2001.

[15] Torres, H.: Generación automática de la prosodia para un sistema de conversión de texto a habla, Tesis Doctoral, Facultad de Ingeniería, Universidad de Buenos Aires. Agosto de 2008.

[16] Torres, H. and Gurlekian, J.: Parameter estimation and prediction from text for a superpositional intonation model. In Proc. of the 20 Konferenz Elektronische Sprachsignalverarbeitung (ESSV 2009), pp. 238-247, TUDpress Verlag der Wissenschaften, Dresden, Germany, September 21-23, 2009.

[17] Torres, H., Informe de técnico: Etiquetado de clase de palabras y su aplicación en un sistema TTS. Informe anual del Laboratorio de Investigaciones Sensoriales, UBA-CONICET. Diciembre de 2011.

[18] Torres, H. and Gurlekian, J., Argentine Spanish segmental duration prediction, In Proc. of 13th Argentine Symposium on Technology, 41 JAIIO, pp. 156-167, La Plata, Agosto de 2012.

[19] Yilmaz, Hüseyin: A theory of speech perception. Bulletin of Mathematical Biology, 29(4):793–825, December 1967.

B.3. Segmentación de audio de radio y televisión. *Torres H.M.*

Informe de técnico: Segmentación de audio de radio y televisión

Autor: Dr. Bioing. Humberto Torres

Lugar: Laboratorio de Investigaciones Sensoriales, INIGEM, CONICET-UBA.

Fecha: 12 de Diciembre de 2013

Declaración: Las tareas realizadas en el marco de este proyecto, así como los resultados obtenidos, son de carácter confidencial. Por lo cual, el presente informe hace un listado resumido y una breve descripción de las tareas realizadas en el período indicado. Para mayor información y/o detalle, o permisos de divulgación, contactarse con el autor.

Objetivos

Crear las herramientas necesarias para la desgrabación y etiquetado de eventos de un corpus de entrenamiento y prueba para el reconocimiento de habla en audios de radio y televisión.

Introducción

La desgravación de audios es una de las primeras tareas en la creación de corpus para el entrenamiento y prueba de sistemas de reconocimiento de habla [4]. Se debe transcribir literalmente lo hablado y además marcar todos los eventos acústicos. También puede ser de interés indicar el cambio de locutor, superposiciones de locutores, la presencia de ruidos, entre otros. Existen herramientas gráficas que permite leer, escuchar, ver, editar y etiquetar audios de locuciones [1]. Este tipo de herramientas presentan varios inconvenientes de uso, por ejemplo: si los audios son de gran longitud el programa se enlentece y/o dejan de funcionar; es necesario marcar el inicio y fin de cada turno de habla antes de desgrabar; es engorroso moverse entre los distintos campos de etiquetado; entre otros problemas detectados. Una alternativa es segmentar los audios antes de su desgrabación manual. El criterio de segmentación puede tener como motivación: pausas, excesivo ruido, por tiempo, por eventos acústicos, ente otros. Una vez realizada la segmentación, los audios resultantes pueden ser guardados en archivos independientes, facilitando su manipulación y desgrabación. En este trabajo nos propusimos segmentar audios de radio y televisión para su posterior desgravación y etiquetado de eventos acústicos. Para ello se crearon dos herramientas: uno que permite segmentar los audios, y otra que permite su desgrabación y etiquetado.

Este trabajo se enmarca dentro del Servicio Tecnológico de Alto Nivel (STAN) “Creación de modelos acústicos y de lenguaje para el reconocimiento de habla en radio y televisión”con la empresa Global News [2].

Segmentación de audio

La segmentación de los audios se realizó utilizando un detector de habla, más conocido como detector de actividad vocal (VAD, del inglés Voice Activity Detection). Luego de un extensa búsqueda bibliográfica se optó por el método propuesto por Ramírez et al. [3]. El método es robusto al ruido ambiente, y originalmente pensado para el reconocimiento de habla.

El algoritmo fue implementado en lenguaje Matlab y adaptado a las necesidades de la tarea: se modificó la forma de inicialización del algoritmo; se ajustaron los valores de los parámetros internos; se agregó un detector de falta de señal; se agregó un filtro para eliminar las activaciones espurreas; se cambiaron los tamaños de las ventanas de análisis; entre otras modificaciones.

En la Figura 1 se presenta una porción de audio de una radio, con un ruido que la enmascara, y la salida del VAD implementado.

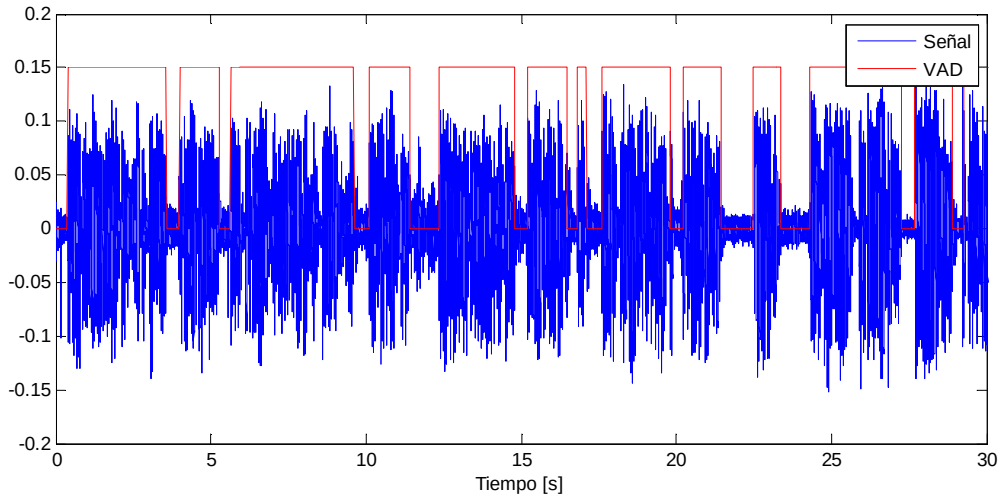


Figura 1: Ejemplo de un una señal de audio radial, inmersa en un ruido, y la salida del VAD implementado.

Una vez detectados los segmentos de audio con habla, se procedió a cortar dichos segmentos y guardarlos en archivos individuales. El nombre de los archivos creados se forma con el nombre original, y los tiempos de inicio y fin del segmento de audio cortado.

Desgrabación de audio

Para desgrabar y etiquetar los eventos acústicos de los archivos de audio generados se adaptó una herramienta desarrollada en el laboratorio por el Dr. Bioing. Diego Evin. Las modificaciones fueron realizadas con dos propósitos: agregar funcionalidades y versatilidad a la herramienta; generar los archivos de desgrabación en el formato especificado.

Conclusiones

Las herramientas desarrolladas cumplen con los objetivos planteados. Actualmente, para la creación del corpus, los archivos de audio de radio y televisión suministrados por la empresa son procesados con la herramienta de segmentación, y los archivos resultantes son entregados a los etiquetadores humanos que utilizan TranscriptorLIS para desgrabar el audio. Los etiquetadores

han reportado que la herramienta es muy útil y fácil de usar. A partir del empleo de las herramientas desarrolladas los tiempos de etiquetado se redujeron en un 80%, con respecto a la aplicación empleada con anterioridad.

Referencias

[1] http://transag.sourceforge.net/

[2] http://www.globalnewsgroup.com

[3] Javier Ramirez *, Jose C. Segura 1, Carmen Benitez, Angel de la Torre, Antonio Rubio, "Efficient voice activity detection algorithms using long-term speech information, Speech Communication 42, pp. 271-287. 2004.

[4] J. Gurlekian, H. Torres: “Informe Técnico: Guía para la segmentación y transcripción fonética para las tecnologías de habla”. Informe anual del Laboratorio de Investigaciones Sensoriales (ISSN 0325-2043), LIS, INIGEM, CONICET-UBA. Diciembre de 2013.

B.4. Internal vs. external standards for the evaluation of F0 perturbation.

Gurlekian J.A., Torres, H.M. y Vaccari, M.E.

Technical Report: Internal vs. external standards for the evaluation of F0 perturbation

Autors: Jorge A. Gurlekian, Humberto M. Torres, María E. Vaccari

Lab: Laboratorio de Investigaciones Sensoriales, INIGEM, CONICET-UBA.

Date: 2 de diciembre de 2013

Abstract

The goal of this work is to explore two protocols for the perceptual evaluation of F0 perturbation. One presentation is based on a scale of limits and uses subject's internal reference and the second is a matching procedure where each stimulus is compared with the entire set to find the match. Responses of three group of listeners -speech therapists both experts and students, and common listeners- are shown when systematic variations of fundamental frequency for vowel /a/ were presented to them. Variations were produced by LPC synthesis using a F0 contour function obtained from a statistical jitter model. Seven jitter values were employed for each of two initial F0 values. Responses to estimation present higher deviation and less repeatability than the obtained for matching. For the second experiment there are no significant intra-subject differences for the answers of experts and students. We conclude that individual differences between speech therapist could be better managed when using the matching paradigm than using a magnitude scale.

Introduction

It is well known between speech therapists the lack of agreement and reliability on the perceptual evaluation of dysphonic voices [1]. Particularly it happens when variations of fundamental frequency and amplitude occur. Clinic experts use their own internal references to judge perturbation. Perceptual judgments using a variety of experimental tasks like equal appearing intervals, visual analog, and magnitude estimation listeners also must rely in an internal reference or scale. So, the advice to them is to develop that internal standard during many years of listening. To perform this task in a reliable way the professional should acquire a kind of absolute ear for frequency perturbation. Even if they accomplish the task in a repeatability way, searching for inter judge's consensus is less probable. Sensation thresholds to F0 change, expected vocal ranges according to experience, location of internal standard are some of the conditions that contribute to the variability, ie: young professionals with less exposure to pathological voices are likely to have different standards. Even for some experienced listeners seems that the internal standard is inherently unstable and the judgments shift relative to the set of stimuli used in the experiment.

In a tutorial and review paper [1] an external reference was referred as a possible solution to the variability sources. This work will present an experiment based on the equal/different discrimination capacity in a series of synthetic stimuli where different references are used. Repeatability and agreement between subjects will be compared with a task based on the use of an internal reference.

Experimental procedure

Stimuli

Based on the acoustic parameters of natural female and male vowels /a/ indicated on Table 1, two synthesized versions were created by a LPC formant synthesizer. Once the F0 was selected (120Hz or 240 Hz) it remains constant for all the duration with a jitter value of 0%. This stimulus will be at one extreme of the continuum. Six additional jitter values completed the linear sequence of values with variation percentages indicated on Table 2.

Table 1: Stimulus F0 & formant values, in Hz.

	Male	Female
F0	120	240
F1	700	900
F2	1250	1450
F3	2500	2500
F4	3500	3500
F5	4000	4000

Table 2: Stimulus Jitter values.

	Intended	Real	Dif
1	0.00	0.00	0.00
2	0.50	0.50	0.50
3	1.00	1.00	1.00
4	1.50	1.50	1.50
5	2.00	2.00	2.00
6	2.50	2.50	2.50
7	3.00	3.00	3.00

To produce these intended jitter values, fundamental frequency variation along time was calculated. One way is to introduce random noise in the glotal source to produce controlled variations [3]. The method which was used here employs a statistical model of jitter [4]. We choose the definition of percent jitter (1) given by the average of the difference between two F0 values, normalized to the average F0 and multiplied by 100.

$$J_{\%} = 100\ \frac{\frac{1}{N-1}\sum_{i=1}^{N}|F0_i - F0_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} F0_i} \qquad (1)$$

Where $F0_i$ is the fundamental frequency cycle to cycle and N the number of cycles.

It is assumed that F0 variations are independent from amplitude variation. According to Titze [5] and Torres et al [6] $F0_i$ values have a Gaussian behavior with normal density probability functions $\mathcal{N}(F0, \sigma_{F0})$. In this way is possible to synthesize vowels from reference F0 values and each intended jitter value. The F0 values set are like a Gaussian noise distribution with average F0 and standard deviation given by the equation (2)

$$\sigma_{F0} = \frac{\sqrt{\pi}\ F0\ J_{\%}}{200} \qquad (2)$$

Once the F0 set which represent the glotal source is defined, the stimuli is synthesized using the LPC method [2]. Stimuli are further verified to control that effectively have the programmed perturbation. Differences between programmed and real values do not exceed 5%. See Table 2.

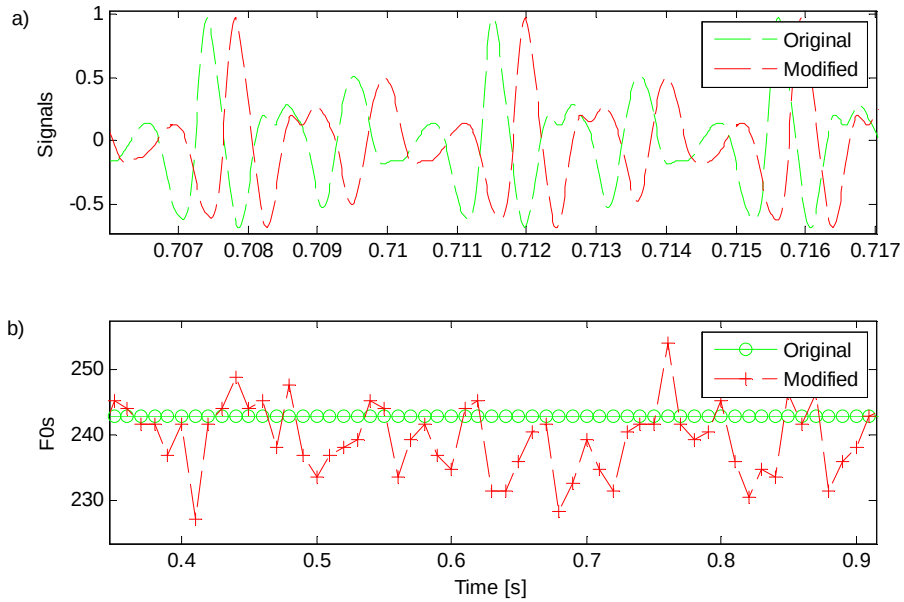


Figure 1: Original and modified stimuli signals, without and with a jitter of 3, respectively. a) signals and b) F0s.

Test protocols

For test #1, stimuli were presented at random for each listener with 5 repetitions each and 4 sec of inter-stimulus interval. Instruction to subjects was: you will hear the vowel /a/, your task consist to assign a number to each stimuli according to a perturbation scale of your preference. Always assign high numbers to stimuli with high perturbation. Before you start you can hear stimuli close to both ends of the scale and one intermediate to get familiar with the stimuli set. We asked also to give a name to the perturbation they heard. By doing this training we also checked that subjects understand the instruction and are capable to do the test.

For test #2, stimuli were presented in random sequences for each listener with 5 repetitions and for each repetition is available for play whenever the subject press a test button. Instruction to subjects was: you will hear the vowel /a/ when you press the test button, your task consist to match the test stimuli with any of the stimuli available to you at the screen through seven different buttons. Once you decide the match you should press next button to hear a new test stimuli and repeat the procedure.

In this way each of the stimuli of the set acts five times as an external reference against the complete set.

Subjects

Group A: Ten highly experienced speech therapists from different Buenos Aires hospitals and clinics.

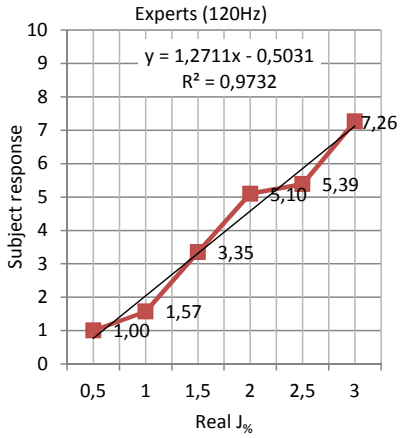
Group B: Ten students from the 3th and 4th year career of Fonoaudiologist at the University of Buenos Aires.

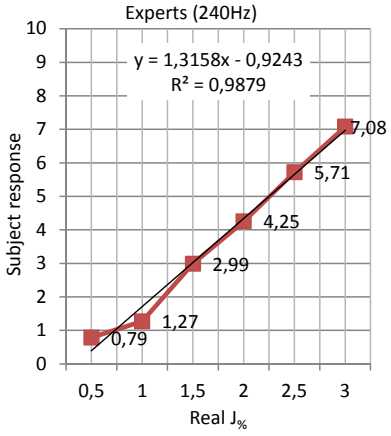
Group C: Ten casual listeners were friends and relatives of the other groups.

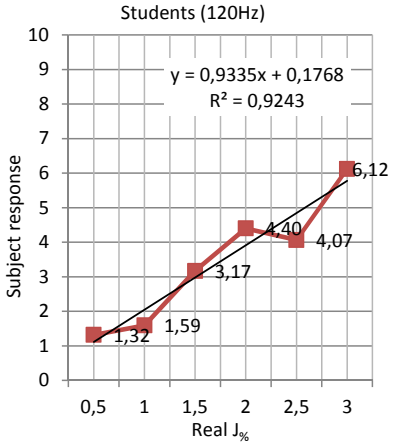
Results

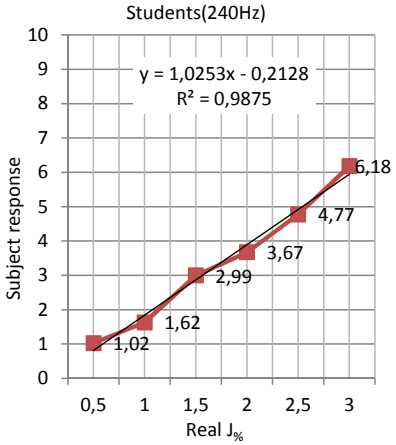
As expected most subjects assign higher perturbation values for high jitter stimuli. See Fig 2. Group A subjects present a wider sensation range (6) than subjects from Group B (5). Subjects of group C present the smallest sensation range (4). When asked for a name for the sensation they said roughness (16), distorted (9) other (5).

Average functions for each group of subjects are shown in Fig. 3.









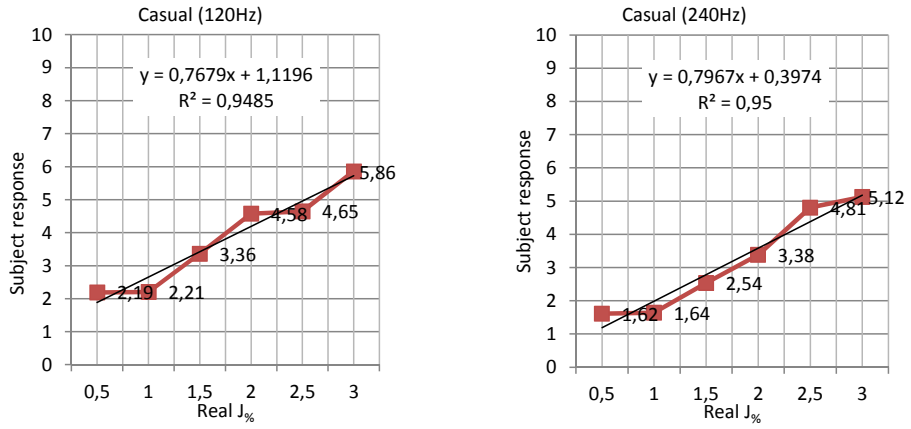


Figure 2: Average results of three groups: experts, students and casual listeners; and two stimuli signals: female (240 Hz) and male (120 Hz).

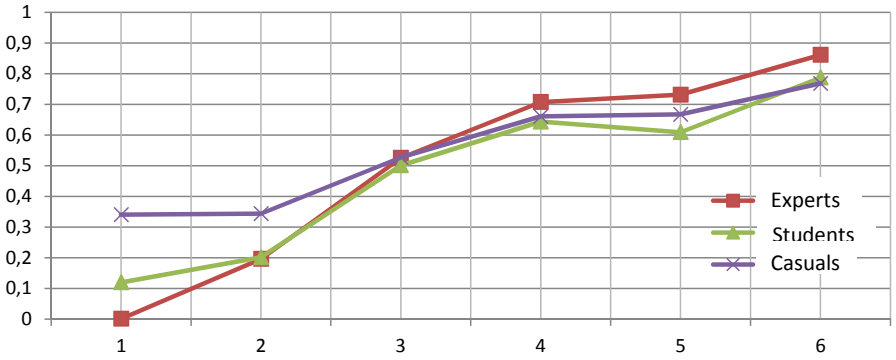


Figure 3: Average results of three groups.

Referencias

[1] Kreiman, J., Gerrat, B. R., Kempster G. B., Erman, A. and Berke, G. S.: Perceptual Evaluation of voice quality: review, Tutorial and frame work for future research. Journal of Speech and Hearing. Research, Vol. 36, pp. 21-40.1993.

[2] Rabiner, L. R., Shafer, R. W.: Digital Processing of Speech Signals, Pearson Education, 2006.

[3] Ruinskiy D. and Lavner Y.: Stochastic models of pitch jitter and amplitude shimmer for voice modification. In Proc. of 25th IEEE Convention of Electrical and Electronics Engineers in Israel (IEEEI'08), Dec. 2008.

[4] Alzamendi, G. A., Schlotthauer, G., Rufiner, H. L. and Torres, M. E.: Evaluation of a new model for vowels synthesis with perturbations in acoustic parameters. Latin American Applied Research, Volume 43, Number 3 - July 2013.

[5] Titze, I. R.: Workshop on acoustic voice analysis: summary statement. Technical report, National Center for Voice and Speech, Denver, USA. 1995.

[6] Torres, M.E., Schlotthauer, G., Rufiner, H.L. and Jackson-Menaldi, M.C.: Empirical mode decomposition. Spectral properties in normal and pathological voices. In Orc. Of IFMBE 2009, pp. 252-255. Munich, Germany, 2009.

B.5. Medición preliminar de la velocidad de conversión del sistema TTS AROMO. *Torres H.M.*

Informe de técnico: Medición preliminar de la velocidad de conversión del sistema TTS AROMO

Autores: Dr. Bioing. Humberto Torres

Lugar: Laboratorio de Investigaciones Sensoriales, INIGEM, CONICET-UBA.

Fecha: 18 de Diciembre de 2013

Declaración: Las tareas realizadas en el marco de este proyecto, así como los resultados obtenidos, son de carácter confidencial. Por lo cual, el presente informe hace un listado resumido y una breve descripción de las tareas realizadas en el período indicado. Para mayor información y/o detalle, o permisos de divulgación, contactarse con el autor.

Objetivo

Realizar una medición preliminar de la velocidad del sistema de conversión de texto en habla AROMO.

Resumen

La síntesis por selección de unidades consiste en elegir, de entre todas las secuencias posibles de unidades que sinteticen el texto de entrada, aquella que en forma conjunta se acerque más a la secuencia predicha y la que tenga menos ruidos por la concatenación. En el sistema de conversión de texto en habla AROMO, se ha implementado un algoritmo de programación dinámica que permite encontrar la mejor secuencia posible de síntesis [1]. Este proceso es el que tiene mayor costo computacional de todo el sistema: aproximadamente 95% del tiempo de proceso. Se han analizado las propuestas para aumentar la velocidad de conversión. Se implementaron una serie de mejoras, algunas de las cuales no modifican la secuencia de unidades óptimas, y otras que si lo hacen reduciendo el ancho de la búsqueda. Se ha logrado reducir el costo computacional en aproximadamente 30%. Al ajustar el ancho de búsqueda para que el sistema trabaje en tiempo real, solo se eligen 8% de las unidades distintas de las óptimas. Resta realizar pruebas perceptuales para determinar en que grado se ve afectada la calidad del habla generada al disminuir al acho de búsqueda.

Introducción

Los sistemas actuales de conversión de texto a voz utilizan como método de síntesis la concatenación de unidades acústicas previamente grabadas. Al momento de generar una frase, el sistema se encarga de buscar las unidades necesarias, las alinea y las une. Las unidades más utilizadas son los difonos, los cuales se definen como el segmento acústico entre dos puntos medios estables entre dos fonos consecutivos[2].

En los primeros sistemas de este tipo, se utilizaba una sola realización acústica de cada una de las unidades, y durante la síntesis, una vez alineadas se las posprocesaba para otorgarle las características de prosodia requeridas, utilizando métodos tales como los basados en el solapado sincrónico con la frecuencia fundamental y suma (PSOLA, del inglés (Pitch Synchronous Overlap Add Method) [3].

Si los cambios que se realizan son relativamente grandes, se genera una pérdida elevada de la calidad del sistema. Para tratar de disminuir esta degradación de la voz sintetizada se propuso tener varias realizaciones de cada una de las unidades, donde cada realización posea distintas características acústicas, de tal forma que ante un requerimiento dado, el sistema pueda seleccionar aquella unidad que más se

acerque a la unidad óptima o deseada[1]. A la búsqueda de las unidades más cercanas a las deseadas, se la conoce como Selección de Unidades [5].

Analizar todas las posibles secuencias tiene un costo extremadamente elevado, y por lo cual es común en estos casos emplear al método basado en programación dinámica [1]. Si el número de unidades en la base de datos es elevado, el costo computacional del algoritmo de búsqueda está lejos del tiempo real. Un sistema se dice que trabaja en tiempo real, si para emitir una frase de un segundo, demora en procesarla un segundo. La Relación de Tiempo Real (xRT, del inglés x Real Time) es una medida de la velocidad del sistema, definida como el cociente entre el tiempo procesamiento y la duración de la frase emitida.

Se han propuesto diferentes aproximaciones para disminuir el costo computacional: reducir el tamaño del corpus [7][8], limitar al ancho de búsqueda del algoritmo [11], precalcular costos de concatenación [8], controlar en forma dinámica la cantidad de subcostos a evaluar [10], entre otros.

En AROMO hemos implementado y adaptado algunas de estas aproximaciones. En una primera aproximación, controlamos el ancho de búsqueda, introduciéndolo como un parámetro de entrada. Además, el algoritmo de búsqueda desecha las uniones, entre la k-ésima unidad del difono objetivo i-ésimo Ui,k y la j-ésima unidad del difono objetivo i-1-ésimo Ui-1,j, si el costo acumulado mínimo actual de Ui,k es menor que el costo acumulado de Ui-1,j.

En una segunda etapa, se optimizó el algoritmo, de forma tal que cada subcosto de concatenación entre las unidades Ui-1,j y Ui,k se estiman secuencialmente, y se verifica que el costo acumulado actual sea menor que el costo acumulado mínimo para el i-ésimo difono objetivo.

Las secuencias de unidades obtenidas es la misma con la versión inicial del algoritmo de búsqueda o con la versión final. Pero el costo computacional se reduce considerablemente.

Experimentos y resultados

Se realizan una serie de experimentos para determinar la velocidad del sistema AROMO, y las mejoras obtenidas con las modificaciones introducidas. Además se analiza la incidencia del ancho de búsqueda en la xTR, y la variación en las unidades seleccionadas con respecto a la secuencia óptima. Aquí tomamos como óptima a la secuencia obtenida con un ancho de búsqueda muy elevado, el cual permite realizar la búsqueda sobre todas las posibles unidades. Hemos fijado ese valor en 999999.

Los experimentos se realizaron sobre un conjunto de 79 frases de prueba, de distintas longitudes, que anteriormente se habían utilizado para realizar una evolución perceptual de la calidad del habla sintetizada por AROMO [6]. Los valores de ancho de búsqueda analizados se fijaron por inspección.

En las siguientes secciones se presentan los resultados obtenidos en la primera y segunda etapa de optimización del algoritmo de búsqueda.

Etapa 1

En la Figura 1 se presentan las Relaciones de Tiempo Real (xRT) en el eje izquierdo, y los Porcentaje de Unidades Diferentes de la secuencia óptima (PUD), y los Porcentaje de Frases Diferentes de las secuencias óptimas (PFD) en el eje derecho, para los distintos anchos de búsquedas analizados. Para obtener una xRT de 1, para que una oración de 1 segundo se convierta en 1 segundo, el umbral debe ser fijado entre 100 y 200, aproximadamente en 160. Para este valor de ancho de búsqueda tendríamos un PUD de 27%, y un PFD de 72%.

En la Figura 2 se presentan los valores estimados para los distintos anchos de búsqueda analizados en función de la longitud de las frases de prueba. Claramente hay oraciones de longitudes que varían de 3 a 7 segundos, y un segundo grupo de oraciones de mayor longitud, de entre 13 y 18 segundos. Para todos los anchos de búsqueda analizados, la xRT no varía significativamente con la longitud de las frases.

En la `Figura 3` se presenta un diagrama de cajas de la xRT para los distintos anchos de búsqueda analizados. La dispersión de los valores estimados de xRT disminuye al decrecer el ancho de búsqueda.

[1] En inglés se utiliza el término target, que se puede traducir como objetivo. Aquí hemos rehuido a utilizar éste vocablo para evitar que se confunda con su acepción relacionada con objetivo/subjetivo.

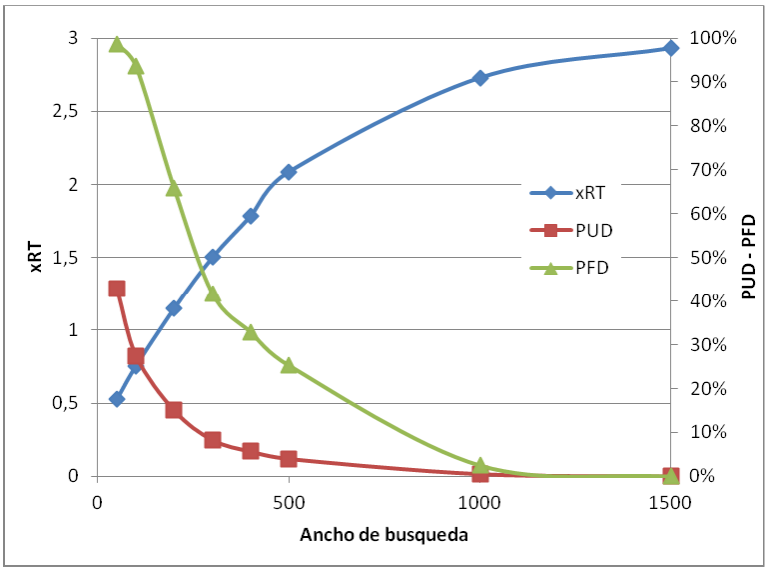


Figura 1: Relación de Tiempo Real (xRT), Porcentaje de Unidades Diferentes de la secuencia óptima (PUD), y Porcentaje de Frases Diferentes de las secuencias óptimas (PFD), para los distintos anchos de busquedas analizados, en la primera etapa de optimización.

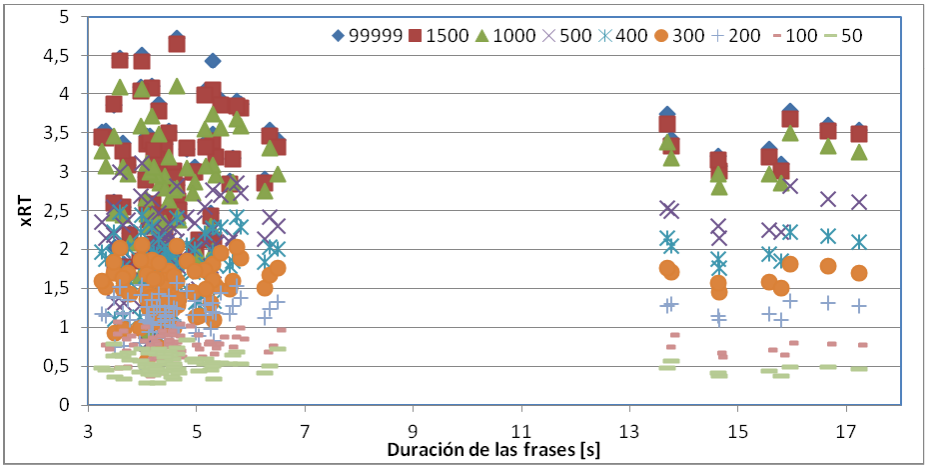


Figura 2: xRT en función de las duraciones de las frases de prueba, en la primera etapa de optimización.

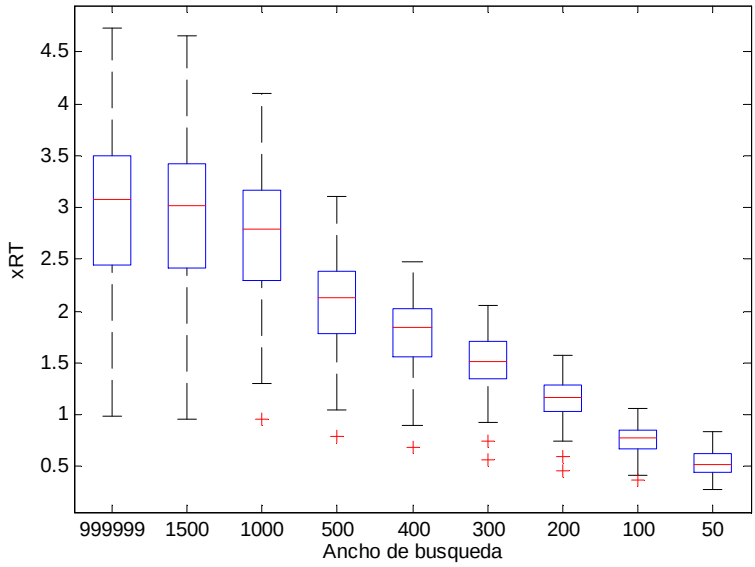


Figura 3: Diagrama de cajas de las xRT obtenidas para distintos anchos de búsqueda, en la primera etapa de optimización.

Etapa 2

En la Figura 4 se presentan las Relaciones de Tiempo Real (xRT) en el eje izquierdo, y los Porcentaje de Unidades Diferentes de la secuencia óptima (PUD), y los Porcentaje de Frases Diferentes de las secuencias óptimas (PFD) en el eje derecho, para los distintos anchos de búsquedas analizados. Para obtener una xRT de 1, el umbral debe ser fijado en 300. Para este valor de ancho de búsqueda tendríamos un PUD de 8%, y un PFD de 32%. Si comparamos estos resultados con los obtenidos en la primera etapa, logramos una reducción relativa del 70% y 55% de PUD y PFD, respectivamente. Si no acotamos el ancho de búsqueda, lo fijamos en infinito y realizamos la búsqueda sobre todas las unidades posibles, el xRT disminuye en un 30%.

En la Figura 5 se presentan los valores estimados para los distintos anchos de búsqueda analizados en función de la longitud de las frases de prueba. Claramente hay oraciones de longitudes que varían de 3 a 7 segundos, y un segundo grupo de oraciones de mayor longitud, de entre 13 y 18 segundos. Para todos los anchos de búsqueda analizados, la xRT no varía significativamente con la longitud de las frases.

En la Figura 6 se presenta un diagrama de cajas de la xRT para los distintos anchos de búsqueda analizados. La dispersión de los valores estimados de xRT disminuye al decrecer el ancho de búsqueda.

Conclusiones

Se realizó una evaluación preliminar de la Relación de Tiempo Real del sistema de conversión de texto a habla AROMO. Los resultados obtenidos indican que la xRT es independiente de la longitud de las frases a convertir. El valor del ancho de búsqueda influye directamente sobre el xRT, pudiendo fijar este valor para obtener la velocidad de conversión deseada. El ancho de búsqueda afecta la secuencia de unidades seleccionadas, y la diferencia con la secuencia óptima aumenta al disminuir el umbral. Esto no es indicador de una pérdida de calidad de conversión, dado que pruebas perceptuales preliminares han demostrado que diferentes secuencias de unidades generan habla de similar calidad. Se deben realizar más pruebas perceptuales para determinar en como se ve afectada la calidad de conversión con los cambios de unidades seleccionadas al variar el ancho de búsqueda.

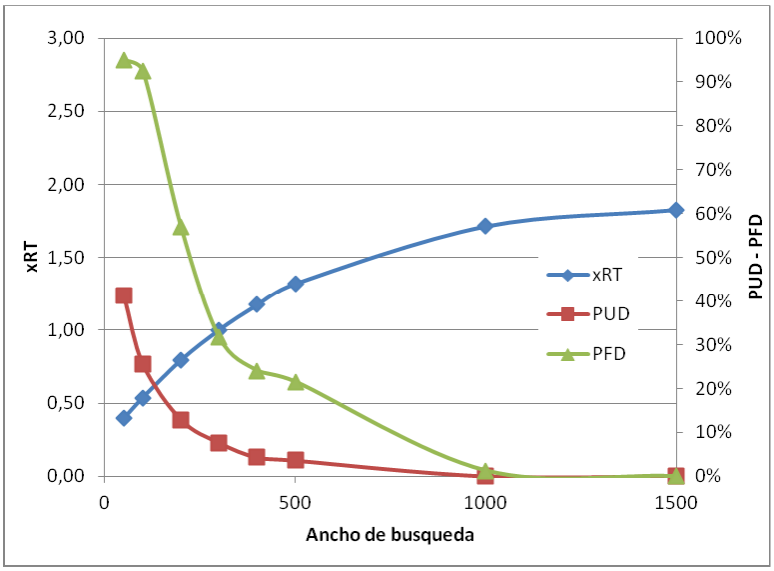


Figura 4: Relación de Tiempo Real (xRT), Porcentaje de Unidades Diferentes de la secuencia óptima (PUD), y Porcentaje de Frases Diferentes de las secuencias óptimas (PFD), para los distintos anchos de busquedas analizados.

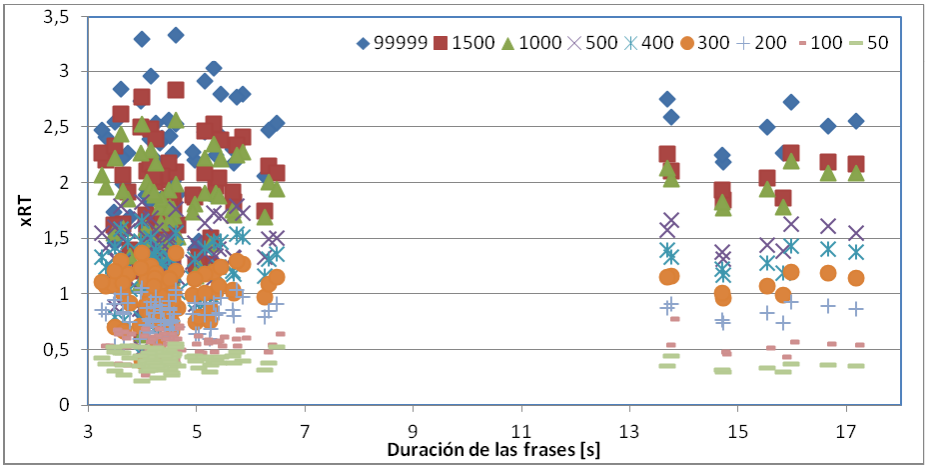


Figura 5: xRT en función de las duraciones de las frases de prueba.

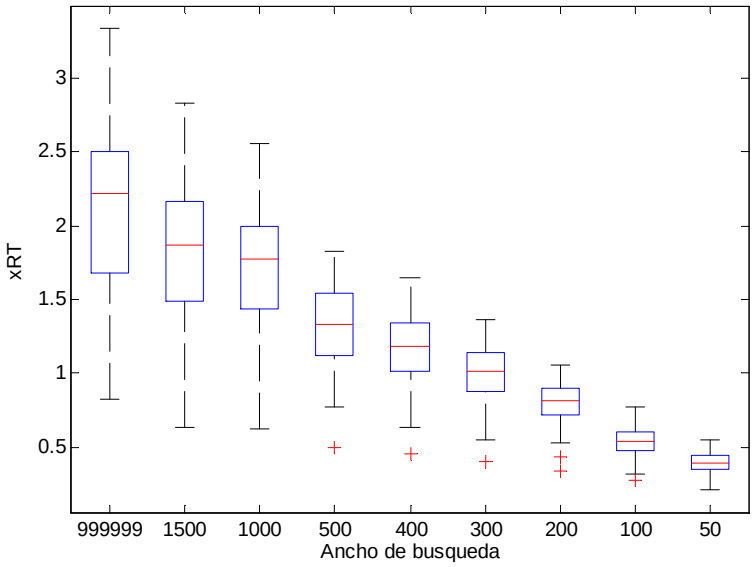


Figura 6: Diagrama de cajas de las xRT obtenidas para distintos anchos de búsqueda.

Referencias

[1] H. Torres, Informe de técnico: Selección de unidades en un sistema de conversión de texto en habla. Informe anual del Laboratorio de Investigaciones Sensoriales (ISSN 0325-2043), LIS, INIGEM, CONICET-UBA. Diciembre de 2011.

[2] Torres, H., Generación automática de la prosodia para un sistema de conversión de texto a habla, Tesis Doctoral, Facultad de Ingeniería, Universidad de Buenos Aires. Agosto de 2008.

[3] Moulines, E. y F. Charpentier: Pitch-SynchronousWaveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. Speech Communication, 9:453–467, 1990.

[4] Prudon, Romain y Christophe d'Alessandro: A selection /concatenation text to speech synthesis system: databases development, system design, comparative evaluation. En Proc. of the 4th Speech Synthesis Workshop (SSW4-2001), Pitlochvy, Scotland, August 2001. paper 138.

[5] Hunt, A.J. y A.W. Black: Unit selection in a concatenative speech synthesis system using a large speech database. En Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96), Vol. 1, pp. 373–376, Atlanta, Georgia, May 1996.

[6] J. Gurlekian, C. Cossio Mercado, H. Torres, and M. Vaccari: Subjective Evaluation of a High Quality Text-to-Speech System for Argentine Spanish. Proc. of. VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, IberSPEECH 2012, pp. 241-250. Madrid, Spain, 21-23 November 2012.

[7] Rohit Kumar, S. P. Kishore: Automatic Pruning of Unit Selection Speech Databases for Synthesis without Loss of Naturalness, International Conference on Spoken Language Processing (Interspeech - ICSLP), October 2004, Jeju Korea.

[8] Isogai, Mitsuaki and Mizuno, Hideyuki: Speech database reduction method for corpus-based TTS system, In INTERSPEECH-2010, pp.158-161.

[9] Jerneja Zganec Gros and Mario Zganec: An Efficient Unit-selection Method for Concatenative Text-to-speech Synthesis Systems, Journal of Computing and Information Technology - CIT 16, 1, pp. 69–78. 2008.

[10] S. Sakai, T. Kawahara and S. Nakamura: Admissible stopping in viterbi beam search for unit selection in concatenative speech synthesis. In Proc. of ICASSP 2008, pp. 4613-4616. Las Vegas, Nevada, USA. March, 2008.

[11] Daniel Tihelka, Jirí Kala, Jindrich Matousek: Enhancements of viterbi search for fast unit selection synthesis. In Takao Kobayashi, Keikichi Hirose, Satoshi Nakamura, Eds. Proc. of INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, pp. 174-177 Chiba, Japan. September, 2010.