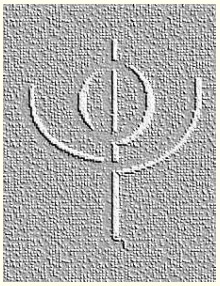
ISSN: 0325-2043



Laboratorio de Investigaciones Sensoriales (LIS)

Informe XLIX–2016

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

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 postdoctorales

- DE MIER Mariela Vanesa, Lic y Dra. en Letras, Becaria CONICET.

Becario y Tesistas Doctorales

- COSSIO MERCADO Christian, Ing. en Informática, Becario CONICET. Tesista de Doctorado UBA.
- MARTINEZ SOLER Miguel, Ing. en Informática, Tesista de Doctorado UBA.
- UNIVASO Pedro, Ing. Electrónico. Tesista de Doctorado UBA.
- ARANCIBIA ARANGIO Gianfranco. Tesista de Doctorado UBA.

Secretaria: SOLINI Gabriela, CONICET.

[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. Identificación Forense de Hablantes

Director: Jorge A Gurlekian, Tesistas: Miguel Martínez Soler, Pedro Univaso.

Los intercambios pasados con Gendarmería Nacional y el Ministerio Público de la Provincia de Buenos Aires permitieron comprender los requerimientos y enfoque que dan estas instituciones al tema de identificación forense. Hacia fines del 2015 la Dirección de Vinculación Tecnológica del CONICET nos invitó a participar en el comité científico para la preparación de un congreso internacional sobre Ciencia y Justicia que se desarrollará el próximo año, en virtud de la experiencia obtenida en los temas de identificación de hablantes para uso forense. La misma Dirección de Vinculación está promoviendo la formación de recursos en el área de las Facultades de Derecho de la provincia de Buenos Aires. En este período se continúa con el desarrollo de dos tesis de doctorado que emplean métodos convencionales y técnicas del estado del arte como los I-vectors para la identificación de hablantes empleando bases de datos propias y bases provistas por NIST.

3.1.1. Trabajos terminados

Data Mining applied to Forensic Speaker Identification
Univaso P., Ale J. (UCA), Gurlekian JA
Resumen:
In this paper we analyze the advantages of using data mining techniques and tools for data fusion in forensic speaker recognition. Segmental and suprasegmental features were employed in 28 different classifiers, in order to compare their performances. The selected classifiers have different learning techniques: lazy orinsta nce-based, eager and ensemble. Two approaches were employed on the classification task: the use of all features and the use of a feature subset, selected with a gainratio methodology. The best performances, with all features, were obtained by three classifiers: Logistic Model Tree (eager), LogitBoost (ensemble) and Multilayer Perceptron (eager). SupportVector Machine (eager) proved to be a good classifier if a Pearson VII function-based universal kernel was used. When low dimensional features were selected, ensemble classifiers exceededthe performance of all others classifiers. Segmental and tone features demonstrated the best speaker discrimination capabilities, followed by duration and quality voice features. Evaluation was performed on Argentine-Spanish voice samples from the Speech_Dat database recorded on a fixed telephone environment. Different recording sessions and channels for the test segmentswere added and the Z-norm procedure was applied for channel compensation.

3.2. UBACyT CM12, "Pruebas clínicas de análisis de la voz y el habla. Mejoras en la evaluación Audio-Perceptual Evaluación objetiva de la Prosodia"

Director: María E. Vaccari (UBA-Fonoaudiología), Co-director: Jorge A Gurlekian, Investigadores: Humberto Torres, Diego Evin, Liliana Sigal.
Período: 2011–2014.

3.2.1. Trabajos terminados

Comparison of two perceptual methods for evaluation of F0 perturbation
Gurlekian J.A.; Torres H.M.; Vaccari M.E.

Abstract:
Objectives. To explore perceptual evaluation of jitter produced by fundamental frequency (F0) variation in a sustained vowel /a/, using two different methods. One is based on listeners internal references and the other is based on external references provided by the experimenter. Methods. We used two methods: one is magnitude estimation-converging limits (ME-CL), which is close to the standard approach used by speech therapists when they use numerical estimations and their own standards, and other is intramodal matching procedure (IMP), where each matched stimulus is to be compared with a fixed-set matching stimuli. Systematic variations were introduced in vowel /a/ by Linear Prediction Coding synthesis using an F0 contour function obtained from a statistical jitter model. Six jitter values were used for each of two reference F0 values.
Three groups of listeners were tested: expert speech therapists, speech therapy students, and naive listeners. Results. Perceptual functions appear to be similar and linear for both methods as the theory predicts. The answers of all groups of listeners tested with ME-CL present higher standard deviations than for IMP. When subjects were tested with IMP, intrareliability and interreliability measurements show a significant improvement for both expert and naive listeners.
Conclusions. Both intraindividual and interindividual differences for expert speech therapists could be better managed when tested with an IMP than when they use numerical estimations and internal standards to evaluate vowel perturbation produced by jitter. This procedure could be the basis for the development of a clinical evaluation tool.

3.3. Desarrollo de un sistema de conversión de texto en habla para su aplicación en sistemas de telecomunicaciones

Director: Jorge Gurlekian, Co-director: Humberto Torres, Investigadores: Diego Evin, Agustín Gravano, Christian Cossio-Mercado.

Se continúa con el desarrollo iniciado con el proyecto PAE-PID 094 Se trabajó en la incorporación de mayor cantidad de datos de la locutora seleccionada. Se trabajó en la mejora de la calidad a nivel de la fusión de palabras. Los aspectos prosódicos fueron investigados en mayor detalle. Este conocimiento se volcó al conversor de texto a habla lográndose una entonación de muy alta calidad prosódica. Todas estas mejoras redundaron en un alto nivel de naturalidad como se puede apreciar en la demostración disponible en:
URL: http://181.28.244.40:8090/LISTools/

Se inicio la transferencia de tecnología del sistema TTS a la empresa BDT Group mediante un STAN. La empresa ha firmado en la actualidad un convenio de confidencialidad. También se presentó el sistema TTS a una segunda empresa Global News.

3.3.1. Trabajos terminados

Novel Estimation Method for Superpositional Intonation Model: From Text to Form
Torres, H.M. & Gurlekian J.A.
Abstract:
Fujisaki's intonation model parameterizes the F0 s contour efficiently and giving its strong physiological basis has been successfully tested in different languages. One problem thathas not been fully addressed is the extraction of the models parameters, i.e., given a sentence, which of the model parameter values best describe its intonation. Most of the proposed methodsstrive to optimize the parameters so as to obtain the best fit for the F0 contour globally. In this paper we propose to use text information from the sentence as the main guide or reference foradjusting the parameters. We present a method that defines a set of rules to fix and optimize the model?s parameters. Optimization never loses sight of the events of the text structure that arouse it. When text information is not enough, the algorithm predicts parameters from F0 contour and tie it to the text. The process of parameter estimation can be seen as a way to go from textinformation to the F0 contour. The parameter optimization is carried out to fit the F0 contour locally. Our novel approach can be implemented manually or automatically. We present examplesof manual implementation and the quantitative results of the automatic one. Tested on three corpora in Spanish, English and German, our automatic method show an performance of 34 % better than other tested methods.

Estudio del Foco
Gurlekian J.A., Torres H.M., Mixdorff H. (Beuth Hochschule für Technik Berlin); Cossio-Mercado C., Güemes M.
Resumen:
Se pretende establecer una conexión entre el significado lingüístico de una emisión indicado por la modalidad y el foco con los múltiples rasgos posódicos subyacentes. De un conjunto de oraciones estructuradas estudiadas en el proyecto AMPER para el Español de Buenos Aires se analizan los datos acústicos de los parámetros tradicionales: energía, frecuencia fundamental, duración y estructura armónica y de los parámetros derivados de los comandos de acento tonal obtenidos con el modelo de entonación de Fujisaki.
Esta información acústica se correlaciona con la evaluación del foco y de las prominencias percibidas y se ensaya la detección automática de prominencia a partir de la combinación de las mediciones obtenidas.
Los resultados verifican una relación entre la máxima prominencia percibida en el objeto con las distintas condiciones de foco. Así mismo se concluye que los parámetros acústicos medidos en la sílaba como los derivados del modelo de entonación contribuyen en forma integrada a la determinación de las prominencias de los casos con un 90.78 % de certeza, utilizando el método de regresión logística. Los parámetros acústicos mejor correlacionados con la prominencia percibida son la duración de la sílaba y la amplitud del comando de acento tonal.

Acoustic Correlates of Perceived Syllable Prominence in German
Mixdorff H., Cossio-Mercado C., Hönemann A., Gurlekian J.A., Evin D., Torres H.M., Interspeech 2015, Dresden, Alemania.
URL: http://www.isca-speech.org/archive/interspeech_2015/i15_0051.html
Abstract:

This paper explores the relationship between perceived syllable prominence and the acoustic properties of a speech utterance. It is aimed at establishing a link between the linguistic meaning of an utterance in terms of sentence modality and focus and its underlying prosodic features. Applications of such knowledge can be found in computer-based pronunciation training as well as general automatic speech recognition and understanding. Our acoustic analysis confirms earlier results in that focus and sentence mode modify the fundamental frequency contour, syllabic durations and intensity. However, we could not find consistent differences between utterances produced with noncontrastive and contrastive focus, respectively. Only one third of utterances with broad focus were identified as such. Ratings of syllable prominence are strongly correlated with the amplitude of underlying accent commands, syllable duration, maximum intensity and mean harmonics-to-noise ratio.

Estimación de las funciones de costo para la selección de unidades en AROMO
Humberto Torres
Resumen:
El objetivo es estimar en forma automática las funciones de costos de selección de unidades para la voz Emilia en el sistema de conversión de texto en habla Aromo.
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 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 un informe anterior presentamos los resultados de la evaluación de distintos atributos de las unidades para la construcción de los costos. También presentamos una propuesta para de la definición de las funciones de costo. En este informe presentaremos una nueva propuesta de estimación de las funciones de costo relacionadas con los atributos de frecuencia fundamental, energía y contenido espectral. Los resultados son obtenidos corresponden a la estimación de las funciones de costo sobre la versión final del corpus Emilia.

Distribución de ocurrencias de fonemas y difonemas
Humberto Torres, Diego Evin y Jorge Gurlekian
Resumen:
Los objetivos son 1) Crear un corpus de texto de transcripciones fonéticas, 2) Analizar la distribución de los fonemas en el habla leída, 3) Analizar la distribución de los difonemas en el habla leída y 4) Estimar la cobertura del corpus Emilia.
La frecuencia con que se repiten los sonidos del habla es de interés teórico y aplicado. Desde el punto de vista teórico, la distribución de las unidades caracteriza a la lengua. Desde el punto de vista de las aplicaciones, esta información es apreciada en el campo de la percepción del habla, reconocimiento automático del habla, síntesis de habla, patologías del habla, identificación del locutor, psicoacústica, procesos cognitivos, neurolingüística, entre otros. Existes varios trabajos que reportan distribuciones de ocurrencias de fonemas para las distintas variantes del español. En particular para el español hablado en argentina, nos remitiremos al trabajo pionero de Guirao y Borzone de 1972, y de Guirao y García que realizaron el conteo

de fonemas sobre el texto de cinco libros de cuentos y novelas. Los tamaños de los córpora utilizados hay ido evolucionando: desde 5.000 fonemas hasta 3.650.000 fonemas.
El corpus Emilia fue diseñado para la creación de una voz para el sistema Aromo de conversión de texto en habla (TTS, del inglés Text-To-Speech). En este contexto, se define la cobertura de un corpus como la capacidad de sintetizar una palabra cualquiera del idioma. Aromo realiza la síntesis del habla mediante el método de selección de unidades previamente guardadas. La unidad empleada es el difono, que se define como el segmento de habla que va desde el punto medio estable de un fono al punto medio estable del siguiente fono. Una forma de estimar si el sistema puede o no sintetizar una palabra es comparando las unidades presentes en el corpus que utiliza el sistema TTS con las unidades necesarias para realizar la tarea de síntesis.

Corpus Emilia
Resumen:
El corpus de texto Emilia fue diseñado para la tarea de conversión de texto en habla. Sus oraciones fueron creadas para contener todas las sílabas del español, en sus formas acentuadas y no acentuadas, su distribución de fonemas sigue la distribución natural de la lengua, con un número de instancias de difónos mínimo de 10. El corpus contiene 2.218 oraciones, con 29.045 palabras, de las cuales 8.919 son distintas. El corpus fue grabado por una locutora profesional para crear la versión oral. El audio fue etiquetado manualmente en varios niveles: palabras, sílabas, fonético con el sistema SAMPA (Speech Assessment Methods: Phonetic Alphabet), en clases de palabras, entre otros. El corpus contiene 141.488 fonemas y 147.750 difonemas. Existen diferencias entre lo que se esperaba que emitiera la locutora y lo que finalmente pronunció. Esto se debe a tres factores: la locutora inserta pausas no marcadas que luego rompen las reglas de producción de alófonos; en algunas ocasiones la locutora cambia algunas reglas de producción de alófonos; en nombre propios de origen extranjero introduce nuevas reglas propias de transcripción de grafemas a fonemas. Así, se puede observar las diferencias entre el etiquetado fonético teórico dado por las reglas de transcripción y lo producido por la locutora.

Corpus de texto Lana
Resumen:
El corpus de referencia Lana se construyó empleando todos los artículos publicados en la edición digital del diario La Nación de la ciudad Buenos Aires entre los años 1996 y 2006. Para obtener el texto de esos artículos se desarrolló un web crawler ad-hoc. Este programa se encarga de obtener de manera metódica las páginas web correspondientes a cada artículo, gestionando la transacción con el servidor de forma tal que el proceso de obtención de datos no perjudique el desempeño del sitio, y de filtrar el contenido textual correspondiente a cada artículo.
Como se mencionó, el propósito de este corpus es permitir un análisis estadístico a nivel grafémico e indirectamente fonético del lenguaje escrito del español de Buenos Aires. Sin embargo, las notas presentan naturalmente diversos símbolos que son propios del lenguaje escrito pero que no se corresponden a palabras. Esto implica la necesidad de llevar todos esos símbolos a sus versiones grafémicas en un proceso denominado normalización o canonización. El módulo de normalización de este material se desarrolló mediante un conjunto de scripts basados en reglas y expresiones regulares que permiten entre otras operaciones convertir símbolos numéricos a palabras, expandir abreviaturas y acrónimos, o eliminar símbolos especiales.
Por otro lado, se decidió eliminar del conjunto final de datos los artículos de la sección política del periódico. Esto se debe a que esa sección involucra muchos personajes que coyuntural-

mente aparecen mencionados muy frecuentemente, por ejemplo, los miembros del gobierno de turno o de la oposición, lo que para un estudio estadístico del lenguaje suponer un artefacto. La versión final del corpus consiste en 55.908.107 palabras, de las cuales 288.779 son distintas.

3.4. El rol de la prosodia y la fluidez lectora: relación entre fonología suprasegmental, reconocimiento de palabras y conexiones causales durante la comprensión del discurso

Director: Vanesa De Mier

Resumen: El objetivo general del presente proyecto es investigar el interjuego entre la producción y la percepción de rasgos prosódicos en el habla y la comprensión del discurso. Específicamente, se intentará clarificar si los rasgos prosódicos y la puntuación del discurso contribuyen a promover la comprensión del discurso oral y escrito en niños de 9 años edad (cuando se espera que comiencen a *leer para aprender*). Con esta finalidad, los objetivos específicos del presente proyecto serán: Examinar las relaciones entre los procesos fonológicos suprasegmentales a nivel léxico y discursivo (acento, entonación, juntura y ritmo) que están involucrados la lectura. Estudiar el rol de la presentación de discurso narrativo en una condición de prosodia normal con foco o una condición de prosodia alterada en la comprensión. Analizar la prosodia productiva en la lectura en voz alta a partir de la puntuación y su incidencia en la comprensión lectora de textos con puntuación alterada y normal con foco. Indagar el interjuego entre la condición de presentación de la prosodia del enunciado (normal con foco o alterada) y su cantidad de conexiones causales.

3.5. Reconocimiento Automático del Habla para el Español de Argentina

Director: Diego A. Evin

Resumen:
El Reconocimiento Automático del Habla (ASR) es el proceso por el cual se convierte la señal acústica de habla en texto. Este proyecto está dedicado al estudio y desarrollo de métodos y sistemas para resolver el problema de ASR para el español hablado en Argentina. Durante el período informado por un lado se trabajó en el problema de reconocimiento de habla proveniente de medios de radiodifusión de la ciudad de Buenos Aires, y dentro de este problema también se trató el subproblema de identificación de palabras clave. Por el otro lado, se trabajó en el estudio y desarrollo de reconocedores de habla embebidos en dispositivos de procesamiento autónomos y portables.

3.5.1. Trabajos terminados

- D. A. Evin, H. M. Torres, J. A. Gurlekian. *Implementación de un Sistema de Keyword Spotting*. Informe Técnico del LIS - INIGEM, pp. 1–13, abril de 2016.
- M. Marufo da Silva, D. Evin, S. Verrastro. “Speaker-Independent Embedded Speech Recognition using Hidden Markov Models”. Congreso Argentino de Ciencias de la Informática y Desarrollos de Investigación, Buenos Aires, Noviembre de 2016.

- Alvarez, D. Evin, S. Verrastro. "Implementation of a Speech Recognition System in a DSC". IEEE Latin America Transactions, Vol. 14, No. 6, pp. 2657-2662, Junio 2016.

3.6. Un rango efectivo de viscosidad

En este trabajo se intenta determinar el rango de viscosidad física que se correlaciona en forma efectiva y sin ambigüedades con la viscosidad percibida. La búsqueda de un rango efectivo conduce a dos interrogantes. Si los valores bajos de la escala física se correlacionan más bien con la fluidez y los altos más bien con la consistencia (entendida como la cualidad de la materia que resiste sin romperse ni deformarse fácilmente) que con la viscosidad. Se utilizarán catorce muestras de siliconas (silicone oils newtonian fluids from Brookfield Eng. Lab., USA) en un rango de valores de 2 a 1 0000 000 centipoises (cps). Los panelistas recibirán las muestras en tubos y las evaluarán observando el movimiento de las soluciones En una sección preliminar los panelistas darán juicios cualitativos. Cada uno dividirá las muestras en grupos de acuerdo a las cualidades (gradientes de viscosidad) que percibe. En las secciones siguientes darán juicios cuantitativos. Se realizarán experimentos en los que se aplicarán diferentes métodos psicofísicos y se observará si las funciones perceptuales que se obtengan se diferencian de acuerdo a los subrangos de la viscosidad física.

3.7. 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

Resumen:
En trabajos anteriores se ha observado que el etanol modifica el dulce de la sacarosa y el agrio del ácido cítrico y que el efecto es en algún modo diferente para uno y otro gusto. En este trabajo se examinan esas posibles diferencias teniendo en cuenta no solo los efectos en la intensidad sino también en la cualidad y en la persistencia. Con respecto a la intensidad el etanol aumenta el dulce de la sacarosa, pero en el ácido cítrico tiene un efecto doble y opuesto: en las concentraciones bajas aumenta el agrio y en las altas el agrio predomina sobre el sabor del etanol.
El etanol aumenta también la duración del dulce y no hay diferencia entre los dos niveles de etanol. En cambio, la duración del agrio depende de las concentraciones y del porcentaje de etanol.
En cuanto a la cualidad el gusto de la sacarosa aún en concentraciones altas, es menos vulnerable al cambio, tiende a mantenerse dulce y mezclada se percibe como un sabor dulce con una nota alcohólica. En cambio, el ácido cambia de modalidad pasa de agrio a trigeminal. Es posible que el ácido cítrico se integre en un nuevo compuesto en el que la nota trigeminal predomine sobre el sabor alcohólico.

3.7.1. Trabajos terminados

Bitter taste modifications by mixing caffeine with ethanol
Guirao M., Calviño A., and Evin D.
Abstract:
Background: The majority of pychophysical research on bitter taste ethanol mixture have

focused on complex models or on a single compound like quinine. Since different bitter compounds give different perceptual responses previous findings may not generalize to ethanol caffeine mixture. Important as it is no much data on the interaction between these two chemicals is available.
Objective: This study examine the potential changes in intensity and duration of aftertaste produced in the taste of caffeine when mixed with ethanol.
Methods: The psychophysical methods of Pair Comparison and Magnitude Estimation were applied to quantify the bitterness intensity of seven concentrations of caffeine (3, 6, 12, 25, 35, 55 and 80 mM) in water solutions tasted alone and mixed with two (8 % and 15 %) ethanol levels in water. The Time-Intensity Method was used to asses intensity and aftertaste of two caffeine (6 and 55 mM) solutions unmixed and mixed with the same ethanol levels.
Results: When taste intensity was rated both levels of ethanol increased bitterness at the lower half part of the caffeine range. In the upper part (up to 35 mM) the bitter taste intensity remained unchanged. When assessing intensity/time responses the mixture gave longer aftertaste than caffeine alone. The difference between the effect of the two ethanol levels is clear in the weak (6 mM) but less pronounced in the stronger (55 mM) caffeine solution.
Conclusions: The addition of ethanol to caffeine enhanced the duration of bitter aftertaste. Taste intensity was increased at weak and moderate concentrations but higher concentrations were unaffected.

4. Docencia

4.1. Cursos de grado

Dr. Humberto M. Torres

Profesor Adjunto de la cátedra *Señales y Sistemas*, Facultad de Ingeniería UBA. Desde el 26 de Septiembre de 2011.

Dr. Jorge A Gurlekian

Dictado del Seminario *Laboratorio de Voz*. Facultad de Medicina. Área Fonoaudiología. Desde el 2001.

Dr. Diego Evin

Docente Auxiliar de 1ra Categoría de la cátedra *Inteligencia Artificial* y de la cátedra *Inteligencia Computacional*. Departamento de Matemática e Informática, Facultad de Ingeniería, Universidad Nacional de Entre Ríos. Desde 2003.

Profesor invitado como Experto Externo en la materia *Procesamiento de Señales Biomédicas*, Instituto Tecnológico de Buenos Aires (ITBA). Octubre-noviembre 2016.

Ing. Christian Cossio Mercado

Jefe de Trabajos Prácticos (JTP) en las materias *Paradigmas de Lenguajes de Programación* y *Teoría de Lenguajes*, Departamento de Computación, Facultad de Ciencias Exactas y Naturales, UBA. Desde marzo de 2015.

4.2. Otros cursos

Dra. Miguelina Guirao

Clase para la carrera de Especialistas en ORL Facultad de Medicina, UBA.
Tema: Interacciones del gusto con otras modalidades sensoriales
En la FASO, 28 de Abril de 2016.

Clase Facultad de Medicina, UBA.
Tema: El gusto: interacciones con otros sistemas sensoriales
En la Asociación Médica Argentina, 30 de Junio de 2016.

5. Intercambio Científico

5.1. MINCYT-COLCIENCIAS

El Dr. Jorge A. Gurlekian realizó el intercambio previsto en el proyecto Bilateral Colciencias CONICET MINCYT en la Universidad Manuela Beltrán de Bucaramanga, Colombia, con el dictado de cursos y conferencias en el tema de la evaluación objetiva del riesgo vocal.

Se recibió la visita de la Fga Melissa Rincon Cediel de la Universidad Manuela Beltran en relación al proyecto bilateral antes mencionado.

5.2. Otra visitas

Se recibió la visita del Prof. Dr. Masuzo Yanagida de la Facultad de Ingeniería, Universidad de Doshisha, Kyo-Tanabe, Japón. Setiembre 2016.

6. Premios y reconocimientos

6.1. Premio Sadosky 2016

El grupo de audición y habla del LIS, junto a GlobalNews Group, resultó finalista del Premio Sadosky 2016 en la categoría Innovación, por el proyecto de reconocimiento automático del habla para monitoreo de medios de comunicación, denominado *Global Voices*.

7. Tesis

7.1. Doctorales finalizadas

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.
Codirector: Dr. Carlos Kukso

Consejero : Prof. Dr. Federico Micheli
Calificación: Sobresaliente

7.2. Doctorales en curso

Evaluación Automática de Calidad del Habla Artificial

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

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

Reconocimiento automático de hablantes empleando información de largo plazo

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

8. Actividades de Divulgación

- Noticias MINCYT Agosto 30, 2016. "Tecnología argentina para el reconocimiento de voz"
  Se refiere a los trabajos que realizan los Dres. Jorge A. Gurlekian. Humberto M. Torres y Diego Evin sobre los modos de comunicación interpersonal para realizar simulaciones en máquinas que hablan y reconocen palabras.
  http://www.mincyt.gob.ar/noticias/tecnologia-argentina-para-el-reconocimiento-de-voz-12069.

- Dra. Miguelina Guirao. Participación en la nota *Los sentidos y la percepción de los colores durante el acto del juego de Antonio Auriti*, Primavera 2016.
  http://www.juegoyreeducacion.com.ar/34.htm.

- Ing. Christian Cossio-Mercado. *Todo con afecto (¡y también la Computación!)*, Programa Exactas va a la Escuela 2016, FCEyN, UBA. Lugar: Diferentes colegios secundarios de CABA y Conurbano bonaerense. Junio, septiembre y noviembre 2016. Dictado de charla sobre Computación Afectiva, área que integra registros del cuerpo humano e información conductual para reconocer emociones y sentimientos, y generar respuestas automáticas que también las denoten.

9. Trabajos que refieren a actividades del LIS

9.1. Publicaciones en revistas

- Piñeda, M.A., y Scherman, P., "S. S. Stevens, M. Guirao y los estudios psicofísicos en Argentina". En *Revista Mexicana de Análisis de la Conducta*, No. 2, Vol. 42, pp. 153–178. Ciudad de México, 2016. http://dx.doi.org/10.5514/rmac.v42.i2.57025.

10. Participación en Congresos

10.1. Simposio Nacional Sobre Ciencia y Justicia

Hacia fines del 2015 la Dirección de Vinculación Tecnológica del Conicet el Dr. Gurlekian participo en el comité científico para la preparación de un simposio nacional sobre Ciencia y Justicia que se desarrollo en el 2016 en virtud de la experiencia obtenida en los temas de identificación de hablantes para uso forense. La misma dirección de Vinculación está promoviendo la formación de recursos en el área de las Facultades de Derecho de la provincia de Buenos Aires.

El Simposio se realizo el 6 y 7 de diciembre de 2016 en el Centro Cultural de la Ciencia, en el Polo Científico Tecnológico de la Ciudad Autónoma de Buenos Aires. La apertura estuvo a cargo del Ministro de Ciencia, Tecnología e Innovación Productiva de la Nación, Dr. Lino Barañao, el Presidente del CONICET Dr. Alejandro Ceccatto y la Defensora General de la Nación, Dra. Stella Maris Martínez.

Las bases de datos constituyen una de las herramientas tecnológicas más ampliamente extendidas en todos los campos de las actividades humanas. Su configuración exhibe características comunes, no obstante los criterios de seguridad varían en función de la confidencialidad de la información almacenada. Su utilización impacta en la investigación judicial.

El Dr. Jorge Gurlekian, CIC-CONICET, expuso sobre el tema *La identificación forense de voces sustentada en bases de datos locales, regionales e internacionales.*

11. Publicaciones

11.1. Capítulos de libros

- Gurlekian, J., Torres, H., Evin, D., Mixdorff, H., Cossio-Mercado, C., y Güemes, M., "Estudio del Foco: las Prominencias Acentuales, Modelado Acústico y la Detección Automática". Capítulo del libro "Reflexiones sobre Aspectos de la Fonética y otros Temas de Lingüística", Vol 53, pp. 209–219. Barcelona, 2016. ISBN: 978-84-608-9830-6. http://stel.ub.edu/labfon/amper/homenaje-eugenio-martinez-celdran/53reflexiones/53reflexiones.pdf

11.2. Revistas

- Gurlekian, J., Torres H., and Vaccari, M. "Comparison of two methods for perceptual evaluation of F0 perturbation". En *Journal of Voice*, Volume 30, Issue 4, pp. 506.e1–506.e8. July 2016. Doi: http://dx.doi.org/10.1016/j.jvoice.2015.05.009.

- Torres H., Gurlekian, J., "Novel Estimation Method for Superpositional Intonation Model". En *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 24, Issue

1, pp. 151–160. January 2016. http://doi.org/10.1109/TASLP.2015.2500728.

- Evin, D., Hadad, A., Solano, A., Drozdowicz, B., "Segmentation Fusion Techniques with Application to Plenoptic Images: A Survey". En *SABI 2015 IOP Publishing Journal of Physics*, Conference Series 705, 2016. http://iopscience.iop.org/article/10.1088/1742-6596/705/1/012026/pdf

- Alvarez A., Evin D., Verrastro S., "Implementation of a Speech Recognition System in a DSC". En *IEEE Latin America Transactions*, Volume 14, Issue 6, June 2016. pp. 2657–2662. http://ieeexplore.ieee.org/document/7555234/

- Güemes, M., Sampedro, B., Cossio-Mercado, C. y Gurlekian J. "La relación entre foco y prosodia: Análisis de la percepción de las prominencias acentuales en un corpus del español". En *Estudios de Lingüística Universidad de Alicante (ELUA)*, pp. 129–139, Universidad de Alicante. ISSN: 0212-7636.

- Guirao, M., y Evin, D.A., "Efecto del etanol en el gusto de la sacarosa y del ácido cítrico". En *La Alimentación Latinoamericana*, No. 321, 2016, pp. 56–51. http://alaccta.org/revista-la-alimentacion-latinoamericana-ed-321/.

11.3. Congresos

- Marufo da Silva, M., Evin, D.A., and Verrastro, S., "Speaker-Independent Embedded Speech Recognition using Hidden Markov Models". En *Proceedings of IEEE CACIDI 2016 - IEEE Conference on Computer Sciences*, Buenos Aires, Noviembre de 2016. https://doi.org/10.1109/CACIDI.2016.7785985.

- Gurlekian, J.A., Mixdorff, H., Torres, H.M., Cossio-Mercado, C., and Evin, D., "Acoustic Correlates of Perceived Syllable Prominence in Argentine Spanish". En *Proceedings of Speech Prosody 2016*, pp. 673–677. Boston, MA, USA, May 31-June 3, 2016. http://www.isca-speech.org/archive/SpeechProsody_2016/pdfs/275.pdf.

11.4. Informes Técnicos

- Evin, D.A., Torres, H.M., Gurlekian, J.A., *Informe Técnico: Implementación de un Sistema de Key Word Spotting*. Laboratorio de Investigaciones Sensoriales, LIS, INIGEM, CONICET UBA. Destinatario: Global News. Abril 2016.

- Torres, H.M., *Informe de técnico: Medición de la velocidad de conversión del sistema TTS AROMO*. Laboratorio de Investigaciones Sensoriales, LIS, INIGEM, CONICET UBA. Destinatario: BDT Group. Abril 2016.

Apéndice

A. Informes Técnicos

A continuación se incluye el texto completo de los informes técnicos realizados para el año 2016.

A.1. Medición de la velocidad de conversión del sistema TTS AROMO. *Torres H.M.*

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

Autor: Dr. Bioing. Humberto Torres

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

Fecha: 28 de Abril de 2016. Primera versión: 18 de Diciembre de 2013.

Destinatario: BDT Group.

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 de la velocidad del sistema de conversión de texto en habla AROMO.

Cuantificar las diferencias en la calidad de habla sintetizada al aplicar un sistema de poda en la selección de unidades.

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. 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.

Las pruebas se realizaron en una computadora de escritorio, con un microprocesador Intel(R) Core(TM) i7-2600k CPU @ 3.40 GHz y 8 GB de memoria RAM, con un sistema operativo Windows 7, Service Pack 1.

En la Figura 1 se presentan las Relaciones de Tiempo Real (xRT) para los distintos anchos de búsquedas analizados. Las barras verticales indican el desvío estándar de la muestra. Se puede observar que en todos los casos la xRT es inferior a la unidad, siendo de 0.84 para el caso de realizar la búsqueda sin poda. Es notable como aumenta la dispersión de los valores de xRT al aumentar el umbral. Esto se puede explicar teniendo en cuenta que el xRT también se ve afectado por las longitudes de las frases a sintetizar: para oraciones largas el xRT es más estable, e inferior, que en el caso de oración largas. Este efecto es más pronunciado para ancho de búsquedas más grandes. En la Figura 2 se ilustra este comportamiento.

Para cuantificar cuanto nos alejamos de la secuencia óptima al reducir la búsqueda, comparamos las secuencias de unidades generadas para cada oración de prueba para los distintos umbrales de poda. En la Figura 3 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.

[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.

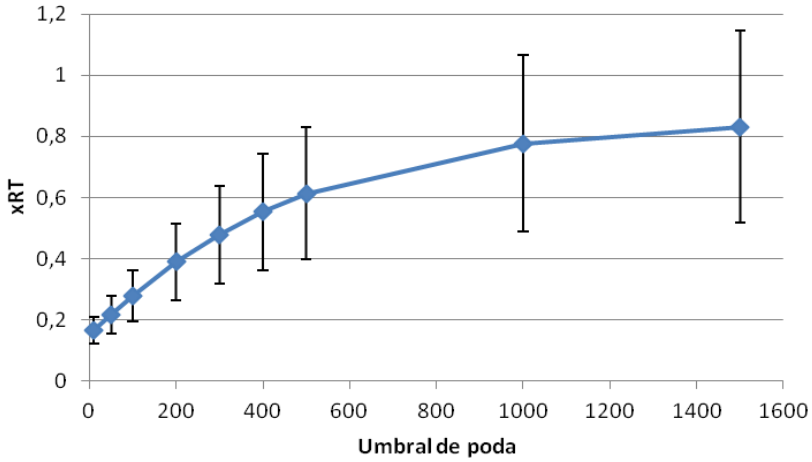


Figura 1: Relación de Tiempo Real (xRT) para los distintos anchos de búsqueda analizados. Las barras verticales indican el desvío estándar de la muestra.

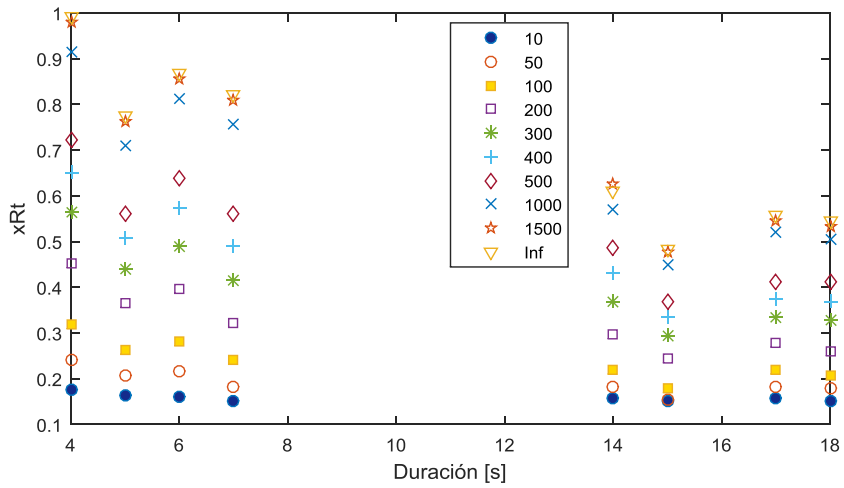


Figura 2: Relación de Tiempo Real (xRT) para oraciones de distintas longitudes, discriminados para los distintos anchos de búsqueda analizados.

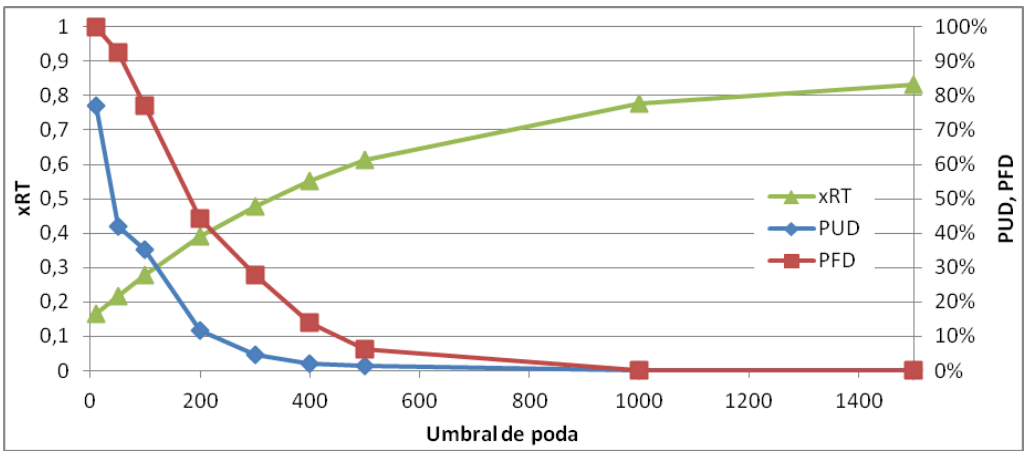


Figura 3: Relación de Tiempo Real (xRT), Porcentaje de Unidades Diferentes de la secuencia óptima (PUD), y Porcentaje de Frases Diferentes de las secuencias óptimas (PFD), para los distintos anchos de búsquedas analizados.

Conclusiones

Se realizó una evaluación de la Relación de Tiempo Real del sistema de conversión de texto a habla AROMO. Los resultados obtenidos indican que la xRT es levemente dependiente de la longitud de las frases a convertir, siendo este efecto más visible a con ancho de búsqueda de unidades más grandes. El valor del ancho de búsqueda de unidades 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.

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.

A.2. Implementación de un Sistema de Key Word Spotting. *Evin, D.A., Torres, H.M., y Gurlekian, J.A.*

Informe Técnico

Implementación de un Sistema de Key Word Spotting

Diego A. Evin, Humberto M. Torres, Jorge A. Gurlekian

Laboratorio de Investigaciones Sensoriales – INIGEM – CONICET

- 01 de abril de 2016 -

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 los autores.

Resumen

Este informe resume el desarrollo de un sistema para la detección automática de palabras clave en registros de emisiones radiales AM y FM de la Ciudad Autónoma de Buenos Aires para la empresa **GlobalNews Group Argentina**. El sistema implementado y transferido a dicha empresa está basado en Modelos Ocultos de Markov (HMM). En este reporte se detallan las características del sistema y los resultados preliminares obtenidos.

Introducción

Desarrollar un sistema para la transcripción automática del habla contenida en un fragmento de audio radial o televisivo es una tarea compleja, que demanda contar con grandes corpus de datos para el entrenamiento de los modelos que permiten la clasificación de patrones acústicos. Una de las consecuencias más probables de entrenar modelos acústicos con escasa cantidad de datos de es obtener sistemas de reconocimiento poco robustos, y cuyas tasas de error en el reconocimiento de palabras sea elevada. Ante este tipo de escenario sin embargo es posible desarrollar sistemas que no busquen identificar todas las palabras de una emisión, sino que permitan saber si un conjunto de palabras de interés aparece o no en una elocución. La detección de palabras claves (keyword spotting), es un sub-campo del reconocimiento automático del habla (RAH) que busca detectar y explotar sólo un fragmento del contenido completo de una frase hablada. Se puede encontrar esta actividad bajo otras

denominaciones, por ejemplo: indexado de audio, indexado de habla, búsqueda fonética, detección de palabras, minería de audio, o recuperación de información oral.

La mayoría de los sistemas de KWS se pueden agrupar en tres métodos distintos:

1. KWS basado en sistemas de RAH de gran vocabulario (LVCSR): emplean un sistema de LVCSR para obtener la transcripción de todo el audio y luego se hace una búsqueda de palabras clave en el texto resultante.
2. KWS Acústico: en este caso el vocabulario de salida está compuesto solamente por las palabras clave compuestas por cadenas de fonos.
3. KWS por búsqueda fonética: en este caso se emplea un reconocedor a nivel de fonos sobre todo el material de audio, y un motor de búsqueda fonética emplea una red o cadena de fonos para determinar la presencia o no de las palabras clave.

Según el escenario de uso cada una de esas estrategias tiene ventajas y desventajas **[1-6]**. Por ejemplo los métodos 1 y 3 tienen la ventaja de eficiencia y practicidad en caso que las listas de palabras clave sean dinámicas. En el caso de esos dos tipos de sistema, la etapa de reconocimiento de habla que resulta muy costosa computacionalmente se lleva a cabo una sola vez, y se guarda su resultado. De esta forma los datos originales de audio ahora se expresan mediante una representación intermedia, sobre la que resulta mucho más fácil de llevar a cabo procesos de búsqueda. En caso que se modifique la lista de palabras clave, ahora no se necesita buscar sobre audio sino sobre texto o cadenas de fonos, un proceso mucho más rápido. En cambio, en el caso de KWS acústico, ante cada nueva lista de palabras clave se debe correr nuevamente el reconocedor de habla sobre el mismo material de audio.

Debido a la cantidad de datos para entrenamiento disponible en este trabajo se decidió implementar un sistema equivalente al propuesto en **[7]**. En esta aproximación se entrenan modelos de subunidades empleando un corpus con sus transcripciones grafémicas o léxicas. Luego, a partir de esas subunidades se construyen los modelos definitivos, compuestos por un modelo oculto de Markov (HMM) para cada palabra clave y HMMs denominados de basura, de relleno o de palabras fuera del vocabulario, que representan el conjunto de las señales de habla que no corresponden a palabras clave. La detección de una palabra clave bajo esta aproximación se efectúa evaluando si el mejor camino de decodificación (empleando el algoritmo de Viterbi) pasa a través del modelo de palabra clave o no. De esta forma se busca que siempre que el usuario diga algo que no constituya una palabra clave, sea capturado por el modelo de rellenos, para ser descartado. Se pueden encontrar diferentes variantes respecto a la forma

de modelar esas palabras fuera del vocabulario, en este trabajo se construye ese modelo conectando completamente todos los modelos de subunidades.

La figura 1 presenta un diagrama en bloques del sistema de KWS desarrollado, y en las secciones siguientes se detalla la construcción de cada uno de sus módulos.

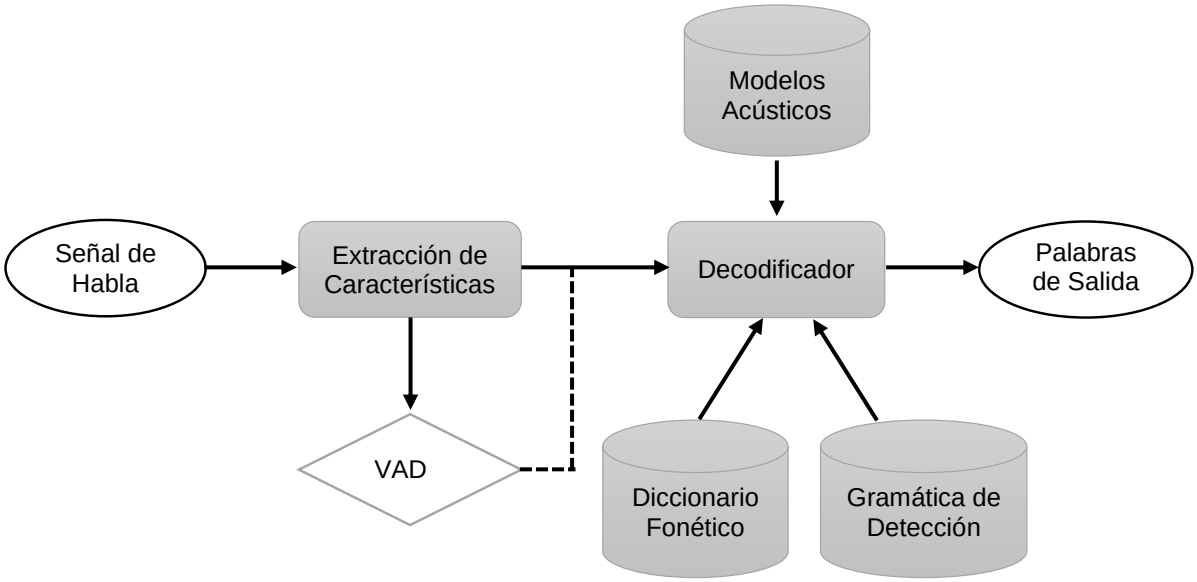


Figura 1. Diagrama en bloques del sistema de KWS desarrollado

Extracción de Características

En esta implementación se utilizó la codificación MFCC estándar, obtenida mediante la función HCopy de HTK **[8]** empleando los siguientes parámetros:

- Ventana utilizada: Hamming
- Ventana de análisis: 25 ms
- Frecuencia de ventaneo: 10 ms
- Coeficiente del filtro de preénfasis: 0.97
- Número de canales en el análisis MFCC: 24
- Número de coeficientes cepstrales utilizados: 12
- Se agrega energía, delta y aceleración
- No se agrega ruido

- No se escala la energía logarítmica
- Coeficiente de realce de coeficientes: 22
- Se sustrae la media temporal
- Se normaliza la energía de la frase

Es decir, como se utilizan 12 coeficientes cepstrales originales más energía son 13 coeficientes. Además, como se agregan los coeficientes de velocidad y aceleración (derivadas primera y segunda de los coeficientes originales), se obtiene un vector de 39 coeficientes cada 10 ms.

Modelos Acústicos

La construcción de modelos acústicos se realizó empleando el conjunto de herramientas para reconocimiento del habla HTK **[8]**.

A continuación, se detallan los pasos específicos realizados:

1. Entrenamiento de monofonos sin pausa corta. Se partió con un conjunto de 30 monofonos descriptos en **[9]** para el español de Argentina, agregando a ese conjunto las versiones acentuadas de las vocales, además de un modelo de silencio y dos tipos de ruidos de locutor. Es decir que inicialmente el sistema empleó 39 modelos acústicos. Todos estos modelos son HMMs con topología de izquierda a derecha sin saltos, compuestos por cinco estados de los cuales solamente tres son emisores. Estos modelos acústicos fueron inicializados con los valores de media y varianzas globales, empleando la función HCompV. Posteriormente estos modelos se pasaron por cuatro fases sucesivas de reentrenamiento mediante la función HERest.
2. Ajuste del modelo de Silencio y generación de modelo de pausa corta. En este paso se generan transiciones en ambas direcciones entre el estado 2 y 4 del modelo de silencio con el objetivo de hacerlo más robusto, permitiendo a los estados individuales absorber distintos tipos de ruidos impulsivos en los datos de entrenamiento. Además, se genera un modelo especial de pausa corta, con un solo estado emisor enlazado al estado central del modelo de silencio, pero con una topología que permite una transición directa entre el estado inicial y el final. Este modelo de pausa corta se usa para permitir un breve silencio opcional entre pares de palabras.
3. Realineamiento de los datos de entrenamiento. El diccionario de pronunciaciones puede contener diferentes pronunciaciones válidas para una misma palabra. Como los datos de entrenamiento tienen la versión ortográfica o léxica de la palabra y no su forma fonética, mientras que el

programa de entrenamiento requiere saber cuáles son los modelos acústicos que corresponden a una palabra de entrenamiento, por defecto supone que la conversión de la transcripción léxica a la fonética es la primera que aparece en el diccionario. En esta fase se realiza un alineamiento empleando el programa HVite, que permite asociar a cada palabra su versión de mayor verosimilitud acústica de las que aparecen en el diccionario.

4. Entrenamiento de monofonos con pausa corta. Se realizan dos nuevas reestimaciones empleando HERest con el modelo de pausa corta y los datos realineados.
5. Entrenamiento con sucesivos incrementos en el número de mezclas de gaussianas. Como paso final, y con el objetivo de mejorar la calidad del sistema, se evaluó el incremento en el número de gaussianas para representar las distribuciones de probabilidades de cada modelo acústico. Este proceso se realizó empleando el comando 'MU' (Mixture Up) dentro del programa HHEd, que incrementa el número de gaussianas. Luego de cada incremento se realizó un par de reestimaciones y se repitió la operación para finalmente evaluar cuál es la cantidad adecuada para el conjunto de datos disponibles.

Gramática de Detección

Se generó una gramática para la detección de palabras clave con la siguiente estructura:

(!SENT-START **{**FILLER **|** Keyword1 **|** Keyword2 **| ... |** KeywordN **}** !SENT-END **)**

Donde el modelo de fillers está compuesto por todos los monofonos entrenados.

A partir de esa gramática se construyó la red de palabras usando HParse. Este programa genera un grafo equivalente a la gramática definida en el paso anterior.

Diccionario de Pronunciaciones

El diccionario de pronunciaciones se implementó usando el alfabeto fonético para fines tecnológicos SAMPA (Speech Assesment Methods: Phonetic Alphabet), adaptado para el Español de Argentina **[10]**, que utiliza un total de 30 unidades fonéticas. En nuestro caso se agregó a este conjunto modelos para 5 vocales acentuadas, un modelo de risa, de otros ruidos del locutor (plops, estornudos, etc), de silencio y de pausa corta.

Para este reconocedor se construyó manualmente el diccionario de pronunciaciones. Se incluyeron como palabras a cada uno de los monofonos y además a las palabras claves con sus posibles pronunciaciones alternativas.

Resultados

Para poder evaluar el resultado del detector de KWS se necesita tener anotaciones con el mismo formato de la red de palabras del sistema. Para obtener estas anotaciones se corrió un procedimiento de alineamiento forzado sobre el conjunto de anotaciones de tal forma de obtener las anotaciones segmentadas. Para esta operación se desarrolló un script Perl: "makeForcedAlignment".

Para efectuar la evaluación del sistema de KWS se desarrolló script Perl "utestKWSPhon" que permite correr una secuencia de reconocimiento sobre el audio de entrada y contrastar las tasas de reconocimiento respecto al número total de instancias que aparecen realmente etiquetadas en ese conjunto de muestra.

La prueba de validación del sistema se efectuó sobre un conjunto de 8 palabras de interés suministradas por la empresa. Se construyó un sistema para ese conjunto de palabras y se hizo una búsqueda de la mejor combinación de parámetros de probabilidad logarítmica de inserción de palabra (p) y factor de escala de la gramática (s) para el programa HVite. Los resultados sobre esa búsqueda se presentan en el Anexo del presente informe. Como se observa en esa sección hay combinaciones de estos parámetros con las que se detectan todas las instancias de palabras clave, pero a expensas de falsas alarmas.

Se desarrolló y suministró a la empresa, el script Perl “BatchKWS” que ante el ingreso de un directorio conteniendo el conjunto de archivos a analizar, filtra aquellos cuyo formato no es similar al de los datos de entrenamiento, realiza la detección de palabras clave y muestra un reporte por palabras de los archivos e instantes potenciales donde pueden estar presentes, ordenadas por un índice de verosimilitud. Además se instruyó a personal de la empresa en los requerimientos y forma de funcionamiento de esas herramientas. Las pruebas de campo por parte de la empresa dieron resultados satisfactorios.

Finalmente cabe mencionar respecto al uso del sistema, que para modificar el conjunto de palabras clave se requiere que se agreguen las nuevas palabras al diccionario de pronunciaciones y que se vuelva a generar la red de palabras de reconocimiento. Sin embargo, el aspecto positivo es que no se debe reentrenar todo el sistema. Este punto es muy importante puesto que en muchas aplicaciones el conjunto de palabras de evaluación no se conoce de antemano o cambia con el tiempo.

Referencias

[1] Szöke I., Schwarz P., Matejka P., "Comparison of Keyword Spotting Approaches for Informal Continuous Speech". Proceedings of the 9th European Conference on Speech Communication and Technology (EUROSPEECH); 2005 4-8 Sept; Lisbon, Portugal.

[2] Šmídl L., Psutka J., "Comparison of Keyword Spotting Methods for Searching in Speech". Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH–ICSLP); 2006 17-21 Sept; Pittsburgh, Pennsylvania.

[3] Wang D., Tejedor J., Frankel J., "A Comparison of Phone and Grapheme-based Spoken Term Detection". Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2008 30 Mar–4 Apr; Las Vegas, Nevada.

[4] Shen W., White C.M., Hazen T.J., "A Comparison of Query by Example Methods for Spoken Term Detection". Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH); 6-10 Sept 2009; Brighton, United Kingdom.

[5] Zhang, Y., Adl, K., & Glass, J., "Fast Spoken Query Detection Using Lower-Bound Dynamic Time Warping on Graphical Processing Units". In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on (pp. 5173-5176). IEEE.

[6] Moyal A., Aharonson V., Gishri M., "Phonetic Search Methods for Large Speech Databases"; 2013; Springer, New York.

[7] R. C. Rose, D. B. Paul, "A hidden Markov model based keyword recognition system". In International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 1, pages 29–32, Albuquerque, NM, USA, 1990.

[8] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland (2006) "The HTK Book (for HTK Version 3.4) ". Cambridge: Entropic Cambridge Research Laboratory.

[10] Gurlekian, J., Colantoni, L. y Torres, H. (2001), "El Alfabeto Fonético SAMPA y el Diseño de Córpora Fonéticamente Balanceados", Fonoaudiológica, vol. 47 (3), pp. 58–70.

Anexo

Pruebas de selección de parámetros de probabilidad logarítmica de inserción de palabra (p) y factor de escala de la gramática (s). Las palabras fueron anonimizadas por cuestiones de confidencialidad. Solo se incluyeron las palabras con alguna instancia en los datos de evaluación.

#Hits: número de aciertos

#Fas: número de falsas alarmas

#Actual: número de veces que aparece la palabra en el conjunto de prueba

FOM: figura de mérito

p=-10.0 s=5.0

---------- Figures of Merit ----------

KeyWord:	#Hits	#FAs	#Actual	FOM
KW1:	1	351	1	0.00
KW2:	1	352	1	0.00
KW3:	4	845	4	0.00
Overall:	6	1548	6	0.00

76547 non keywords found in test files----------

p=-5.0 s=5.0

---------- Figures of Merit ----------

KeyWord:	#Hits	#FAs	#Actual	FOM
KW1:	1	154	1	0.00
KW2:	1	208	1	0.00
KW3:	4	458	4	0.00

Overall: 6 820 6 0.00

93495 non keywords found in test files----------

p=0.0 s=5.0

---------- Figures of Merit ----------

KeyWord:	#Hits	#FAs	#Actual	FOM
KW1:	1	60	1	0.00
KW2:	1	88	1	0.00
KW3:	4	193	4	0.00
Overall:	6	341	6	0.00

115596 non keywords found in test files----------

p=2.0 s=5.0

---------- Figures of Merit ----------

KeyWord:	#Hits	#FAs	#Actual	FOM
KW1:	1	35	1	4.08
KW2:	1	58	1	0.00
KW3:	4	133	4	0.00
Overall:	6	226	6	0.68

126801 non keywords found in test files----------

p=2.5 s=5.0

---------- Figures of Merit ----------

KeyWord:	#Hits	#FAs	#Actual	FOM
KW1:	1	31	1	20.07
KW2:	1	51	1	0.00
KW3:	4	120	4	0.00
Overall:	6	202	6	3.34

130332 non keywords found in test files----------

p=2.7 s=5.0

---------- Figures of Merit ----------

KeyWord:	#Hits	#FAs	#Actual	FOM
KW1:	1	30	1	24.06
KW2:	0	46	1	0.00
KW3:	4	119	4	0.00
Overall:	5	195	6	4.01

131807 non keywords found in test files----------

p=3.0 s=5.0

---------- Figures of Merit ----------

KeyWord:	#Hits	#FAs	#Actual	FOM
KW1:	1	26	1	36.05
KW2:	0	41	1	0.00

KW3: 4 108 4 0.00

Overall: 5 175 6 6.01

134223 non keywords found in test files----------

p=4.0 s=5.0

---------- Figures of Merit ----------

KeyWord: #Hits #FAs #Actual FOM

KW1: 1 19 1 52.04

KW2: 0 32 1 0.00

KW3: 4 83 4 0.00

Overall: 5 134 6 8.67

143503 non keywords found in test files----------

p=5.0 s=5.0

---------- Figures of Merit ----------

KeyWord: #Hits #FAs #Actual FOM

KW1: 1 13 1 68.03

KW2: 0 26 1 0.00

KW3: 4 64 4 0.00

Overall: 5 103 6 11.34

155035 non keywords found in test files----------

p=8.0 s=5.0

---------- Figures of Merit ----------

KeyWord:	#Hits	#FAs	#Actual	FOM
KW1:	0	3	1	0.00
KW2:	0	11	1	0.00
KW3:	4	26	4	9.08
Overall:	4	40	6	6.05

192321 non keywords found in test files----------

p=10.0 s=5.0

---------- Figures of Merit ----------

KeyWord:	#Hits	#FAs	#Actual	FOM
KW1:	0	3	1	0.00
KW2:	0	6	1	0.00
KW3:	4	10	4	62.03
Overall:	4	19	6	41.35

212384 non keywords found in test files----------

p=15.0 s=5.0

---------- Figures of Merit ----------

KeyWord:	#Hits	#FAs	#Actual	FOM
KW1:	0	0	1	0.00
KW2:	0	1	1	0.00

KW3:	2	2	4	47.00
Overall:	2	3	6	31.34

247878 non keywords found in test files----------