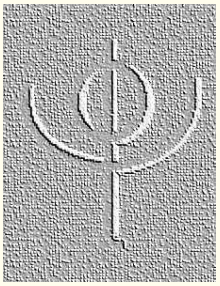
ISSN: 0325-2043



Laboratorio de Investigaciones Sensoriales (LIS)

Informe XLV–2012

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/.

El 14 de septiembre de 2011, el LIS y otros laboratorios del Hospital de Clínicas-UBA constituyeron el *Instituto de Inmunología, Genética y Metabolismo* (INIGEM), dependiente del CONICET y de la Universidad de Buenos Aires.

2. Personal

Investigadores

- 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.
- TOLEDO Guillermo, Lingüista, Dr. en Filosofía y Letras.
- VACCARI María Elena, Lic. en Fonoaudiología.

Becarios

- EVIN Diego, Bioingeniero, Dr. en Ciencias de la Computación. Becario Posdoctoral CONICET
- COSSIO MERCADO Christian, Ing. en Informática, Becario FONCYT. Tesista de Doctorado UBA
- MARTINEZ SOLER Miguel, Ing. en Informática, Becario FONCYT. 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

Tesistas

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

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 Mincyt-BMBF: Extracción y Modelación de los Parámetros Prosódicos para el Análisis, Síntesis y Reconocimiento del habla

Nombre en alemán: *Prosodische Parameterextraktion und Modellierung für die Sprachanalyse, -synthese und -erkennung*

Directores: Jorge A. Gurlekian y Hansjörg Mixdorff. Período: 2009-2011.
Unidad de Ejecución: Laboratorio de Investigaciones Sensoriales y Department of Computer Sciences and Media.
Institución de la que depende la Unidad de Ejecución: CONICET y Technische Fachhochschule Berlin (TFH)
Este proyecto se integra con el proyecto Nombre: PAE Nro: 37122, PID 2007. Nro. 094.

FONCYT. Desarrollo de un sistema de conversión de Texto a Habla. Director: Jorge A. Gurlekian. Período: 2009-2011. Unidad de Ejecución: Laboratorio de Investigaciones Sensoriales.

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: 2009-2011
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

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

5. Docencia

5.1. Cursos de posgrado

Docente: Dra. Miguelina Guirao
Para la Carrera de Especialistas en ORL Facultad de Medicina UBA.
Tema: Mecanismos sensoperceptuales del sistema gustativo
Lugar: Asociacioń Médica Argentina, Buenos Aires, Argentina.
Fecha: 15 de Septiembre 2012

5.2. Otros cursos

Nombre: Identificación forense de voz
Director: Jorge A. Gurlekian
Docentes: Dr. Humberto Maximiliano Torres, Dr. Diego Alexis Evin, Lic. Alejandro Renato, Ing. Christian Cossio Mercado, Ing. Miguel Martinez Soler
Lugar: LIS, INIGEM, CONICET-UBA
Fecha: Julio a octubre de 2012
Duración: 20 horas (teórico-práctico)
Audiencia: Gendarmería Nacional Argentina.

5.3. Seminarios en el laboratorio

- Martes 7 de agosto de 2012: Prof. Richard M. Stern (Professor in the Electrical and Computer Engineering, Carnegie Mellon University). "The current state-of-the art in speech and language".

6. Intercambio Científico

6.1. Visita de Investigadores al LIS

Dr. Hansjörg Mixdorff, mayo 2011

Dentro del Programa de investigación conjunta entre MINCyT Y BMBF de Alemania, se trasladó al LIS el Prof. Dr.-Ing. habil. Hansjörg Mixdorff de la Beuth-Hochschule für Technik Berlin (University of Applied Sciences).

6.2. Estadías en el Exterior de Investigadores del LIS

Dr. Diego Evin, California (EE.UU.), 8 de abril a 8 de octubre de 2013

El Dr.Diego Evin realizó una estadía posdoctoral de seis meses en SRI International.

7. Tesis

7.1. Doctorales

Desarrollo de pruebas de evaluación de la inteligibilidad del habla. *Gurlekian, J.A.*

Dr. Jorge A. Gurlekian

El dia 2 de Octubre de 2012 el Ing. Dr. Jorge A. Gurlekian presentó la defensa de su tesis de doctorado en la Facultad de Medicina de la Universidad de Buenos Aires. El tema de la disertación "Desarrollo de pruebas de evaluación de la inteligibilidad del habla" se refiere al desarrollo de una prueba de evaluación de la inteligibilidad en condiciones de ruido altamente interferente. Se trata de una nueva prueba rapida y de fácil aplicación en diversos ámbitos donde concurren escolares. Tambien puede utilizarse para evaluar el deterioro de la percepción y de la producción del habla.

7.2. 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 semi-automáticos implementados mediante técnicas del estado del arte.

Reconocimiento automático de hablantes empleando información de largo plazo

Tesista: Ing. 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: Lic. 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

- Demostración del sistema de conversión de texto a habla, TN Ciencia, Todo Noticias. Fecha: Junio 2012. Se realizó una demostración del funcionamiento del sistema de conversión de texto a habla del proyecto, como parte de una entrevista realizada por el programa TN Ciencia, del canal de cable Todo Noticias (TN).
- Cossio-Mercado, C.. Tecnologías del Habla, Espacio de la UBA, Tecnópolis. Agosto a noviembre 2012. Realización de experimento y presentación de resultados de prueba de riesgo vocal con el público, como parte de las actividades del Hospital de Clínicas (UBA). Demostración y presentación de sistema de conversión de texto a habla. Introducción al funcionamiento de la percepción humana del habla.

9. Publicaciones

9.1. Revistas

- GUIRAO, M. , GRECO DRIANÓ, E., EVIN, D.A. and CALVIÑO A: "Psychophysical assessments of sourness in citric acid- ethanol mixtures". In *Perception and Motor Skills* (en prensa).
- GURLEKIAN, J.A. y MOLINA, N: "Índice de perturbación, de precisión vocal y de grado de aprovechamiento de energía para la evaluación del riesgo vocal". Revista de logopedia, foniatría y audiología. (AELFA). Editorial Elsevier España. 2012 (en Prensa).

9.2. Congresos

- EVIN, D., GURLEKIAN, J.A., TORRES, H.M.: "Phonological Phrase Segmentation Based On Acoustic Information". In TIMELY Workshop on Dynamical systems for psychological timing and timing in speech processing, Vietri sul Mare, Italy, May 2012.
- TORRES, H.M., GURLEKIAN, J.A.: "Argentine Spanish segmental duration prediction". In Proc. of 13th Argentine Symposium on Technology, 41 JAIIO, pp. 156–167, La Plata, Agosto de 2012.
- GURLEKIAN, JA., COSSIO-MERCADO, C., TORRES, H.M. AND VACCARI, M.E.: "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.
- TORRES, H.M., GURLEKIAN, J.A. AND COSSIO-MERCADO, C.: "Aromo: Argentine Spanish TTS System". In. Proc. of. VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, IberSPEECH 2012, pp. 416–421. Madrid, Spain, 21–23 November 2012.
- UNIVASO, P., MARTÍNEZ-SOLER, M., EVIN, D., GURLEKIAN, J.A.: "A Preliminary Approach to Forensic Speaker Recognition Using Phonemes". In. Proc. of. VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, IberSPEECH 2012, pp. 123–132. Madrid, Spain, 21–23 November 2012.

9.3. Informes Técnicos

- UNIVASO, P., MARTÍNEZ-SOLER, M., EVIN, D.A., y Gurlekian, J.A.: "An approach to forensic speaker recognition using phonemes", 2012.
- TORRES, H.M. y GURLEKIAN, J.A.: "Conversión de Grafemas a Fonemas para un Sistema TTS", 2012.
- TORRES, H.M., "Creación de un corpus de texto para la construcción de un sistema TTS", 2012.

Apéndice

A. Resúmenes de Trabajos

A.1. Argentine Spanish segmental duration prediction. *Torres, H.M. y Gurlekian, J.A.*

Humberto Torres y Jorge Gurlekian

Abstract

In this paper we model the segmental duration of Spanish spoken in Buenos Aires, considering its application in a text-to-speech system. The work was performed on two hand labeled databases. We use artificial neural networks as predictor, and all the input features can be extracted automatically from the speech text. We experimented with a neural network for all phonemes and one neural network for phoneme. In both cases the results are very promising for the two databases used. The order of importance of input features revealed to be different for each of the methods tested and different according to the speaker style.

Keywords

Phone duration prediction; Prosody prediction; Text-To-Speech.

A.2. Subjective Evaluation of a High Quality Text-to-Speech System for Argentine Spanish. *Gurlekian, J.A. et al.*

Jorge Gurlekian, Christian Cossio-Mercado, Humberto Torres y M. Elena Vaccari

Abstract

This work summarizes the perceptual evaluation of our recently developed text-to-speech system (S1), based on unit concatenation. We compare it with two commercially available systems (S2 and S3) using three different evaluation methods. One is the P.85 recommendation by the International Telecommunication Union (ITU), the second method, called Syntactically Unexpected Sentences (SUS), is known to be the most strict for intelligibility evaluation, and the third is the Mean Opinion Score (MOS) scale. Results of ITU test showed better quality and intelligibility responses for system S2. General quality evaluated by MOS and intelligibility evaluated by SUS presented no appreciable differences between S1 and S2. It is concluded that high quality performance is related to a complete intonation modeling of different type of phrase length and styles. S1 has high intelligibility and quality for general information sentences but presented lower scores for specific tasks as proposed in ITU tests, where short phrases coverage should be introduced in the intonational modeling of our system.

Keywords

TTS Evaluation, Intelligibility, Speech Quality, ITU P.85, Mean Opinion Score, MOS, Syntactically Unexpected Sentences, SUS

A.3. Aromo: Argentine Spanish TTS System. *Torres, H.M. et al.*

Humberto M. Torres, Jorge A. Gurlekian, y Christian Cossio-Mercado

Abstract

This paper introduces Aromo text-to-speech system for Argentine Spanish, which was designed for telephony applications and is based on unit selection and concatenation. The system operates as a client-server engine that supports MRCP, SIP and SSML technologies. Perceptual evaluation results show that Aromo's voice achieve high performances in both naturalness and intelligibility.

Keywords

Text-to-Speech; Argentine Spanish TTS; Unit-selection Synthesis

A.4. A preliminary approach to forensic speaker recognition using phonemes. *Univaso, P.*

Pedro Univaso, Miguel Martínez-Soler, Diego Evin y Jorge Gurlekian

Abstract

Present work focus on speaker identification using phonemic information. An ASR system -based on HMMs- is employed to extract acoustic information from phonemes as speaker identity features. Main contribution resides on proposing phoneme forced alignment which could be performed at the forensic environment. This approach with manual intervention, is further combined with a GMM approach, resulting in an expected improvement on performance. Tests on a fixed-phone database of 136 Argentine-Spanish speakers were performed to evaluate the proposed approach in two conditions: first, using the same recording session and channel for both the suspect and evidence and second adding babble noise to the evidence. Results for the ideal condition show an error rate reduction of 68 % relative to the GMM baseline system. The difference in favor of HMM with forced alignment against GMM holds for different SNR conditions.

Keywords

Automatic speaker recognition; Hidden Markov Models (HMM); Gaussian Mixtures Models (GMM); Logistic Regression; Support Vector Machines (SVM)

B. Informes Técnicos

A continuación se incluye el texto completo de los informes técnicos realizados para el año 2012.

B.1. Creación de un corpus de texto para la construcción de un sistema TTS. *Torres, H.M.*

Informe de técnico: Creación de un corpus de texto para la construcción de un sistema TTS

Autores: Dr. Bioing. Humberto M. Torres

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

Fecha: 1 de Diciembre de 2012

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

Diseñar e implementar un corpus de texto para la creación de una base de datos acústica para un sistema de conversión de texto a habla, que utilice síntesis de voz por concatenación de unidades. La unidad de concatenación elegida es el difonema, y se debe asegurar un recubrimiento mínimo de cinco ocurrencias de cada una de estas. La distribución de las unidades debe ser semejante a las encontradas en diarios de circulación local. Debe incluir frases del tipo declarativas, y en un porcentaje menor del tipo interrogativas. El número total de frase debe ser minimizado, teniendo en cuenta los requerimientos anteriores.

Creación del corpus textual

El diseño del texto del corpus a utilizar para crear un sistema de conversión de texto en habla (TTS, del inglés Text-to-Speech) es de vital importancia por su efecto sobre la calidad final del habla generada. No importa que métodos o tecnologías utilicemos en un sistema de TTS, si el corpus de texto no fue diseñado e implementado en forma adecuada, la calidad del habla generada será pobre e insuficiente (Möbius, 2000; Chalamandaris et al, 2011).

Existen diversas aproximaciones a la hora de diseñar un corpus de texto. Algunos prefieren comenzar con una gran volumen de texto, y luego aplicar diferentes técnicas para reducir el corpus al tamaño deseado (Santen and Buchsbaum, 1997; Kelly et al. 2009). En cambio, otros proponen comenzar con un texto pequeño, que cumpla determinadas especificaciones, por ejemplo que estén presentes todas las unidades de síntesis, para luego ir agregando mas texto hasta alcanzar el volumen deseado (Breen and Jackson, 1998). En Matousek et al 2001 se propone crear un corpus que siga la distribución del habla natural, y presentan un algoritmo iterativo que permite seleccionar un subconjunto de oraciones que contengan un mínimo de ocurrencia para todas las unidades de concatenación. Una objeción a este algoritmo es que el número de oraciones necesarias para lograr un mínimo de ocurrencia para cada unidad, puede ser extremadamente grande y difícil de implementar. En otros trabajos se formulan la pregunta si el texto debe ser seleccionado bajo ciertas consignas o debe ser totalmente aleatorio (Lambert et al, 2007).

Una métrica importante en un sistema TTS por selección de unidades es el índice de cobertura, que mide la probabilidad de poder sintetizar una frase al azar con las unidades presentes en el corpus. Nuestra

unidad de síntesis es el difonema, y partiendo de que nuestro alfabeto fonético tiene 31 elementos, incluido el silencio, con solo 961 difonemas podríamos sintetizar cualquier texto, es decir, tendríamos un recubrimiento del 100%. Pero en la selección de unidades no solo se tiene en cuenta la identidad fonética si no también otros rasgos como el carácter acentuado o no, el contexto fonético, los parámetros acústicos, entre otros. Por lo cual, asegurar el recubrimiento de un corpus no es una tarea trivial (Santen and Buchsbaum, 1997).

Nuestro corpus de textol fue creado en tres etapas, las cuales se detallan a continuación.

Primera etapa

La primera etapa, tenía como objetivo lograr oraciones con todas las sílabas del español, en las dos condiciones de acento (sílaba acentuada y no acentuada) y en todas las variantes posicionales (inicial, media y final) dentro de la palabra. Para ello se elaboraron 741 oraciones declarativas que emplean el 97% de las sílabas del español. El 70% de las oraciones fueron obtenidas de los periódicos que se publican en Buenos Aires. El resto fue creado por maestros de lengua quienes recibieron la instrucción de elaborar oraciones con palabras que contuvieran las sílabas menos frecuentes (Gurlekian et al, 2001). Esta primera aproximación tenía como objetivo principal estudiar la prosodia del español hablado en Buenos Aires (Gurlekian et al, 2001).

Segunda etapa

La segunda etapa tenía como objetivo lograr un recubrimiento mínimo de 5 realizaciones de cada uno de los difonos utilizados en el español. Además, se requirió que la distribución de difonos del corpus sea semejante a la del habla natural. También se incluyeron oraciones del tipo interrogativas.

Primero se estimó la identidad de los difonemas que usualmente utilizamos en el español hablado en Buenos Aires, y su distribución relativa. Para ello se obtuvieron grandes volúmenes de texto extraídos de diarios locales, en un total de 2.896.666 oraciones. Luego se los transcribió fonéticamente con una herramienta especialmente diseñada (Gurlekian et al. 2001). Con estas transcripciones se estimaron los difonemas a emplear. Una dificultad es que no todos los difonemas encontrados se corresponde a palabras del español. El filtrado de estas se hizo en forma manual. Otro problema es la baja tasa de ocurrencia de ciertos difonemas, y los cuales también deben ser analizados en forma manual. Por último, se calculó el histograma de los difonemas resultantes para obtener una distribución de referencia.

Luego, del total de las oraciones disponibles se seleccionaron aquellas que no tenían palabras extranjeras, y que tuvieran entre cinco y diez palabras de longitud. Este es un proceso iterativo y semiautomático, dado que no se puede conocer a priori las palabras con difonos que no corresponden al español. La restricción de longitud de las frases se tomó para que sean similares a las creadas en la primera etapa de construcción del corpus.

Después, se sintonizó el histograma de las frases resultantes con el de referencia. En este proceso también se tuvo en cuenta la ocurrencia de por lo menos cinco realizaciones de cada difonema. Esto último es una tarea muy laboriosa, teniendo en cuenta la baja tasa de ocurrencia de determinadas unidades. Para llevar adelante esto, se utilizó el diccionario de la Real Academia Española, y sustituyendo palabras de las frases preseleccionadas.

Por último, pensando en las posibles aplicaciones del corpus, se agregaron frases con connotación de cortesía.

El corpus de texto resultante sumó 652 frases declarativas y 200 del tipo interrogativa. Si bien en forma teórica se cumple con la cobertura requerida, luego de la grabación y etiquetado de las oraciones, se debe realizar una verificación manual, dado que la transcripción teórica de grafemas a fonemas no siempre coincide con la locución.

Tercera etapa

Con el corpus obtenido al final de la segunda etapa se construyó una voz para nuestro sistema de TTS (Torres et al, 2012). La calidad de la voz obtenida por el TTS fue evaluada perceptualmente de acuerdo a tres normas (Gurlekian et al 2012). La primera fue la P.85 establecida por la International Telecommunication Union (ITU-T Recomendation P.85.), que realiza una exhaustiva y completa evaluación del habla generada. La segunda fue la Syntactically Unexpected Sentences (Benoit et al, 1996), que se centra en el aspecto de inteligibilidad. Y la tercera fue la escala Mean Opinion Score comúnmente empleada para medir la naturalidad del habla (ITU-T Recomendation. P.800). A partir del análisis de los resultados de las evaluaciones perceptuales, se resolvió incorporar nuevas oraciones al corpus, con el fin de mejorar algunas de las falencias encontradas. La mayoría de los errores detectados se debían a dos razones: la ausencia o la escasa presencia de algún difono, generalmente asociado a algún extranjerismo; frases con estructuras no presentes en el corpús, por ejemplo, oraciones con muchas frases entonativas o con frases entonativas de pocas palabras.

A partir de este análisis se sumaron al corpus 198 nuevas oraciones, creadas en forma manual. Además, se agregaron 35 oraciones interrogativas, incluyendo preguntas frecuentes, y en todas las terminaciones fonéticas posibles.

Análisis estadístico del corpus

En la Tabla 1 se presentan las estadísticas del corpus textual al final de cada etapa de construcción. Los datos presentados se desprenden del etiquetado semiautomático realizado por fonoaudiólogas con entrenamiento musical y licenciados en lingüística.

Tabla 1: Estadísticas del corpus textual.

	1ª etapa	2ª etapa	3ª etapa
Nro. de oraciones	741	1.593	1.826
Minutos de habla	37.41	88.99	141.74
Nro. de oraciones declarativas	741	1.393	1591
Nro. de oraciones interrogativas	0	200	235
Nro. de frases entonativas	1.224	2.388	4.429
Nro. de frases entonativas por oración	1.65	1,50	2,43
Nro. de palabras	5280	13495	21.693
Nro. de palabras por frase	4.31	5.65	4,90
Nro. de difonos distintos	480	525	660
Nro. de difonos	28.343	68.324	110.789
Nro. de fonos distintos	30	30	30
Nro. de fonos	27.120	65.938	106.364

Hay dos observaciones que podemos realizar a partir del análisis de la Tabla 1. Primero, el aumento del número de difonos distintos de una etapa a la otra. Esto es así, por que en la primera etapa no se tuvo en cuenta el recubrimiento a nivel de difonos; y en la tercera etapa se incluyeron palabras extranjeras, particularmente nombres propios derivados de otras lenguas, que generan combinaciones de fonos que no ocurren normalmente en el español. Con respecto al número de posibles combinaciones de fonos, que es 961, no se logra en el corpus dado que muchas combinaciones no están permitidas, como por ejemplo [nb]. La segunda observación está relacionada con el número de frases entonativas por oración, que varía notablemente de la etapa 2 a la etapa 3. Esto se explica teniendo en cuenta que en las oraciones agregadas en la etapa 3 fueron especialmente diseñadas para tener una estructura que no estaban presentes en las etapas anteriores.

En la **¡Error! No se encuentra el origen de la referencia.**.a) se presenta el número de ocurrencias de los fonemas, según el alfabeto SAMPA. El fonema más representado es el [e], con 14.029 instancias, y el menos es el [g], presente con 159 casos. En este gráfico podemos observar el grado de disparidad en el número de ocurrencia de cada fonema en el corpus. Esto está de acuerdo con los estudios previos de Guirao y Garcia Jurado, 1993.

En la Figura 1.b) se puede observar el número de difonos agrupados por rangos de número de ocurrencias en el corpus. Los rangos fueron definidos a priori con el objetivo de ilustrar la distribución. El grupo de los difonos con menos o igual a cinco ocurrencias son 148, que representan los difonos menos probables y que solo se dan en situaciones muy específicas, como los difonos generados por la fusión de un nombre propio, que termina en una consonante y la palabra que le sigue que comienza con una consonante, por ejemplo el difono [fn] en el contexto de las palabras "… Budasof nació …". El grupo de los 25 difonos con más presencia en el corpus, con más de mil ocurrencias cada uno, está integrado por difonos de alta frecuencia en el habla, por ejemplo [el], [es], [en], [la], entre otros.

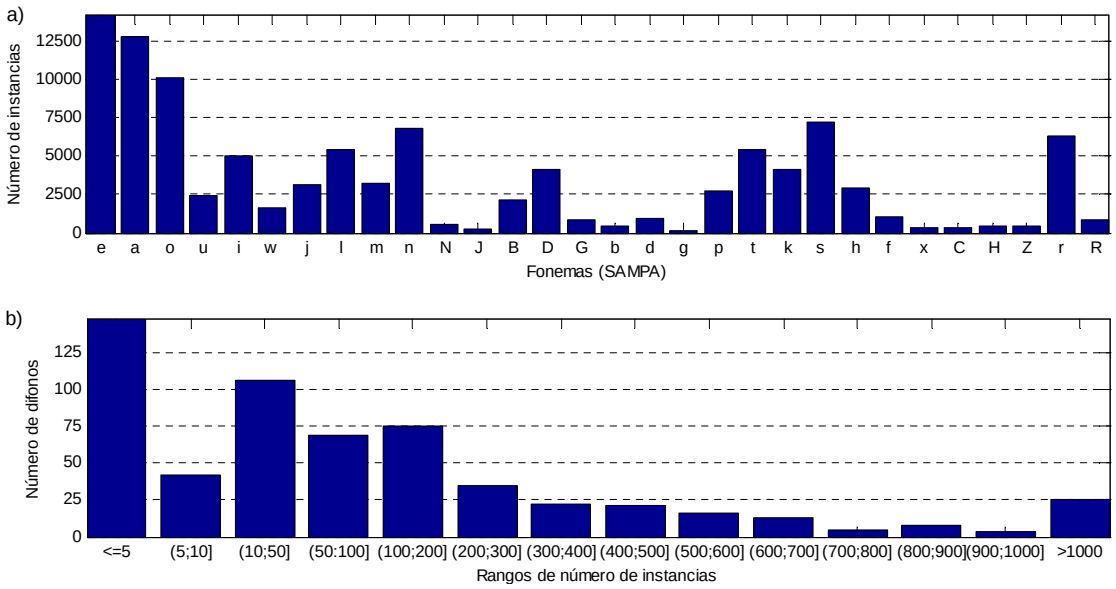


Figura 1: a) Número de ocurrencias de los fonemas según el alfabeto SAMPA; y b) Número de difonos según rangos de número de ocurrencias.

Bibliografía

Bonafonte, A. , Höge, H., Kiss, I., Moreno, A., Ziegenhain, U. , Van Den Heuvel, H., Hain, H., Wang,X. and Garcia, M., "TC-STAR:Specifications of Language Resources and Evaluation for Speech Synthesis". In Proc. of the fifth International Conference on Language Resources and Evaluation, Genoa, Italy, May 2006.

Breen, Andrews P. and Jackson, P."Non-uniform unit selection and the similarity metric within BT's Laureate TTS system". In Proc. of the Third ESCA Workshop on Speech Synthesis, pp. 373-376. Jenolan Caves, Australia, 1998.

Chalamandaris, A., Tsiakoulis, P., Raptis, S. and Karabetsos, S.; "Corpus Design for a Unit Selection TtS System with Application to Bulgarian". In Human Language Technology. Challenges for Computer Science and Linguistics. Lecture Notes in Computer Science Volume 6562, 2011, pp 35-46.

Guirao,M. y M.A. Garcia Jurado: Estudio estadístico del español. CONICET, Buenos Aires. Argentina,

1993.

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. (http://www.asalfa.org/fono.html).

Gurlekian, J., Rodriguez, H., Colantoni, L. and Torres, H., "Development of a Prosodic database for an Argentine Spanish text to speech system". In Proc. of the IRCS Workshop on Linguistic Databases, University of Pennsylvania, Philadelphia, USA, December 11-13, 2001, Ed. for Bird, Buneman & Liberman. pp. 99-104.

Gurlekian, J., Cossio-Mercado, C., Torres, H. and Vaccari, M., "Subjective Evaluation of a High Quality Text-to-Speech System for Argentine Spanish", En Proceedings of VII Jornadas en Tecnolog´ del Habla and III Iberian SLTech Workshop, IberSPEECH 2012, pp. 241–250, Madrid, 2012.

Kelly, A., Berthelsen, H., Campbell, N., Chasaide A. and Gobl, C., "Corpus Design Techniques for Irish Speech Synthesis". In Proc. of China Ireland ICT Conference Maynooth, Ireland, August 2009.

Lambert, T., Braunschweiler, N. and Buchholz, S., "How (Not) to Select Your Voice Corpus: Random Selection vs. Phonologically Balanced". In Proc. of 6th ISCA Workshop on Speech Synthesis, pp. 22-24, Bonn, Germany, August 2007.

Matousek, J. and Psutka, J., "Design of Speech Corpus for Text-to-Speech Synthesis". In Proc. of Eurospeech 2001, pp. 2047-2050. Alborg, 2001.

Möbius, Bernd, "Corpus-Based Speech Synthesis: Methods and Challenges", Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS 6 (4), pp.87–116., 2000.

Oliveira, L., Paulo, S., Figueira, L., Mendes, C., Nunes, A. and Godinho, J., "Methodologies for Designing and Recording Speech Databases for Corpus Based Synthesis". In Proc. of the Sixth International Conference on Language Resources and Evaluation, pp. 2921-2925. Marrakech, Morocco, May 2008.

Torres, H., Gurlekian, J. and Cossio Mercado, C., "Aromo: Argentine Spanish TTS System". In. Proc. of. VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, IberSPEECH 2012, pp. 416-421. Madrid, España, 21-23 November 2012.

Van Santen, J. and Buchsbaum, A., "Methods for optimal text selection". In Proc. of the European Conference on Speech Communication and Technology; vol. 2, pp. 553-556, Rodhos, Greece, 1997.

B.2. Conversión de grafemas a fonemas. *Torres, H.M. y Gurlekian, J.A.*

Informe de técnico: Conversión de grafemas a fonemas

Autores: Dr. Bioing. Humberto M. Torres; Ing. Dr. Jorge A. Gurlekian

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

Fecha: 1 de Diciembre de 2012

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

Crear reglas que permitan la conversión de grafemas a fonemas, para su posterior implementación en un sistema de conversión de texto en habla. Este mapeo debe abarcar todas las posibles combinaciones de grafemas, de forma tal que podamos sintetizar cualquier texto con los difonos presentes en el corpus de datos.

Reglas para la conversión de grafemas a fonemas

Un alfabeto fonético es un conjunto de símbolos que representan gráficamente los sonidos que pronunciamos al hablar, en forma biunívoca y reflejando todas las posibles variaciones. Podemos considerarlos como una representación simbólica del habla. Algunos ejemplos de alfabetos fonéticos son el Alfabeto Fonético Internacional (IPA, del inglés International Phonetic Association) (IPA, 1999), Métodos para la Evaluación del Habla: Alfabeto Fonético (SAMPA, del inglés, Speech Assessment Methods: Phonetic Alphabet) (ESPRIT, 1989), entre otros.

La transcripción fonética consiste en representar en forma escrita lo que hablamos. A la transcripción fonética a partir de un texto se la denomina conversión de grafemas a fonemas.

En un sistema de conversión de texto a habla necesitamos un módulo que convierta el texto de entrada en la secuencia de sonidos que debe pronunciar, es decir una transcripción de grafemas a fonemas. Este módulo también debe considerar las variantes de cada fonema, es decir, sus alófonos.

En trabajos anteriores (Gurlekian et al 2001), hemos presentado una adaptación del alfabeto fonético SAMPA para el español hablado en Argentina, y también desarrollamos una herramienta que permite convertir en forma automática un texto en su representación fonética. En la Tabla 1 se reproduce el alfabeto SAMPA propuesto, y en la Tabla 2 se presentan las reglas de transcripción propuestas. Aquí presentamos las reglas adaptadas a la voz EMILIA de nuestro sistema de conversión a habla AROMO. Las reglas originales fueron adecuadas a la locutora que grabó el corpus para implementar la voz de EMILIA, con condición adicional de poder emitir cualquier texto con las unidades presentes en el corpus.

Tabla 1: Alfabeto SAMPA para el español de la Argentina. Extraído de Gurlekian et al 2001.

SAMPA para Argentina					
N	***IPA***	**SAMPA**	**Modo, Punto, F0**	**Transcripción**	**Palabra**
1	i	i	Vocal, cerrada frontal, sonora.	b**i**s	Bis
2	e	e	Vocal, cerrada frontal, sonora.	m**e**s	Mes
3	a	a	Vocal, central abierta, sonora.	m**a**s	Más
4	o	o	Vocal, media posterior redondeada, sonora.	t**o**s	Tos
5	u	u	Vocal, posterior cerrada redondeada, sonora.	t**u**l	Tul
6	j	j	Aproximante, palatal, sonora	laB**j**o	Labio
7	ω	w	Aproximante, labial velar, sonora.	aG**w**a	Agua
8	l	l	Lateral, ápico-gingival, sonora.	**l**oBo	Lobo
9	m	m	Nasal, bilabial, sonora	**m**esa	Mesa
	ɱ	m	Nasal, labiodental, sonora.	e**n**fermo	Enfermo
10	n	n	Nasal, ápico-gingival, sonora.	**n**aDa	Nada
11	ŋ	N	Nasal, velar, sonora.	o**N**Go	Hongo
12	ɲ	J	Nasal, dorso-palatal, sonora.	ni**J**o	Niño
13	β	B	Fricativa, bilabial, sonora.	tu**B**o	Tubo
14	ð	D	Fricativa, ápico-interdental, sonora.	a**D**a	Hada
15	ɤ	G	Fricativa, dorso-velar, sonora.	a**G**wa	Agua
16	b	b	Oclusiva, bilabial, sonora.	**b**eso	Beso
17	d	d	Oclusiva, ápico-dental, sonora.	**d**ar	Dar
18	g	g	Oclusiva, dorso-velar, sonora.	**g**ula	Gula
19	ɾ	r	Vibrante simple, ápico-gingival, sonora.	pe**r**o	Pero
20	r	R	Vibrante múltiple, ápico-gingival, sonora.	ca**R**o	Carro
21	ʎ	Z	Lateral, dorso-prepalatal, sonora.	**Z**uBja	Lluvia
	ʒ	Z	Fricativa, prepalatal, sonora.	A**Z**er	Ayer
	ʤ	Z	Fricativa, dorso-palatal, sonora.	**Z**uBja	Lluvia
	ʃ	Z	Africada, dorso-prepalatal, sonora.	Kon**Z**uGe	Cónyuge
	ɹ	Z	Fricativa, dorso-prepalatal, sorda.	**Z**uBja	Lluvia
22	H	h	Fricativa, laríngea, sorda.	a**h**ta	Hasta
23	P	p	Oclusiva, bilabial, sorda.	**p**ala	Pala
24	T	t	Oclusiva, ápico-dental, sorda.	**t**ierra	Tierra
25	K	k	Oclusiva, dorso-velar, sorda.	**k**ilo	Kilo
26	ʧ	H	Africada, dorso-prepalatal, sorda.	te**H**o	Techo
27	s	s	Fricativa, ápico-gingival, sorda.	**s**ala	Sala
28	F	f	Fricativa, labio-dental, sorda.	**f**e	Fe
29	X	x	Fricativa, dorso-velar, sorda.	**x**wes	Juez
30	ç	C	Fricativa, palatal, sorda.	ar**C**entina	Argentina

Tabla 2: Reglas de conversión de grafemas a fonemas.

Grafema	Contexto	SAMPA
a	Todos	a
e	Todos	e
i	Vocal + i_sin_acento; i_sin_acento + Vocal	j
	Restantes	i
o	Todos	o
u	Vocal + u_sin_acento; u_sin_acento + Vocal	w
	Restantes	u
b	b_inicial; Consonante_Nasal + b	b
	Restantes	B
c	c + e; c + i	s
	Ck	k
	Restantes	k
ch	Todos	H
d	d_inicial; Consonante_Nasal + d; l + d	d
	Restantes	D
f	Todos	f
g	g_inicial; Consonante_Nasal + g	g
	gu + e; gu + i	g
	g + ü	g
	g + e; g + i	C
	Restantes	G
h	Todos	ELIMINAR
j	j + a; j + o; j + u	x
	a + j + Consonante; o + j + Consonante; u + j + Consonante	x
	j + e; j + i	C
	e + j + Consonante; i + j + Consonante	C
	Restantes	x
k	Todos	K
l	Todos	l
ll	Todos	Z
m	Todos	m
n	n + f; n + b; n + p; n + v	m
	n + Consonante_Velar	N
	Restantes	n
ñ	Todos	J
P	Todos	p
q	Todos	k
	qu + e; qu + i	k
r	r_inicio_de_palabra; n + r; r + r	R
	l + r; s + r; h + r	R
	Restantes	r
S	s + vocal	s
	sh; sch	Z
	Restantes	h
T	Todos	t

Grafema	Contexto	SAMPA
V	v_inicial; Consonante_nasal + v	b
	Restantes	B
W	Todos	w
X	Todos	ks
Y	y + vocal	Z
	y_fin_de_palabra	i
Z	z + Consonante	h
	Restantes	s

En la Tabla 3 se puede observar el número de ocurrencias de las difonos encontrados en el corpus de datos. En esta tabla, se encuentran las identidades del semifono izquierdo de los difonos en las filas, y en las columnas el semifono derecho. Según el alfabeto SAMPA para el español hablado en la Argentina existen 22 fonemas (ver Tabla 1). Estos componentes son utilizados en todos los países de habla hispana. Para el español peninsular se agregan los fonemas /z/ y /v/ no utilizados en Latinoamérica. Si consideramos los 8 alófonos, más frecuentes en la Argentina, mas el silencio inicial/final, el número de difonos posibles es 961. Pero no todas estas combinaciones dan a lugar a un difono, ya que algunas combinaciones de fonemas no son posibles debido a restricciones fonotácticas. Además, hay que considerar que algunos fonemas son más probables que otros (Guirao y García Jurado, 1993), a tal punto que algunos difonos no se utilizan en el español y solo se dan en por la presencia de palabras de otros idiomas. Finalmente, hay que contemplar la posibilidad que la locutora genere difonos propios al producir alófonos diferentes, modificando las reglas propuestas en la Tabla 2. Los alófonos que han sido cambiados por la locutora son [D] por [d] en el difono [lD] y [s] por [h] para los difonos [sh] y [sf], y en el contexto de [nhp], [nht] y [nhk] que se transforman en [nsp], [nst] y [nsk], respectivamente.

Estos considerandos se han reflejado en la Tabla 3 mediante el color de sus celdas, según se indica a continuación:

- Las celdas en color blanco agrupan los difonos que la locutora ha generado de acuerdo con las reglas de transcripción universales (RTU) y que son frecuentes de hallar en el habla real.
- Las celdas en color azul contienen a los difonos que también son posibles de realizar, pero tienen una baja probabilidad de ocurrencia. El número de ocurrencias de estos difonos es inferior al mínimo estipulado de cinco. Estas representaciones son adecuadas debido a que las probabilidades de aparición en el habla real son muy bajas (Torres 2012).
- Las celdas en color verde y celeste son posibles de generar según las RTU, pero en los textos analizados no se hallaron ejemplos. La ausencia de ejemplos de difonos de las celdas en verde se debe a cambios que realiza la locutora en las RTU. Estas modificaciones no generan nuevos difonos, y se listan en la Tabla 5. En las celdas celestes agrupamos a los difonos que pueden formarse por palabras extranjeras, o por la fusión de nombres propios, terminados en consonantes, con la palabra siguiente. Para asegurarnos la cobertura fonética del corpus, creamos un conjunto de reglas de posprocesamiento que se listan en Tabla 4.
- Las celdas en color naranja y amarillo se generan a partir de un cambio en las RTU causado por las características propias de la locutora. Las celdas en naranja son imposibles según las RTU, pero el elevado número de ocurrencias de estos difonos, en combinación con la escasa ocurrencia de los difonos en las celdas amarillas, impulsan cambios en las reglas de transcripción, las cuales se listan en la Tabla 5.
- Las celdas en color rojo resaltan a los difonos imposibles de componer según las RTU. Se puede observar que existen algunas ocurrencias de los difonos en las celdas en rojo, donde la locutora rompe con las reglas, pero su número no es significativo como para reescribir alguna RTU.

Las reglas de transcripción fueron exitosamente implementadas en AROMO. Las pruebas realizadas hasta ahora prueban la consistencia y el recubrimiento del sistema de transcripción propuesto.

Tabla 3: Distribución de las ocurrencias de los difonos en el corpus de datos.

Identidad del primer fono	Identidad del segundo fono																														
	e	a	o	u	i	w	j	l	m	n	N	J	B	D	G	b	d	g	p	t	k	s	h	f	x	C	H	Z	r	R	-
e	355	487	137	20	25	130	297	1677	564	2201	147	96	461	686	236	0	31	0	314	454	763	1586	1108	153	80	99	60	62	1363	174	442
a	378	255	106	21	44	245	324	1201	585	1005	70	73	565	1248	134	4	47	0	434	483	669	1420	579	170	84	98	35	142	1415	216	744
o	353	247	74	31	7	95	135	437	665	1421	99	4	341	591	81	7	28	0	343	238	488	1291	687	111	24	29	182	39	1007	105	891
u	3	6	1	2	3	1	33	196	181	567	31	5	154	154	32	5	17	1	108	91	117	166	138	25	13	11	73	29	223	35	20
i	7	269	24	0	2	8	1	372	304	569	148	17	275	474	111	0	7	0	99	363	464	610	345	108	20	31	17	51	278	15	51
w	799	327	23	0	63	0	0	8	38	203	25	0	5	20	4	1	0	0	1	49	7	22	26	5	0	0	7	3	27	0	6
j	852	544	970	46	2	1	0	27	83	210	27	0	13	74	11	0	2	0	18	29	48	72	26	8	4	6	5	7	13	6	29
l	684	1555	681	142	440	47	94	31	129	60	0	5	111	192	85	1	2	0	164	249	221	156	0	55	10	5	12	9	0	52	230
m	549	637	446	171	388	12	93	1	1	10	0	2	19	0	0	287	3	1	398	1	5	2	0	150	0	0	1	2	0	2	38
n	728	879	801	148	315	212	63	243	90	68	0	4	0	39	0	0	538	0	5	1472	0	658	2	10	0	0	32	6	0	50	362
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	139	0	0	366	0	19	0	23	43	0	0	0	0	0
J	15	64	138	9	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	538	407	239	55	299	39	135	170	1	5	0	2	2	6	2	0	0	0	6	8	2	15	1	5	2	4	1	5	168	0	14
D	1659	648	849	71	422	12	176	6	17	8	0	2	7	5	5	0	5	0	13	8	14	9	1	4	3	1	1	3	61	3	95
G	23	229	186	73	34	56	9	25	11	27	0	2	1	1	0	0	0	0	0	1	1	1	0	1	0	0	2	0	149	0	17
b	88	61	32	20	41	73	65	8	0	0	0	0	0	0	0	3	0	0	0	2	0	9	0	0	0	2	0	0	71	0	0
d	425	81	175	27	93	2	54	0	0	0	0	0	1	1	1	0	51	0	0	1	0	1	0	0	0	0	0	0	21	0	5
g	4	43	25	21	2	8	3	9	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	37	0	3
p	426	608	587	151	112	116	68	155	0	5	0	2	0	1	2	1	1	1	1	36	2	28	0	1	0	0	2	1	437	0	12
t	1142	1312	1021	213	579	50	170	6	8	5	0	1	6	4	2	1	0	1	3	2	3	4	0	4	1	0	2	1	807	0	26
k	653	762	1267	204	135	281	68	77	2	17	0	1	0	4	2	1	1	0	2	258	2	243	1	2	0	3	2	0	144	1	17
s	1585	814	644	575	940	93	1006	18	10	17	0	0	13	31	2	1	5	0	57	69	35	158	0	81	3	3	0	0	0	1	1087
h	0	0	0	0	0	3	0	107	186	116	0	3	66	340	59	8	8	0	410	1129	425	1	0	2	10	5	5	3	0	48	0
f	142	108	141	53	230	121	33	47	0	1	0	1	0	1	0	1	0	1	0	6	1	0	0	0	1	0	0	1	104	0	6
x	0	108	103	60	0	21	0	2	1	1	0	1	1	1	1	0	0	0	0	1	1	0	0	0	1	1	2	1	0	1	5
C	265	0	0	0	101	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
H	96	131	150	13	37	7	4	1	0	0	0	1	1	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0	0	1	11
Z	104	145	91	25	20	3	1	2	2	2	0	2	2	0	2	0	0	1	2	1	2	1	0	0	0	0	1	0	0	1	10
r	998	1506	940	129	406	17	196	230	216	99	0	1	78	221	74	2	0	0	54	265	174	288	1	35	11	49	8	7	0	12	308
R	516	87	70	30	56	6	26	3	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-	822	474	131	131	239	5	48	363	125	145	1	11	8	15	3	152	192	14	323	156	340	505	0	67	23	6	7	48	0	75	0

	Nro. de difonos	Nro. de instancias
Posibles según las reglas	444	109577
Posibles según las reglas, con pocas instancias.	143	342
Posibles por cambio de reglas de la locutora	3	431
Generado artificialmente	1	1
Posibles según reglas, sin instancias, transformados en el posprocesamiento	76	0
Imposibles por cambio de reglas de la locutora, sin instancias.	14	0
Imposibles por cambio de reglas de la locutora, descartados.	3	4
Imposible según las reglas	275	442

	Nro. de difonos	Nro. de instancias
Totales	961	110790
Finales	594	110343

Tabla 4: Reglas adicionales de posproceso para la transcripción de grafemas a fonemas para asegurar la cobertura fonética del corpus.

Combinación	Transformación	Combinación	Transformación
m + x	m + - -+ x	x + H	x + - - + f
m + C	m + - - + C	C + m	C + - - +m
m + r	m + R	C + n	C + - - +n
n + h	n + s	C + N	C + - - +N
J + j	J + i	C + J	C + - - +J
G + G	G + - -+ G	C + B	C + - - + B
G + p	G + - -+ p	C + D	C + - - + D
G + h	G + s	C + G	C + - - + G
G + x	G + - -+ x	C + p	C + - - + p
G + C	G + - -+ C	C + t	C + - - + t
G + Z	G + - -+ Z	C + k	C + - - + k
G + R	G + r	C + s	C + - - + s
p + m	p + - - + m	C + h	C + - - + h
P + B	p + - - + B	C + f	C + - - + f
p + h	p + s	C + x	C + - - + x
p + x	p + - - + x	C + C	C
p + C	p + - - + C	C + H	C + - - + H
p + R	p + r	C + Z	C + - - + Z
t + h	t + s	C + r	C + - - + R
t + C	t + - - + C	C + R	C + - - + R
t + R	t + r	H + m	H + - - + m
k + B	k + - - + b	H + n	H + - - + n
k + x	k + - - + x	H + N	H + - - + N
k + Z	k + - - + Z	H + D	H + - - + D
f + m	f + - - + m	H + G	H + - - + G
f + N	f + n	H + p	H + - - + p
f + B	f + - - + B	H + k	H + - - + k
f + G	f + - - + G	H + h	H + s
f + p	f + - - + p	H + x	H + - - + x
f + s	f + - - + s	H + H	H
f + h	f + - - + h	H + Z	H + - - + Z
f + f	f	Z + N	Z + n
f + C	f + - - + H	Z + D	Z + - - + D
f + H	f + - - + C	Z + h	Z +s
f + R	f + r	Z + f	Z + - - + f
x + N	x + n	Z + x	Z + - - + x
x + p	x + - - + p	Z + C	Z + - - + C
x + s	x + - - + s	Z + Z	Z
x + h	x + - - + h	-h	-s

Tabla 5: Reglas adicionales de posproceso para la transcripción de grafemas a fonemas propias de la locutora.

Combinación	Transformación
i + u	j + u
w + u	u + u
w + w	u + u
w + j	w + i
w + J	u + j
w + x	u + x
w + C	u + c
w + R	u + R
j + j	j + i
j + J	i + J
B + R	B + r
x + r	x + R
H + r	H + R
Z + r	Z + R
l + d	l + D
s + h	s + s
h + f	s + f
n + h + p	n + s + p
n + h + t	n + s + t
n + h + k	n + s + k

Bibliografía

M. Guirao y M A García Jurado “Estudio estadístico del Español”. Conicet. 1993

J. Gurlekian, L. Colantoni y H. Torres “*El alfabeto fonético SAMPA y el diseño de corpora fonéticamente balanceados*”, Fonoaudiológica Nº 3, pp. 58-69. Diciembre de 2001. (http://www.asalfa.org/fono.html).

J. Gurlekian, H. Rodriguez, L. Colantoni and H. Torres, “Development of a Prosodic database for an Argentine Spanish text to speech system”. In Proc. of the IRCS Workshop on Linguistic Databases, University of Pennsylvania, Philadelphia, USA, December 11-13, 2001, Ed. for Bird, Buneman & Liberman. pp. 99-104.

IPA, International Phonetic Association. (1999). Handbook of the International Phonetic Association. A guide to the use of the International Phonetic Alphabet. Cambridge: Cambridge University Press.

ESPRIT. 1989. http:www.phon.ucl.ac.uk/home/sampa/home.htm

H. Torres, “Informe de técnico: Creación de un corpus textual para la construcción de un sistema TTS”. Laboratorio de Investigaciones Sensoriales, UBA-CONICET. Diciembre de 2012.

B.3. An approach to forensic speaker recognition using phonemes. *Univaso, P. et al.*

Technical Report: An approach to forensic speaker recognition using phonemes

Pedro Univaso, Miguel Martínez Soler, Diego Evin, and Jorge Gurlekian

Laboratorio de Investigaciones Sensoriales, Facultad de Medicina, U.B.A.,
Córdoba 2351 Piso 9 Sala 2, Buenos Aires, Argentina
punivaso@yahoo.com.ar,{miguelmsoler,diegoevin}@gmail.com,jag@fmed.uba.ar
http://www.lis.secyt.gov.ar/

Abstract. The present work is focused on building up reference patterns for forensic speaker recognition which include phonemic information as identity attributes. An HMM based ASR system extracts precise acoustic information from phonemes by forced alignment, which is further combined with the standard method of Gaussian mixtures GMM, resulting in a new method that we have called GHMM. In order to combine the classifiers two methods were employed. One based on logistic regression and the second based on a linear combination. For a mobile-phone database of 136 Argentine-Spanish speakers, results show a reduction of errors measured by EER of 7% relative to the HMM system alone and by 69% relative to the conventional GMM system.

Keywords: Automatic speaker recognition, Hidden Markov Models (HMM), Gaussian Mixtures Models (GMM), Logistic Regression, Support Vector Machines (SVM).

1 Introduction

Speaker recognition main application fields are found in forensics and in security systems. Speaker identification as seen in forensics is initiated from voice recordings produced at a criminal situation. These recordings are named dubitable or evidence, which are later matched with recordings called indubitable or suspicious that belong to a known person. Standard techniques employed by forensic recognition experts are based on those known as speaker verification by speech scientists. These techniques propose a Bayesian approach to solve a likelihood proportion between matching hypothesis. Some authors [1] claim that forensic situations must be solved using a robust method that extends this approach by holding the principle of innocence of the suspect candidate. Others [2] indicate to change the standard verification binary decision of acceptance or refusal for the a posteriori probability of the hypotheses, due to the high environment variability and to several factors that affect forensic recognition [3].

In order to represent each speaker voice, generic acoustic probabilities based on GMMs [4] are employed in actual state of art speaker recognition systems. During their evolution, the use of distinctive phonetic features in characteristic

vectors can be found in some previous work. Some of them started by using phoneme based models represented by HMMs [5]. Experiments presented by Kajarekar et al. [6] arranged clusters of wide phonetic categories to produce each speaker models.

Nevertheless, this approach was discarded both in the academic community and the application developments when contrasted with the performance obtained with GMM based systems, because of practical advantages of GMM solutions, as their computational efficiency and text independence of samples.

A drawback of GMMs system implementations is that they usually do not take advantage of linguistic information both phonological and phonetic present in running speech. Phoneme discriminative power has been studied by several authors [7, 8] but the way for incorporating this knowledge in speaker recognition systems still is an open issue. High level information of prosodic features as intonation, rhythm and accent are not adequately exploited. Hansen et al. [9] provided a useful background to the present work as they showed improvements relative to standard GMM approach by using a GMM for each phoneme. Some investigations from Stanford Research Institute evaluated the introduction of high level features in GMM based systems. It is worth to mention two works recently presented by Stolcke et al. [10] and Kajarekar [11] who added words, syllables and phonemes as parameters for the speaker recognition task. SRI systems showed the best performance during the NIST evaluation contest. Systems developed at SRI as well as other systems which used models based on high level features [12] produce a final model as a combination between both high and low level models. The result is always an improvement relative to the isolated model. Model combination is achieved in different ways: by neural networks, supporting vector machines, logistic regression or some other particular classifiers [13].

Inspired by this successful approach, we present a new contribution on speaker recognition at the forensic domain that evaluates the use of acoustic information present at phoneme level. Our proposal is to produce a more precise phoneme model using forced alignments in a HMM. This can be done because at the forensic domain we can produce almost perfect segmentation and labelling to catch later on phoneme information in a stochastic way by mean of HMMs. Our results are also compared with a GMM based system doing the same job. Finally a combination of both systems is obtained by a logistic regression classifier. A linear combination classifier is also employed for this purpose. This new proposed method will be referred as GHMM.

The rest of this paper is as follows: Section 2 presents the proposed method based on phonological information and its relation with a standard automatic method. Section 3 shows the strategy of evaluation. Section 4 presents results obtained, which are discussed in section 5, to finally present our conclusions and future work in sections 6 and 7, respectively.

2 Methodology

2.1 Gaussian Mixture Models

GMM systems are used on speaker recognition applications as a generic probabilistic model for multivariate densities that is able to represent arbitrary densities, making them useful for text independent applications. At the same time they are known statistical models, computationally efficient but insensitive to temporal characteristics inherent of speaker dependent features extracted from acoustic segmental data and/or linguistically oriented data.

Speaker Verification. Verification of speaker L given a speech segment S, tries to determine if S was emitted by L generally using a likelihood ratio as a test [14]. The basic hypothesis are:

- H_0: S was said by L
- H_1: S was not said by L

$$\frac{p(S|H_0)}{p(S|H_1)} \begin{cases} \geq \theta \text{ accept } H_0 \\ < \theta \text{ reject } H_0 \end{cases} \tag{1}$$

In the preceding formula, p is the probability density function for hypothesis H_i, evaluated for the observed segment S. θ is the decision threshold to accept or reject H. The goal is to determine the techniques that compute both likelihoods $p(S|H_0)$ and $p(S|H_1)$.

Once the distinctive feature vector X of the speaker to be detected is obtained, we can calculate the likelihoods of H_0 and H_1. H_0 is represented by a model called MH_0 that represents speaker L in the space of features of X. If a gaussian distribution is assumed for the feature vector of H_0, MH_0 could be seen as representing the mean value and covariance matrix distribution. The alternative hypothesis H_1 is represented by the model MH_1. The likelihood ratio is then:

$$\frac{p(X|MH_0)}{p(X|MH_1)} \tag{2}$$

and its logarithm:

$$\Lambda(X) = log\ p(X|MH_0) - log\ p(X|MH_1) \tag{3}$$

Speaker Model Adaptation. In forensic tasks a small amount of data is collected from the suspect speaker during controlled interviews or collected from certified recordings from that speaker. In recent years GMM based speaker recognition systems used a Bayesian adaptation procedure, starting from a universal reference model (UBM) [14]. Bayesian "adaptation" refers to Bayesian learning or maximum a posteriori estimation (MAP). Another adaptation methods are linear transformations, which solved the maximization problem with the maximum expectation algorithm. Techniques using linear transformations can be: Maximum Likelihood Linear Regression (MLLR) and Maximum Likelihood Linear Regression Restricted (CMLLR).

Proposed Method. It can be seen in Fig. 1 a simplified schematic of the speaker recognition system based on the methodology GMM-UBM with adaptation to the suspect speech used in this work. The GMM can be viewed as a

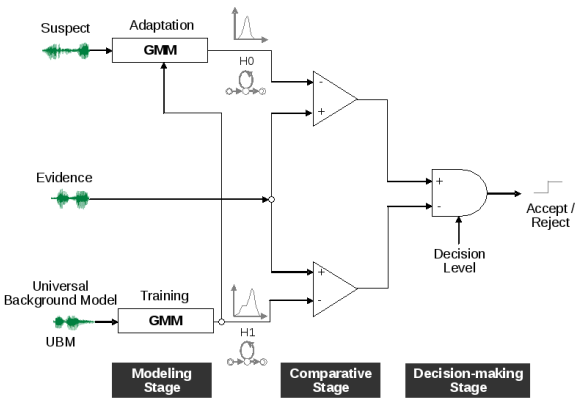


Fig. 1. GMM-UBM methodology.

one-state HMM with a density of observation composed of Gaussians mixtures. Whereupon the acoustic model that represents each speaker is reduced to a single HMM, being the GMM a special case of HMM. This property is used in this work to implement Gaussian mixtures modeling through a standard HMM generation system. It employs a training process in stages starting with a one state HMM with multiple components of Gaussian mixtures and considered a single unit of speech to represent the voice of each speaker. Both in the GMM as in the HMM methods subsequently described, silences are removed in the calculation of the model to be employed. The proposed modeling system for the generation of GMM and HMM includes both a short pause and silence model who is not employed in the final stage of comparison, but they are useful for automatically discriminating passages speech from silence. The HMM model adaptation of the suspect was made through linear transformations of their parameters, using an integrated training version of Baum-Welch algorithm. Starting from the UBM model and re-estimating the same with information on the suspect, it is obtained the new set of adapted HMMs to be used in the speaker recognition system.

2.2 Hidden Markov Models

HMMs are stochastic models widely used for the implementation of ASR systems. They have also been used in the field of text-dependent speaker recognition. For this application, each word is modeled by a HMM, the system being limited to the recognition of a restricted vocabulary that must be trained specifically. Because of its performance and computational economy, the GMMs have been

replacing HMMs in current research and commercial implementations for text-independent speaker recognition. Fig. 2 shows the procedure used in this study to recognize speakers using an HMM-based ASR system. The generation of the universal reference (UBM) speakers acoustic models was performed through a training process in stages according to the methodology proposed by Young [15] for ASR systems.

Adapting the model of the suspect was conducted similarly to the case of GMM, re-estimating each HMM corresponding to UBM with information of the suspect. The step of automatic speech recognition is performed using a general purpose word recognizer based on the Viterbi algorithm. This algorithm optimizes the comparison between an unknown speech utterance, with a network formed by the HMM models obtained in the training stage and whose sequence represents a word in the dictionary, resulting in a transcription of the unknown utterance. The recognizer makes this comparison taking into account both the acoustic and the language model. The optimization in the comparison is made maximizing the likelihoods of the compound word networks as the sums of the likelihoods of each of the triphones that shape them.

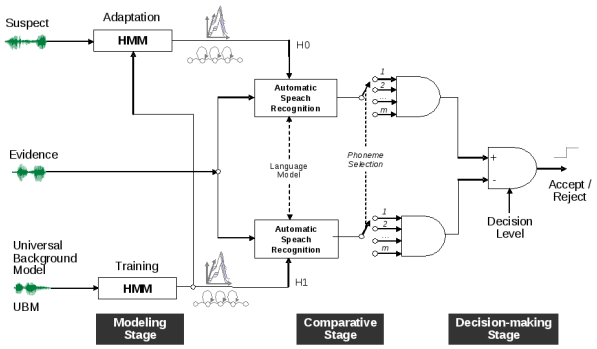


Fig. 2. HMM-UBM methodology.

An alternative to the above method is to employ a forced alignment, instead of the word recognition, for which the Viterbi algorithm uses a word network to recognize the corresponding transcriptions. This alternative is more accurate and fast and does not require the generation of a language model. In the case of forensic applications generally we have these transcriptions (transcript of evidence) and this approach is to be used extensively in this work. In the final stage of comparative analysis, we calculate the average of the likelihoods of each of the phonemes that were recognized to finally obtain an overall average value for each of the comparisons. These two values are representative of the similarity between the evidence and the UBM (H_1) and between evidence and the suspect (H_0), which are used in the final decision analysis using the likelihood ratio test described above.

Considering an alphabet consisting of m phonemes, each of the factors of equation (2) takes the form:

$$p(X|MH_0) = \prod_{i=1}^{m} p(X_i|MH_0) \tag{4}$$

$$p(X|MH_1) = \prod_{i=1}^{m} p(X_i|MH_1) \tag{5}$$

Where X_i is the likelihood of each of the m phonemes, that appear t_i times in the evidence utterance, and is calculated as:

$$p(X_i|MH_0) = \prod_{j=1}^{t_i} p(X_{i,j}|MH_0) \tag{6}$$

$$p(X_i|MH_1) = \prod_{j=1}^{t_i} p(X_{i,j}|MH_1) \tag{7}$$

Being the likelihood ratio given by:

$$\Lambda(X) = log \sum_{i=1}^{m}\sum_{j=1}^{t_i} p(X_{i,j}|MH_0) - log \sum_{i=1}^{m}\sum_{j=1}^{t_i} p(X_{i,j}|MH_1) \tag{8}$$

2.3 Combined method (GHMM)

The first combined methodology proposed here is the result from the application of a logistic regression classifier trained on the output of a SVM classifier. The input data were obtained at the output stage of the comparative analysis methods GMM and HMM, expressed in the formulas (3) and (8) respectively. This methodology was called GHMM (logistic regression).

The second methodology, called GHMM (linear comb.) is a linear combination of these previous data, affected by factor in the interval $[0, 1]$ for the HMM system output and its complement applied to the GMM system output, according to:

$$\Lambda(X)_{LinearComb} = \alpha\Lambda(X)_{HMM} + (1-\alpha)\Lambda(X)_{GMM} \tag{9}$$

3 Evaluation

3.1 Database

The database used, is part of the SALA I Project (SpeechDat Across Latin America) [16]. The style of speaking corresponds to read paragraphs taken from newspapers and books of Argentina or developed by linguists. Recordings were

made through the fixed telephone network through a computer equipped with an AVM-ISDN-A1 board and a basic access interface ISDN (BRI).

For this work we selected sentences of continuous speech from the SALA database Argentina, South region, which will be called SALA I-Continuous. This region includes the provinces of Buenos Aires, Santa Fe, Entre Ríos, La Pampa, Neuquén, Rio Negro, Chubut, Santa Cruz and Tierra del Fuego. The corpus was delimited to 1,301 words, with a total of 9,948 words, corresponding to a vocabulary of 2,722 different words, issued by 136 speakers (47 males and 89 females) for 99 minutes of recording. For final comparisons between the three speaker recognition systems, a 10-fold cross validation method was used. For each partition we used the data from 124 speakers for the formation of UBM and 12 speakers for testing identities.

The average duration of the recordings of each speaker was 44 seconds. 39 seconds were used as the suspect speech and 5 seconds on average as evidence. While the paucity of evidence can lead to low recognition rates, also allows to represent one of the real drawbacks of current forensic field. Using data from women only, improvements were of 54%, but due to the small amount of data, final experiments were performed using the complete database.

3.2 Reference system

HMM models were built using a tool developed by the University of Cambridge: HTK Toolkit ver. 3.4 [15], free for academic use. Digitizing the audio signal is performed at a sampling frequency of 8 kHz and 16 bits, subtraction of the mean time, so as to eliminate any offset from the analog recording stage. We employed an analysis window of 25 ms Hamming type at a frequency of 10 ms windowing, using a pre-emphasis filter with coefficient 0.97, having normalized energy of the sentence and using a logarithmic scale for energy.

The parameters used for creating the models were 13 MFCC coefficients: Mel-Frequency Cepstral Coefficients, including zero coefficient whose value is approximately proportional to the fundamental frequency F0, to which were added delta and acceleration transformations, forming a total of 39 parameters. The selection of the phoneme as the unit of speech to be used in this work was based on the phonetic alphabet SAMPA Speech Assessment Methods: Phonetic Alphabet, adapted for Spanish in Argentina [17]. In addition to the 30 units of SAMPA phonetic alphabet formed with the 22 phonemes of Spanish and 8 allophones often used in Argentina, a model of silence generally present between phrases was included, which is associated with a model of short pause between words, completing a total of 31 monophones that represent the speech of Spanish in Argentina. HMM was created as a 3-state left to right for each of the 31 monophones, except the short pause that is associated with the central state silence model. The construction of context-dependent models (triphones) was performed by grouping monophones by a clustering based on phonetic decision trees, from which we obtained a total of 1,010 clusters. Eventually older models were transformed in associated Gaussian mixture models, designated in the literature as

semi-continuous models (SC-HMM) using a splitting method to achieve a number of GMM by model 256 (equal to the amount of the GMM standard contrast method used in comparative experiments). The best fit is achieved with only one re-estimation, increasing the speaker recognition error by 41% when using 2 re-estimates.

For the comparative stage all the phonemes and states of the HMM were used, since the use of phonetic vowel groups, and groups of consonants or in combination produced greater errors (see Fig. 3).

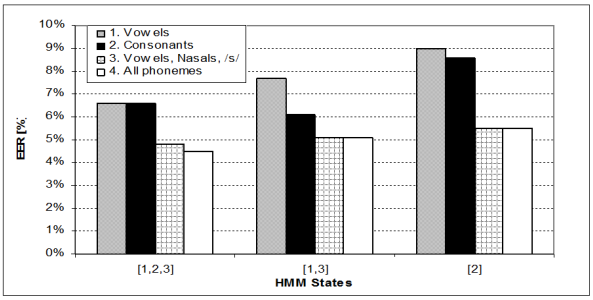


Fig. 3. Results of the HMM methodology with different phonemes and HMM states in the process of comparative analysis.

4 Results

Table 1 shows comparative results of equal error rate (EER) and minimum cost rate of detection ($C_{Det-min}$), produced by the standard methods and the proposed method with different combination procedures.

Table 1. Results of equal rate error and minimum cost rate in the recognition of speakers for each method tested.

Method	Equal error rate EER[%]	Minimum cost rate $C_{Det-min}$ [%]
GMM (baseline)	13.4	8.7
HMM - word recognizer	15.5	6.7
HMM - forced alignment	4.5	2.0
GHMM logistic regression	5.9	1.9
GHMM - linear comb. EER opt.	4.2	2.0
GHMM - linear comb. CDet-min opt.	4.2	1.8

HMM methodology produced more errors than the GMM baseline, as expected. The forced alignments in HMMs introduced the most significant change with an improvement 66% in EER, and 77% in minimum-cost rate related to GMM. Then, various GHMM combination strategies were built and tested based on HMM-forced alignment and GMM. Parameters required for the logistic regression on SVM were determined by a previous run on the training data. The linear combination optimizing EER was achieved with a factor α=0.85, and the optimization of $C_{Det-min}$ was obtained with α=0.9. GMM-HMM combinations using both logistic regression and a linear combination algorithm achieve further improvements of 1% and 2% respectively.

DET (detection error trade-off) curves in Fig. 4 show overall results from the experiments for GMM, HMM, and GHMM methods.

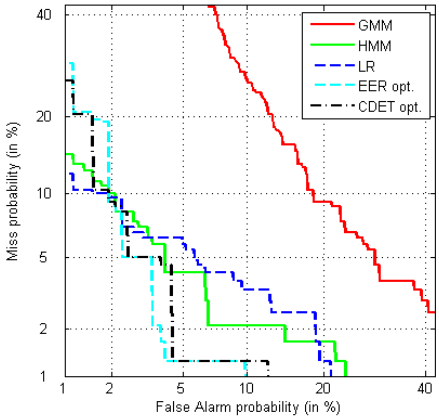


Fig. 4. Average DET curves over the 10 folds for GMM, HMM with forced alignment, linear regression (LR), and both linear combinations (optimizing EER and $C_{Det-min}$, respectively).

5 Discussion

Results of the new method show the best performance between all experiments we made. The use of forced alignments has proved in this work, as it could be supposed in advance, a clear improvement over both GMM and HMM phoneme modeling. Considering that the scientific community has been working mainly on verification tasks in a fully automatic way, the forensic field could benefit from this basic idea of using the known text to produce forced alignments and then to continue with the verification paradigm using a universal background model of our language region.

Regarding a comparison with state of art found in the literature at the forensic domain, we should mention a 2007 contribution by Gonzalez-Rodriguez,et al.[19]. From their work we can deduce an EER about 5%. They use a set of formant patterns of English diphthongs in h-d context produced by 171 Australian speakers. Also, 2010 NIST evaluation of speaker recognition used a 10 sec waveform of spontaneous speech to train and test. This is the closer task to our present contribution but by automatic means and using a big data base for the UBM. From this task the deduced EER is over 10%[18].

Using this approach Castaldo et al. reported in 2011 an EER of 10.44%[20]. Besides, they also used a joint factor analysis to make UBM GMM model adaptations to cope the problem of channels, sessions, microphones and so on. In a similar work based on an i-vector system, Mandasari et al. reported an EER of 14,68%[21].

In summary, our results for the fully automatic approach are comparable to the state of art in terms of percentage EER using GMMs. For the semiautomatic approach our results are highly promising. Regarding manual intervention we could only find the contribution previously cited[19], where the 5% EER is higher than ours. No $C_{Det-min}$ was presented in that work. Forensic approach should also consider other metrics like C_{llr} related to the adjustment of a reported conditional probability instead of a hard decision response.

6 Conclusions

Determining the most efficient way to use speaker intrinsic information to improve speaker identification is still a debt we have today. The methodology presented here is an alternative approach that attempts to contribute to solve this issue. The way phonemes are produced by the speaker and their variations are highly speaker dependent.

Modeling speaker voice using HMM contributes to enhance phonological information present at phonemic level. The use of forced alignment in HMMs is possible for forensic purposes, where the court requests the use of speaker identification techniques using transcriptions made in advance. System combinations produce a further improvement, confirming what was stated in other studies [10], and where the linear combination methodology is simpler and more effective than logistic regression considering EER metric.

7 Future Work

First, increasing the amount of training data by using the full SALA database will permit to adequately assess performance with triphones, given that there is a compromise between the quality of representation of the context and number of samples available to adequately estimate models for each of these contexts. Also we expect to correlate speaker profiles with gender, age, dialect, type of telephone channel and environment situation. Other methodologies will be analyzed for

adaptation (MLLR, CMLLR) instead of the re-estimation used, based on Baum-Welch algorithm.

We expect to evaluate the role of individual phonemes. Giving them differential weights inspired in perceptual judgments and in previous work [8] where phonetic groups as nasals, vowels and fricatives have already demonstrated to produce better results in speaker discrimination. Also, the proposed method will permit research in a flexible manner on the influence of other parameters that contribute to speaker recognition, such as: phoneme durations and silences, distribution of phonemes and words used in spontaneous conversations.

Acknowledgments. To project PID 35891 "Development of techniques for speaker recognition", awarded by the Ministry of Science, Technology and Productive Innovation of Argentina.

References

1. Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., Ortega-Garcia, J.: Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. Computer Speech and Language, 20(2-3), 331-355 (2006)
2. Campbell, W., Reynolds, D., Campbell, J., Brady, K.: Estimating and evaluating confidence for forensic speaker recognition. In: Proc. ICASSP, pp. 717-720, Philadelphia (2005)
3. Campbell, J., Shen, W., Campbell, W., Schwartz, R., Bonastre, J., Matrouf, D.: Forensic speaker recognition. IEEE Signal Processing Magazine, 26(2), 95-103 (2009)
4. Reynolds, D. A.: Speaker Identification and Verification Using Gaussian Mixture Models. Speech Commun., 17, 91-108 (1995)
5. Matsui, T., Furui, S.: Concatenated Phoneme Models for Text-Variable Speaker Recognition. In: Proc. of ICASSP-93, Vol. 2, pp. 391-394, Minneapolis (1993)
6. Kajarekar, S. S., Hermansky, H.: Speaker Verification Base on Broad Phonetic Categories. In: Proceedings of Speaker Odyssey 2001, pp. 201-206, Crete, Greece (2001)
7. Sambur, M. R.: Selection of Acoustic Features for Speaker Identification. Acoustics, Speech and Signal Processing, IEEE Transactions on, 23(2), 176-182 (1975)
8. Eatock, J. P., Mason, J. S.: A Quantitative Assessment of the Relative Speaker Discriminating Properties of Phonemes. In: Proceedings of ICASSP 1994, Vol. 1, pp. 133-136, Adelaide, Australia (1994)
9. Hansen, E., Slyh, R., Anderson, T.: Speaker Recognition using Phoneme-Specific GMMs. In: Proceedings of Speaker Odyssey 2004, pp. 179-184, Toledo, Spain (2004)
10. Stolcke A., Shriberg E., Ferrer L., Kajarekar S., Sonmez K., Tur G.: Speech Recognition as Feature Extraction for Speaker Recognition. In: Proc. SAFE 2007: Workshop on SP Applications for Public Security and Forensics, pp. 39-43, Washington, D.C. (2007)
11. Kajarekar S., Scheffer N., Graciarena M., Shriberg E., Stolcke A., Ferrer L., Bocklet T.: The SRI NIST 2008 Speaker Recognition Evaluation System. In: Proc. ICASSP, pp. 4205-4208, Taipei (2010)

12. Campbell, J. P., Reynolds D. A., Dunn, R. B.: Fusing High- and Low-Level Features for Speaker Recognition. Proceedings of Eurospeech 2003, pp. 2665-2668 Geneva, Switzerland (2003)
13. Ferrer, L., Graciarena M., Zymnis, A., Shriberg, E.: System Combination using Auxiliary Information for Speaker Verification. In: Proc. ICASSP, pp. 4853-4856, Las Vegas (2008)
14. Reynolds D., Quatieri T., Dunn R.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 10(1-3), 19-41 (2000)
15. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtech, V. Wooland, P.: The HTK Book. Cambridge University Press (2006)
16. Moreno A.: SALA: SpeechDat Across Latin America. In: Proceedings of the I Workshop on Very Large Databases, Athens, Greece (2000)
17. Gurlekian, J., Colantoni, L., Torres, H., Rincn, A., Moreno A., Mario J.: Database for an automatic Speech Recognition System for Argentine Spanish. In: Proc. Of the IRCS Workshop on Linguistic databases, pp. 92-98, Pennsylvania (2001)
18. NIST SRE10 Results. `http://www.nist.gov/itl/iad/mig/sre10results.cfm`
19. Gonzalez-Rodriguez, J., Rose, P., Ramos, D., Toledano, D. T., Ortega-Garcia, J.:Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition. IEEE Transactions on Audio, Speech and Language Processing, 15(7), 2104-2115 (2007)
20. Castaldo, F., Colibro, D., Vair, C.: Loquendo-Politecnico di Torinos 2010 NIST speaker recognition evaluation system. In: Proc. ICASSP, pp. 5464-5467, Torino, Italy (2011)
21. Mandasari, M., McLaren, M.: Evaluation of i-vector Speaker Recognition Systems for Forensic Application. In: 12th Annual Conference of the International Speech Communication Association, pp. 21-24, Florence, Italy (2011)