Subjetividad y Procesos Cognitivos, Vol. 14, Nº 2, 2010
Pág. 89-98, ISSN impreso: 1666-244X, ISSN electrónico: 1852-7310

Sistemas de análisis acústico y de reconocimiento automático en habla espontánea

Acoustic analysis and automatic recognition systems in spontaneous speech

Jorge A. Gurlekian, Diego Evin, Humberto Torres y Alejandro Renato*

Resumen

En este trabajo se presentan dos sistemas de análisis acústico del habla con aplicaciones a la descripción de segmentos de discurso espontáneo y un sistema de reconocimiento automático de habla espontánea orientado a la detección de palabras. El primer sistema de análisis presenta detalladamente todos los rasgos instintivos segmentales y suprasegmentales del habla en forma simultánea asociados a la frecuencia, energía y duración. El segundo presenta automáticamente los parámetros físicos asociados a la entonación en una superficie que cuantifica el campo vocal del hablante y mide el rango vocal y dinámico en el discurso hablado. Se presenta un histograma de la frecuencia fundamental útil para comparar las tendencias entonativas de sesión a sesión. Finalmente se ha desarrollado una herramienta de reconocimiento con modelos acústicos para el español hablado en la Argentina. El mismo transcribe los sonidos grabados a texto y posibilita la aplicación de otras herramientas para el procesamiento de lenguaje natural.

Palabras clave: análisis acústico, reconocimiento automático del habla, frecuencia fundamental.

Summary

In this paper two acoustic speech analysis systems are presented with applications to the description of spontaneous speech segments and a system of automatic spontaneous speech recognition oriented to word detection. The first analysis system presents in detail all segment and supra-segment instinct speech features simultaneously and associated frequency, power and duration. The second automatically displays the physical parameters associated to intonation in a surface that quantifies the vocal field of the speaker and measures the vocal and dynamic range in spoken discourse. A histogram of the fundamental frequency proves useful to compare intonation tendencies from session to session. Finally a recognition tool with acoustic models

* Laboratorio de Investigaciones Sensoriales Hospital de Clínicas, Facultad de Medicina de la UBA. Dirección: Av. Córdoba 2351 Piso 9º, Sala 2 (C1120AAF), Ciudad de Buenos Aires, Argentina. Tel.: (011) 5950-9024, E-mail: jag@fmed.uba.ar, diegoevin@gmail.com, alejandrorenato@gmail.com

was developed for Spanish spoken in Argentina. It transcribes the recorded text sounds and enables the application of other tools for natural language processing.

Key words: acoustic analysis, automatic speech recognition, fundamental frequency.

1. Introducción

La percepción del habla permite extraer información lingüística (lexical, sintáctica semántica y pragmática), paralingüística (intencional, actitudinal y estilística) y extralingüística (emocional y física) en una proporción mucho mayor que la representada solo en forma de texto. Los cálculos arrojan una proporción de 100:1 (Gurlekian, Franco y Toledo, 1983). Por otra parte el oído humano no puede procesar toda la información acústica en forma simultánea y debe elegir solo una pequeña proporción (50 bits en un segundo de habla), cuando la señal acústica que recibe tiene muchísima más información (5000 bits por segundo). La respuesta a esta enorme diferencia la da el nivel cognitivo del oyente, el grado de predicción de lo que vendrá y su concentración o focalización en determinados intereses. Para complementar al oído humano y ofrecer una medida objetiva de la toda la información que transporta el habla, se utilizan sistemas de análisis acústico como los que se presentan a continuación, incluyéndose la descripción los avances logrados para el español hablado en la Argentina.

Los sistemas de análisis acústico operados manualmente permiten la visualización de los rasgos distintivos del habla y de sus alteraciones en segmentos breves como los fonemas, sílabas y palabras. Características como el temblor, los niveles de escape de aire en la fonación y el grado de periodicidad, dan información relevante sobre cambios involuntarios en el estado emocional del sujeto. El análisis segmental permite individualizar eventos de interés tales como pausas, pausas llenas, dudas propias o provocadas y eventos especiales localizados por el profesional durante el discurso. La medición de los formantes, con sus energías y anchos de banda se han empleado en trabajos que correlacionan estos rasgos con con la depresión y riesgo suicida (France et al., 2000). Para las frases y oraciones se ha producido un avance en la descripción de las características prosódicas (ritmo, acento y entonación) del español. Se han descripto los correlatos acústicos de estas sensaciones (Guirao, 1980; Toledo, 1988, 2001; Colantoni y Gurlekian, 2004) y presentado un modelo de entonación para el español de Buenos Aires (Gurlekian, Colantoni y Torres, 2001). El análisis de segmentos seleccionados de habla permite extraer características de largo plazo como índices de la calidad vocal. La descripción dinámica en un mismo sujeto de los rangos de frecuencia e intensidad que presenta el sistema de análisis son indicadores de cambios en la expresión.

En el área de reconocimiento automático del habla se han desarrollado sistemas que se ajustan a las características específicas de los locutores de Argentina (Gurlekian, et al., 2001; Univaso, Gurlekian, y Evin, 2009). Estos sistemas están basados en la creación de modelos probabilísticos para cada unidad acústica del lenguaje, modelos estadísticos de las palabras que podrá utilizar el usuario y modelos de pronunciaciones

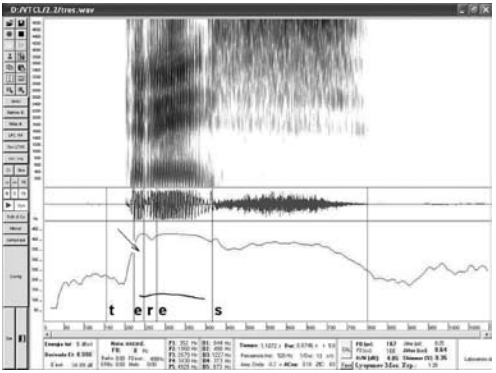
que indican cómo se relacionan las unidades acústicas para conformar palabras. El proceso de reconocimiento consiste en una búsqueda de los modelos acústicos y palabras que con mayor probabilidad se haya presentado dentro de la señal de habla. El desempeño de los reconocedores depende de la calidad de las grabaciones que se utilicen para llevar a cabo la tarea, al tipo de habla y a los rasgos que presente cada locutor. Los porcentajes de reconocimiento obtenidos en el laboratorio utilizando locutores profesionales y un ambiente especial para realizar las grabaciones superan el 97%. En habla espontánea las tasas de reconocimiento empleando vocabularios medianos se encuentran alrededor del 70%.

El reconocimiento automático de emociones ha recibido últimamente aportes del área de inteligencia computacional con la selección automática de rasgos y con la exploración de sistemas de clasificación para categorizar estados emocionales en voces neutrales y con estrés (Casale, Russo, and Serrano, 2007; Albornoz, Milone y Rufiner, 2008; Albornoz, Crolla y Milone 2010).

2. Sistemas de análisis acústico

El primer sistema de análisis acústico que proponemos es ANAGRAF (Gurlekian, 1992) que permite visualizar y cuantificar todas las características del habla. Tanto de los sonidos percibidos como los no percibidos. Los sonidos no percibidos pueden contener medidas de duración, intensidad y frecuencia menores a los umbrales perceptuales o poseer diferencias menores a los umbrales apenas perceptibles (JND). Por ejemplo en habla corriente, el fonema /r/ necesita para su realización que existan dos sonidos periódicos a sus adyacencias. Por ejemplo en la palabra "tres" la realización fonética es [teres]. Donde la primer vocal /e/ se realiza como vocal epentética, (ver Figura 1).

Figura 1: Representación del espectrograma (arriba), forma de onda (medio) y contornos de Energía Total (abajo, en rojo) y Frecuencia Fundamental (abajo, en azul) de la palabra "tres". Se observa la vocal epentética [e] en la transcripción y en sincronía con el resto de los gráficos

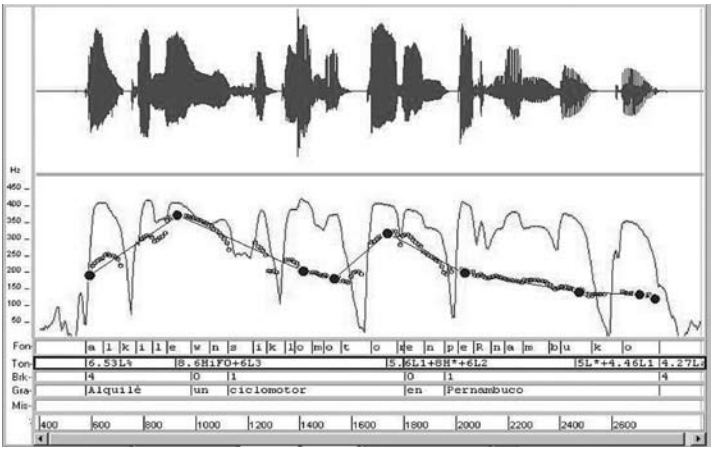


UCES Subjetividad y Procesos Cognitivos, Vol. 14, Nº 2, 2010
Pág. 89-98, ISSN impreso: 1666-244X, ISSN electrónico: 1852-7310

La cuantificación de los sonidos se realiza en función del contexto, posición en la palabra/oración y si posee acento o no. Las mediciones permiten realizar perfiles prototípicos con los cuales es posible comparar las realizaciones no convencionales respecto del valor medio. Por ejemplo, las vocales se extienden en la posición nuclear en la oración pero en algunos casos esta extensión corresponde a una pausa llena es decir una extensión exagerada de la vocal final debido a una pausa no deseada.

El análisis permite además obtener el espectro de largo plazo para evaluar la calidad vocal (voz apagada vs. brillante), medir los índices de perturbación de la voz: Jitter[1],Shimmer[1] (ambos por falta de control en la fonación), relación armónico ruido y grado de aprovechamiento de la energía. La medición del grado de fonación desde la categoría dura o tensa hasta la hiperelajada se observa en los ataques vocales con o sin escape de aire y con o sin golpe glótico en el contorno de energía. El sistema Anagraf permite la transcripción y la manipulación de los parámetros prosódicos y la síntesis con los parámetros modificados. Ver Figura 2.

Figura 2. Forma de onda de la frase "Alquile un ciclomotor en Pernambuco" (arriba), contorno de F0

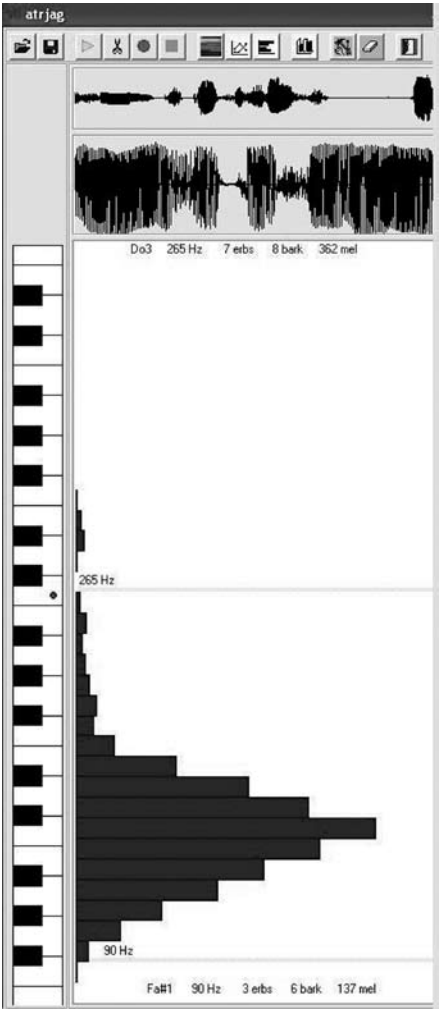


Entonación (abajo, en azul) y contorno de Energía (abajo, en rojo). Los circulitos y la línea continua indican el contorno estilizado para la síntesis. Más bajo: los niveles de transcripción: fonética, tonal, de pausas, grafémico, misceláneas, clases de palabra y sintácticos.

[1] Variación ciclo a ciclo de la frecuencia y la amplitud.

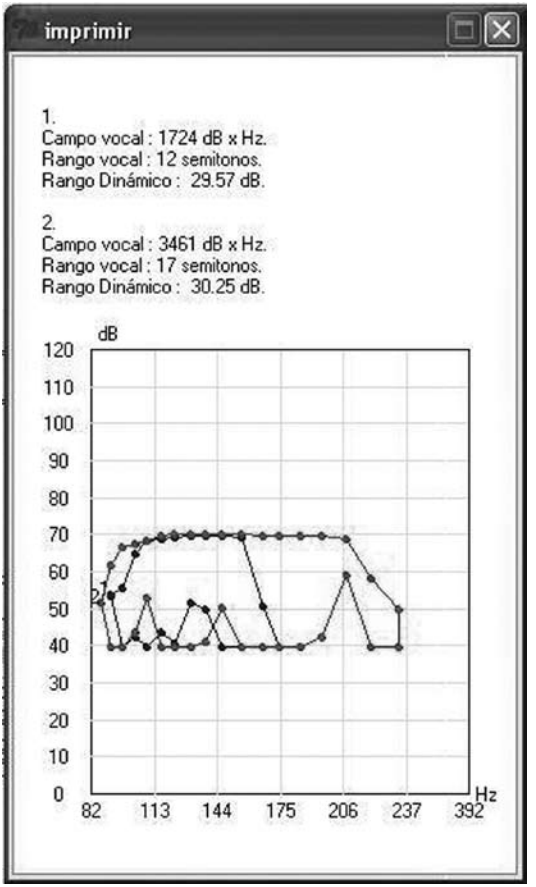
El segundo sistema de análisis llamado ATR-FONETOGRAMA realiza una elaboración y presentación particular de los datos obtenidos con el primer sistema de análisis. Permite la visualización y cuantificación de los rangos de variación de la Energía y la Frecuencia Fundamental (F0) en forma automática. La evaluación del contorno de F0 -parámetro asociado a la entonación- se realiza con el histograma. (Ver Figura 3).

Figura 3. Histograma de la Frecuencia Fundamental para un discurso prolongado de habla



El fonetograma da una información instantánea sobre el grado de expansión del tono fundamental. La comparación de las variaciones del tono fundamental asociadas a la energía, informan sobre los cambios de expresividad tonal del sujeto. Estos cambios se han vinculado con diferentes estados del paciente. Ver Figura 4.

Figura 4. Representación del Fonetograma. 1: Segmento de habla con pocos cambios tonales y 2: Segmento con mayores cambios tonales. Se cuantifica: el campo vocal (superficie), el rango vocal (F0) y el dinámico (energía)



3. Reconocimiento automático de habla

La mayoría de las herramientas de procesamiento de lenguaje natural requieren de un texto de entrada en una escala que no es fácilmente tratable en forma manual. Nuestra

propuesta consiste en obtener el texto a partir de grabaciones en forma automática, mediante la adaptación de un sistema de reconocimiento de habla ya desarrollado, que podrá ser adaptado para su aplicación en psicología.

3.1. Segmentación y etiquetado

En la etapa previa a la preparación de los modelos acústicos, se utiliza una herramienta de segmentación que permite detectar los segmentos de habla del paciente y del profesional. Las frases son segmentadas de acuerdo a las pausas largas.

Luego, se continúa la preparación de la base de datos mediante el etiquetado manual. Las etiquetas empleadas son los símbolos ortográficos y un conjunto de etiquetas asociadas a las pausas internas a las frases, pausas llenas, ruidos del paciente**.** Estas etiquetas dan información sobre las disfluencias en su doble función, dudas y auto-reparaciones que realiza el paciente. El inventario completo se presenta en el glosario.

3.2. Modelos acústicos

El equipo de investigación de Laboratorio de Investigaciones Sensoriales ha desarrollado diversos modelos acústicos para el español hablado en la Argentina. Uno de estos modelos y el modelo de lenguaje utilizado se basan en un corpus de datos registrado en condiciones de oficina de bajo ruido y con un micrófono dinámico vocal de buena respuesta en frecuencia equivalente a la relación señal a ruido de un consultorio.

Los modelos acústicos representan a los fonemas de la lengua en todos los contextos en que estos aparecen por lo que no requieren adaptaciones por el nuevo vocabulario. Las adaptaciones que se necesitan están vinculadas a las características del canal de grabación. La relación señal a ruido del ambiente y la respuesta en frecuencia de los micrófonos.

3.3. Modelos del lenguaje

Por otra parte el modelo de lenguaje debe ser adaptado a la aplicación. El modelo del lenguaje considera la estructura sintáctica y el contenido semántico del discurso. Esta información se obtendrá de las grabaciones existentes que deben ser transcriptas mediante herramientas de análisis acústico que permitan la transcripción grafémica y consideren los eventos que no constituyen palabras del diccionario.

3.4. Modelos de pronunciaciones

Estos modelos tienen en cuenta las omisiones, inserciones y cambios típicos que suceden en el habla. En principio, corresponden a un conjunto de palabras mal pronunciadas que se usan frecuentemente. Ej. 1: Obsesión: [opsesión, osesion]; Doctor: [dogtor, dotor]; Somos: [somoh]; Envío: [emBio]; Salud: [salu]. Se consideran las variantes alofónicas de los sonidos de acuerdo a la región dialectal o estilo de los hablantes. Ej. 2: Lluvia: [shuBia, iuBia, liuBia, yuBia, chuBia]

4. Conclusiones

- Las herramientas de análisis acústico pueden utilizarse por el profesional de Psicología y permiten obtener datos objetivos asociados a eventos temporales específicos de corto y largo plazo.

- El laboratorio de Investigaciones Sensoriales se halla comprendido entre los grupos científicos del CONICET capaces de realizar transferencia tecnológica a distintos profesionales e instituciones.

- Los sistemas de reconocimiento se justifican para obtener corpus de texto que serán procesados con otras herramientas de procesamiento lingüístico.

- La disponibilidad de grandes bases de datos de habla espontánea puede lograrse con la participación de profesionales interesados en este tipo de estudios y respetando los criterios de confidencialidad.

- Se invita a los profesionales de psicología a desarrollar tesis de doctorado en estas disciplinas en el ámbito del Laboratorio de Investigaciones Sensoriales.

Bibliografía

Albornoz, E.M.; Crolla, M.B. & Milone, D.H.; "Recognition of emotions in speech". *XXXIV Conferencia Latinoamericana de Informática*. Sep. de 2008, pp. 1120-1129,

Albornoz, E.M.; Milone, D.H. & Rufiner, H.L. "Multiple Feature Extraction and Hierarchical Classifiers for Emotions Recognition". *Development of Multimodal Interfaces: Active Listening and Synchrony*, Vol. 5967, 2010, pp. 242-254.

Casale, S., Russo, A., Serrano, S. (2007). "Multistyle classification of speech under stress using feature subset selection based on genetic algorithms". *Speech Communication*, 49, pp. 801-810.

France, D.; Shiavi, R.; Silverman, S.; Silverman, M. and Wilkes, D.M. (2000). "Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk". *IEEE Transactions on Biomedical Engineering*. Vol. 47, No. 7, pp. 829-837.

Guirao; M. (1980). *Los sentidos, bases de la percepción*. Madrid: Alhambra.

Gurlekian, J.A.; Franco, H.E. y Toledo, G.A. (1983). "Procesamiento de señales de habla. El hombre dialoga con la máquina". *Revista Quid,* Informe Especial, Nro. 14, pp. 119-134.

Gurlekian, J.A. (1997). El laboratorio de Audición y Habla del LIS. En: Guirao, M. (editor). *Procesos sensoriales y cognitivos*. Buenos Aires: Dunken, pp. 55-81.

Gurlekian, J.A., Colantoni, L., Torres, H., Rincon, A. and Mariño. (2001). Database for an automatic Speech Recognition System for Argentine Spanish. Proc. Of the IRCS Workshop on Linguistic databases (Bird& Liberman eds.). Pennsylvania: University of Pennsylvania, pp. 92-98.

Kramer, E. (1963). Judgment of personal characteristics and emotions from nonverbal properties of speech, Psychological Bulletin. Vol. 60(4), pp. 408-420.

Toledo, Guillermo (1988). *El ritmo en el español*. Madrid: Gredos.

Toledo, G.A. (2001). "Acentos en el español: un corpus de conversación". *Estudios de Fonética Experimental XI*, pp. 121-142.

Univaso, P.; Gurlekian, J.A. y Evin, D. (2009). "Reconocedor de habla continua para el Español de la Argentina". *Revista Clepsidra*, Enero-Junio, No. 8, pp.13-22.

6. Glosario

Símbolos para el etiquetado para uso tecnológico: Alfabeto SAMPA. i,e,a,o,u,p,t,k,b,d,g,s, f,x(j),ts(ch),m,n,J(ñ),l,r,R(rr),z(y). Alófonos frecuentes de Argentina: B,D,G,h,N,C,j,w.

Unidades:

Grafema: símbolo ortográfico. "v" Fonema: símbolo perceptual. /b/ Fono: realización acústica. [b]
Alófono: realización acústica alternativa. [B]
Frase entonativa (IP): conjunto de frases intermedias que constituyen una oración.
Frases intermedias (ip): Segmentos de habla separados por silencios o cambios tonales.

Sistema de transcripción ToBI: Acentos tonales: H*, L* y Bitonales Acentos de Frase: H-,L-
Tonos de Juntura: L%, H% Niveles de pausa: 0-4
Misceláneas Niveles sintácticos Clase de palabras

Normas de Etiquetado Grafémico: Ruidos del Locutor
No identificables: [spk]
Identificables: Incluye sonidos producidos por el hablante: tow (tos), riw (risa), esw(estornudo), jaw
(jadeo), exw (explosiones), etc. Otros ruidos
Ruidos cortos: [int]
Ruidos Prolongados: [sta] Palabras Mal Pronunciadas:
*sonido_oído palabra_correcta
Pausas Internas: sil

Pausas llenas: aah, eeh, iih, ooh, uuh, emh, mmh, imh, nnh, llh, ssh, ffh, jjh, estemh, bah, eah, euh, auh, opah, eh, ah, ahah, ayh, etc
Fragmentos, sonido Ininteligible: ** Palabras Extranjeras:
@sonido_oido palabra_correcta
Archivos cortados: -
Cuando la grabación comienza o termina cortada, o hay cortes internos en el audio, se coloca una marca al comienzo de la oración.

Fecha de recepción: 15/12/09
Fecha de aceptación: 10/05/10