Construcción de patrones prosódicos para el reconocimiento automático del habla

Enrique M. Albornoz y Diego H. Milone

Grupo de investigación en señales e inteligencia computacional
Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral

Resumen Se ha avanzado mucho en el reconocimiento automático del habla y se ha incorporado información importante a distintos niveles de análisis del habla. Por otra parte y siendo el fundamento de este trabajo, los rasgos prosódicos que presenta el habla no son generalmente ejes en estos análisis y su incorporación al reconocimiento automático del habla aún es incipiente. Esta investigación tiene como fin encontrar, dentro de estas manifestaciones prosódicas físicas, la información necesaria para mejorar el rendimiento de los sistemas de reconocimiento automático del habla. Se propone un método para caracterizar a las palabras según sus estructuras prosódicas y, aplicando modelos de lenguaje variantes en el tiempo, se utiliza esta información para desambiguar hipótesis en el proceso de reconocimiento mediante modelos ocultos de Markov.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz & D. H. Milone; "Construcción de patrones prosódicos para el reconocimiento automático del habla"
7th Argentine Symposium on Artificial Intelligence, pp. 225-236, 2005.

1. INTRODUCCIÓN

Por medio del habla, el ser humano logra el proceso de comunicación más avanzado entre los seres vivos y es precisamente esta una de las características principales que lo distinguen como el ser más inteligente.

La señal de voz es producida por el aparato fonador y se transmite mediante ondas de presión propagadas por el aire [1]. En el caso de los fonemas sonoros (o tonales), la frecuencia con que vibran las cuerdas vocales se denomina frecuencia fundamental (F_0) y su sensación auditiva es el tono de la voz o entonación, la que varía con el hablante según sea mujer o varón, adulto o niño, etc. [2].

Debido a la variabilidad temporal del tracto vocal, la señal de voz es una señal no estacionaria. Entonces, para su análisis, luego de digitalizarla y dado que no tendría sentido analizarla muestra a muestra y tampoco en períodos de varios segundos, se hace valer la hipótesis de estacionariedad por tramos de la señal en relación a la velocidad de variación de la morfología del tracto y ésta se la analiza en tramos cada 10 a 30 milisegundos.

El reconocimiento automático del habla (RAH) es una disciplina que se encarga de la concepción y realización de sistemas automáticos que convierten las señales acústicas procedentes de un locutor humano en (secuencias de) categorías lingüíticas de un universo dado. Los modelos ocultos de Markov (MOM) son la técnica más utilizada y con mejores resultados en RAH. Su ventaja principal es la capacidad de modelar bien la variabilidad temporal de la señal de voz [3].

Tabla 1. Algunos niveles de la organización jerárquica del habla.

Fonemas:	/e//s/ /u//n/ /b//a//R//k//o/
↓	
Acentuación:	/A/ /A/ /T/ /A/
↓	
Sílabas:	/es/ /un/ /bar/ /co/

Algunos investigadores han utilizado rasgos prosódicos en sistemas de traducción automática [4,5], o para detectar eventos espureos y fines de frases o palabras [6,7,8,9]. En [10] se propuso un método para incorporar información adicional a un sistema de RAH mediante la penalización adaptativa del modelo de lenguaje. Luego, se utilizó esta técnica con éxito para incorporar información acentual en un sistema de RAH continua, pero se observó una débil asociación entre el acento prosódico y el acento ortográfico, lo que imponía una cota superior en las mejoras que podían realizarse [11]. En este trabajo se propone una nueva definición de la acentuación prosódica y su aplicación al RAH continua mediante el mismo método de penalización antes citado.

A continuación se introducen brevemente la organización estructural del habla y el RAH mediante MOM. En la Sección 2 se tratan los aportes principales y algunos detalles de implementación. En la siguiente sección se discuten los resultados del método y por último se presentan las conclusiones y los trabajos con que continuará esta investigación.

1.1. Organización estructural del habla

El habla puede organizarse según distintas estructuras jerárquicas de acuerdo con el aspecto que se considere como central. La lingüítica provee de una jerarquía en base a la que se pueden desarrollar muchos otros estudios [12]. El objeto de estudio es principalmente la estructura del mensaje, despojándolo de los mecanismos que lo han generado. En este sentido, la fonética y la fonología estudian los sonidos elementales de una lengua tanto en lo que respecta a su acústica como a su función en el sistema de comunicación. En este trabajo no se considera el significado que transmiten estos sonidos y los símbolos asociados, y se analiza la prosodia a niveles de suprasegmentos y sílabas, aunque ésta abarque en su estudio niveles superiores en la estructura del habla. Las manifestaciones de los distintos niveles pueden ser unidades disímiles e independientes que ocurren sin modificar los rasgos característicos a cada nivel.

Fonemas, suprasegmentos y sílabas. En la Tabla 1 se mencionan algunos de los niveles jerárquicos de la organización estructural del habla.

Del análisis del proceso de generación y el resultado acústico se establecen modelos para los sonidos elementales del habla y se los denomina fonemas. Este es el nivel en el que pueden distinguirse las primeras unidades del habla.

Los suprasegmentos están relacionados con la expresión y representados principalmente por el acento, la cantidad y la entonación [12]. Estas estructuras poseen diversas manifestaciones físicas y sus correspondientes modelos y símbolos lingüíticos asociados. Las reglas que rigen su uso se agrupan bajo la denominación general de prosodia. El suprasegmento es una estructura de duración mayor a la de fonemas y menor a la de morfemas o palabras, que es afectada por rasgos prosódicos comunes. En este rango de tiempo se encuentra la sílaba; que no es un suprasegmento, pero se le aproxima en su duración.

Una sílaba se constituye por un núcleo sonoro o vocálico y su contexto. El núcleo generalmente es el que posee la mayor apertura articulatoria y debe permitir la extensión de su duración. La división en sílabas del español está definida por un conjunto de reglas sencillas basadas en su representación ortográfica [13]. La acentuación refiere a una representación de los suprasegmentos, en la que las distintas sílabas, según sean acentuadas o no, se caracterizan como Tónicas (T) y Átonas (A) respectivamente.

Entonación. El término entonación se utiliza, en un sentido amplio, para hacer referencia a un conjunto de fenómenos lingüíticos relacionados directamente con la F_0 de las emisiones de voz. La diversidad de niveles a los que se estudia la entonación incluye: F_0, tonema, grupo de entonación y curva melódica [14]. La F_0 se mide en cada tramo de análisis y constituye el nivel más elemental de estudio poseyendo la menor duración en el análisis. Además, es punto de partida del análisis en los niveles superiores [15].

1.2. El RAH mediante modelos ocultos de Markov

En esta aplicación el objetivo es entender como se genera y entiende naturalmente el habla, aprender de este proceso y luego poder diseñar buenos sistemas de RAH.

El RAH es un problema multidisciplinar, relacionado con: procesamiento de señales, acústica, teoría de la comunicación y de la información, estadística, matemática, lingüítica, fisiología, informática (especialmente reconocimiento de formas e inteligencia artificial), etc. Hay buenas razones para suponer que el proceso del habla se puede modelar adecuadamente como un proceso estocástico:

- El mismo *sonido/fonema/palabra* suena diferente con cada pronunciación.
- Podemos suponer que, al hablar, se transita aleatoriamente entre diferentes configuraciones del tracto vocal y en cada configuración se emiten fonemas siguiendo alguna distribución de probabilidades.

Las cadenas de Markov y los MOM son modelos estadísticos que proporcionan descripciones de secuencias de eventos. Pueden entrenarse con muchas pronunciaciones y, al decodificar, el costo computacional depende básicamente del número de modelos y no del número de pronunciaciones con que fueron entrenados [16].

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz & D. H. Milone; "Construcción de patrones prosódicos para el reconocimiento automático del habla"
7th Argentine Symposium on Artificial Intelligence, pp. 225-236, 2005.

2. DESARROLLO

2.1. Métodos propuestos

En el español, la acentuación definida por las reglas ortográficas guarda una débil relación con las manifestaciones prosódicas del habla [11]. La idea principal de este trabajo es pasar la información de la acentuación definida según las reglas ortográficas a un segundo plano y hallar relaciones claras entre los rasgos prosódicos y las palabras que se pronunciaron, para poder definir una nueva forma de clasificar las prominencias acentuales del idioma. Una vez obtenida esta forma de clasificación, podremos incluirla en un sistema RAH para mejorar el reconocimiento con un método de penalización similar al propuesto en [10]. Para alcanzar estos objetivos se plantea una técnica de clasificación basada en histogramas. El método consiste en:

1. Identificar correctamente las posiciones de los fonemas en las frases: mediante un sistema de RAH previamente entrenado y, siendo conocidas las transcripciones, se realiza una alineación forzada con el algoritmo de Viertbi.
2. Extraer los rasgos prosódicos principales (F_0, energía, duración, etc) de cada señal: directamente aplicando los métodos clásicos de análisis por tramos de señales (ventanas con 10 ms de paso y 50 ms de ancho).
3. Seccionar las frases en sílabas y asociar los valores prosódicos correspondientes: en el punto 1 se obtuvo la segmentación en palabras y fonemas y, como antes se ha mencionado, en el español la separación silábica se obtiene mediante la aplicación directa de las reglas ortográficas.
4. Calcular para cada palabra los mínimos, medias y máximos prosódicos de cada sílaba.
5. Tomar todas las ocurrencias de la misma palabra y contabilizar las distintas estructuras que se presentan.
6. Generar histogramas y clasificar las palabras según sus patrones prosódicos más característicos.
7. Obtener una estimación de la probabilidad de la palabra dado su patrón prosódico: a partir de los histogramas de cada palabra.
8. Incorporar esta información en el sistema de RAH, en el proceso de decodificación por el algoritmo de Viterbi [10].

A continuación se detallan las características del reconocedor utilizado para llevar a cabo la etapa 1.

2.2. Sistema de RAH para segmentar la base de datos

Dentro de los muchos pasos que son más bien comunes a todos los desarrollos de sistemas RAH, a continuación se comentan los más destacados y las modificaciones propias realizadas para la segmentación de la base de datos en este trabajo.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz & D. H. Milone; "Construcción de patrones prosódicos para el reconocimiento automático del habla"
7th Argentine Symposium on Artificial Intelligence, pp. 225-236, 2005.

Tabla 2. Características de la base de datos.

Total de elocuciones	1000
Total de frases con textos diferentes	500
Total de palabras	9448
Total de palabras diferentes	277
Hablantes femeninos	6
Hablantes masculinos	6

Las señales se analizaron con ventanas de 50 ms de ancho y con un paso de 10 ms. Los parámetros elegidos para ser extraídos fueron 12 coeficientes ceptrales en escala de mel (MFCC) y la energía. También se utilizaron coeficientes delta y aceleración [17].

Para la creación y diseño del prototipo MOM se seleccionaron, como es habitual para fonemas, modelos de 5 estados. Las variables estadísticas que modelan la distribución de probabilidades de observación de estos estados con respecto a los coeficientes de las distintas frases de entrenamiento fueron *media* y *varianza*, modelando así con gaussianas esféricas en $\mathbb{R}^{39}$. Para la fase de entrenamiento del MOM se provee al modelo con las distintas frases parametrizadas y sus transcripciones.

Finalmente, a partir de las transcripciones de las frases y las propias señales de voz se obtienen las segmentaciones en palabras y fonemas mediante alineación forzada por el algortimo de Viterbi [18].

2.3. Herramientas y materiales

Para el desarrollo e implementación de los MOM se utilizó un conjunto de herramientas denominado *Hidden Markov Toolkit* (HTK)[1] [17]. Se utilizaron rutinas del ToFy[2] para la extracción de energía y F_0 de las señales, esta última calculada con un algoritmo similar al de Noll [19], basado en coeficientes ceptrales. Otras rutinas para el cálculo y las estadísticas fueron implementadas en *GNU C++* y *Free Pascal*.

Las frases utilizadas fueron extraidas de la base de datos Albayzin [20], creada por cinco Universidades españolas. Ésta se desarrolló con el objetivo de contribuir al desarrollo y la evaluación de sistemas de reconocimiento y procesamiento del habla. Los hablantes pertenecen a la variedad central del castellano, en su mayor parte de las comunidades de Castilla-La Mancha, Castilla-León, Cantabria y Madrid, con mujeres y varones de entre 18 y 55 años de edad. En la Tabla 2 se muestran los datos del subconjunto utilizado de esta base de datos.

[1] Desarrollado en el Speech and Vision Robotics Group en la Universidad de Cambridge, disponible en `http://htk.eng.cam.ac.uk`

[2] Desarrolladas en el Laboratorio de Cibernética de la Universidad Nacional de Entre Rios (Argentina), disponible en `http://www.milone.tk`.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz & D. H. Milone; "Construcción de patrones prosódicos para el reconocimiento automático del habla"
7th Argentine Symposium on Artificial Intelligence, pp. 225-236, 2005.

3. RESULTADOS Y DISCUSIONES

3.1. Clasificación según rasgos prosódicos

Las palabras fueron preseleccionadas, eligiendo aquellas que tenían un número suficientemente alto de ocurrencias en la base de datos utilizada. Antes de continuar es menester explicar el proceso de clasificación de las palabras, introducido en la sección anterior. Una vez asociada la palabra a sus rasgos prosódicos, se calculan para éstos los valores máximos, mínimos y medios de cada sílaba para todos los sucesos de la palabra en la base de datos. Tomando los sucesos de cada palabra por separado y evaluando un rasgo particular se obtienen clases codificadas en n dígitos, siendo n el número de sílabas de la palabra. El código indica en forma relativa la magnitud medida (por ejemplo: máximo de F_0) para cada sílaba. Paso siguiente, se contabilizan los sucesos de cada clase que poseen las palabras. A modo de ejemplo, la clase 321 (para palabras de tres sílabas) indica que la primer sílaba tiene valor máximo y la última valor mínimo; así la clase 213 indica que el valor máximo está en la última sílaba y mínimo en la segunda.

En primera instancia se puede ver que cada palabra, para sus distintos sucesos, puede pertenecer a alguna de las $n!$ clases que se forman de intercambiar las distribuciones de las cantidades dentro de la palabra. Afortunadamente para este método, casi todas las palabras pertenecen a unas pocas clases y están caracterizadas en su mayoría por una única clase.

Para ejemplificar, se presentan algunas gráficas de valores relativos con resultados de la caracterización para distintos rasgos prosódicos. Cabe mencionar que, para simplificar las gráficas, se han eliminado las clases para las que una palabra tiene cero sucesos.

En la Fig. 1 se observan las clases que caracterizan la palabra *dime* para la media de energía. Aquí se caracteriza completamente a la palabra con la clase *12* para 256 palabras computadas y se interpreta como: *la palabra dime se caracteriza por tener un valor mayor de energía media en la segunda sílaba.*

En la Fig. 2 se ven las clases que definen la palabra *longitud* para el rasgo prosódico mínimo de energía. Con 134 palabras computadas, se ve claramente que la clase *321* define completamente esta palabra.

En la Fig. 3 se observan las clases para la media de F_0 de la palabra *valenciana*, que queda caracterizada por la clase *4321* para 34 palabras computadas.

También para este método se ven palabras que no son clasificadas por ninguno de los rasgos prosódicos propuestos. En la Fig. 4 se observa que para la palabra *cúbicos* (con 62 sucesos computados) no se encontró una clase, de al menos un rasgo prosódico, que la caracterice. Un caso similar se presenta con la palabra *comunidad*, en la Fig. 5 se ven las clases para el rasgo mínimo de energía, con 256 sucesos computados de la palabra. En estos casos los rasgos prosódicos no aportarían ninguna información importante al sistema de RAH.

Se puede observar un resumen de los resultados en la Tabla 3. La columna *Total de palabras* da cuenta de la cantidad total de palabras diferentes que posee la base de datos; la columna *Palabras evaluadas* muestra cuántas palabras del

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz & D. H. Milone; "Construcción de patrones prosódicos para el reconocimiento automático del habla"
7th Argentine Symposium on Artificial Intelligence, pp. 225-236, 2005.

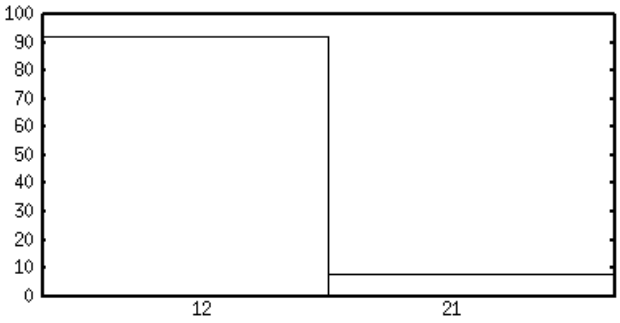

Figura 1. Clases para la palabra *dime*

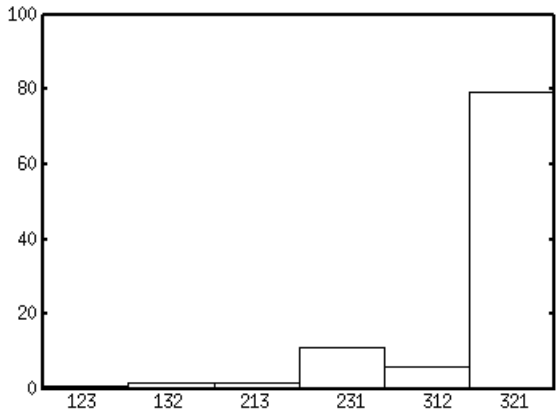

Figura 2. Clases para la palabra *longitud*

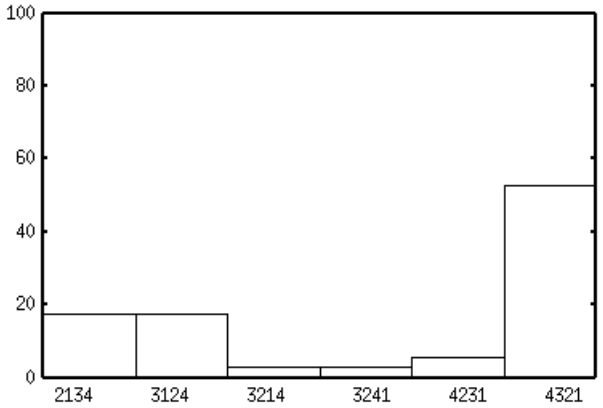

Figura 3. Clases para la palabra *valenciana*

Figura 4. Clases para la palabra *cúbicos*

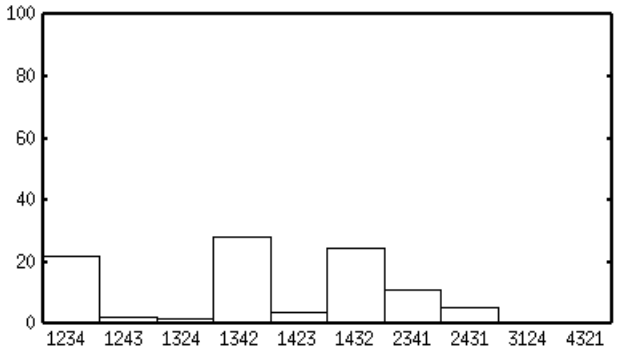

Figura 5. Clases para la palabra *comunidad*

Tabla 3. Porcentajes de caracterización

Cantidad de sílabas	Total de palabras evaluadas	Palabras diferentes	Porcentaje	Diferencia
2 sílabas:	1646	25	96.00 %	≥80 %
3 sílabas:	398	6	83.33 %	≥40 %
4 sílabas:	792	9	77.78 %	≥30 %
5 sílabas:	196	1	100.00 %	≥20 %

total cumplieron los requisitos y se evaluaron; la columna *Porcentaje* muestra cuántas palabras este método puede clasificar correctamente. También se ve que la capacidad de discriminación del método decrece marcadamente con la cantidad de sílabas, esto puede deberse a que no contamos con palabras de muchas sílabas que son representables por sus rasgos prosódicos o bien el método no es eficiente para esas cantidades de sílabas. Esto último podría comprobarse analizando una base de datos más extensa. La columna *Diferencia* indica la diferencia relativa suficiente que debe presentar cada palabra en la clase que la caracteriza para ser considerada distinguible por el método. Vimos que acorde al incremento del número de sílabas de la palabra, se incrementa la cantidad de clases a la que puede pertenecer la palabra y entonces se debe reducir la tolerancia en las diferencias. Una característica interesante detectada en el análisis y que se da más frecuentemente en las palabras de tres o más sílabas, es que aparecen 2 o 3 clases dominantes en las palabras para un parámetro prosódico determinado, esto es, de las $n!$ clases que caracterizan a una palabra hay 2 o 3 que claramente se destacan de las otras. Esta última observación puede ser muy interesante desde el punto de vista que restringe también la caracterización de una palabra, si bien no a una, a varias clases prosódicas bien definidas.

3.2. Reconocimiento con información prosódica

Para la experimentación se ha utilizado como sistema de referencia un reconocedor del habla continua con idénticas características al descripto en la Sección 2.2 y con la única diferencia de que en el procesamiento de señales se utilizó una ventana de 25 ms de ancho.

Para ilustrar las mejoras obtenidas mediante el método de penalización adaptativa del modelo de lenguaje, con los rasgos prosódicos propuestos en este trabajo, se presentan a continuación algunas de las frases reconocidas. Se pueden identificar tres tipos de resultados:

- Tipo 1 “mejoras totales”: dan una solución completa y permiten un reconocimiento 100 % correcto de la frase.
- Tipo 2 “mejoras parciales”: dan una solución en cuanto proponen qué hipótesis de palabra no es correcta, aunque no corrigen todos los errores de reconocimiento en la frase.
- Tipo 3 “mejoras indirectas”: entre estos casos, por ejemplo, se incluyen situaciones en que, si bien no se corrige una palabra prosódicamente incorrecta,

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz & D. H. Milone; "Construcción de patrones prosódicos para el reconocimiento automático del habla"
7th Argentine Symposium on Artificial Intelligence, pp. 225-236, 2005.

se selecciona correctamente entre varias hipótesis de la misma palabra pero con distintos tiempos de inicio y fin, lo que indirectamente beneficia al resto de la red de palabras.

Ejemplos del Tipo 1

En la frase *bxge3113*, cuya transcripción correcta es:

`Dime el nombre de los mares que bañan la Comunidad de Andalucía.`

Mientras que el reconocedor sin información prosódica reconoce:

`Dime el nombre de los mares que baña la Comunidad de Andalucía.`

El reconocedor con prosodia reconoce correctamente debido a que la palabra `baña` es penalizada en la red y la hipótesis de `bañan` pasa a una tener mayor probabilidad.

Una corrección similar se realiza en *euge0139*, cuya transcripción correcta es:

`Dígame si hay algún río que pase por tres Comunidades Autónomas.`

Mientras que el reconocedor sin prosodia reconoce:

`Dígame segundo río que pase por tres Comunidades Autónomas.`

El reconocedor con prosodia reconoce correctamente debido a que la palabra `segundo` es penalizada.

En la frase *ruge0199*, cuya transcripción correcta es:

`Dime los nombres de los ríos con más de cien kilómetros de longitud.`

Mientras que el reconocedor sin prosodia reconoce:

`Dime el nombre de los ríos con más de cien kilómetros longitud.`

El reconocedor con prosodia reconoce correctamente, penaliza a `nombre` y soluciona la extensión de `kilómetros`, que con una buena segmentación permite incorporar a la palabra `de`.

Ejemplos del Tipo 2:

Para *ilge0071*, cuya transcripción correcta es:

`Hay algún río que nazca y desemboque en la misma Comunidad.`

Mientras que el reconocedor sin prosodia reconoce:

`Caudal de un río que nazca y desemboque en la misma Comunidad.`

El reconocedor con prosodia reconoce:

`Cada algún río que nazca y desemboque en la misma Comunidad.`

Aquí se ve que el descartar la hipótesis de `caudal` ayuda a corregir otra palabra y un error por inserción.

Para el archivo *ikge0064*, cuya transcripción correcta es:

`En qué comunidad se halla el Duero?`

Mientras que el reconocedor sin prosodia reconoce:

`Que en que comunidad se halla el Ebro?`

El reconocedor con prosodia reconoce:

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz & D. H. Milone; "Construcción de patrones prosódicos para el reconocimiento automático del habla"
7th Argentine Symposium on Artificial Intelligence, pp. 225-236, 2005.

Que en que comunidad se halla el Duero?

Aquí se ve que al penalizar la hipótesis de `Ebro` resulta más probable la palabra `Duero`.

Ejemplo del Tipo 3:

Para la frase *nxge3161*, cuya transcripción correcta es:

`Lugar donde desemboca el Jucar.`

Mientras que el reconocedor sin prosodia reconoce:

`Lugar donde desemboca Jucar`

El reconocedor con prosodia reconoce correctamente, dado que la penalización de hipótesis mal ubicadas temporalmente de `desemboca` permite que se haga más probable la hipótesis que tiene a esta palabra con la ubicación correcta en la frase y así deja lugar para insertar la palabra `el`.

4. CONCLUSIONES Y TRABAJOS FUTUROS

Se ha presentado un método de caracterización de las distintas palabras en base a histogramas, discriminadas por grupos según su separación silábica y luego clasificadas por su estructura prosódica. Se hace evidente que esta caracterización distingue a las palabras y brinda esa información relevante buscada, que permite mejorar el desempeño de un sistema de RAH.

Como trabajos futuros se preve extender la técnica de histogramas a la combinación de éstos, tomando de a dos o más variable prosódicas y formando nuevos histogramas a partir de todas las combinaciones posibles de medidas de estas variables. Otra propuesta para clasificar mejor a las palabras es extraer más información prosódica de las palabras, como *cadencias, anticadencias y mesetas* de entonación [12].

Como tarea pendiente se planea extender los experimentos de reconocimiento a otros corpus de habla, con diferentes acentos regionales, en condiciones de ruido y, más a largo plazo, realizar estos mismos estudios en otros idiomas.

Referencias

1. Deller, J.R., Proakis, J.G., Hansen, J.H.: Discrete-Time Processing of Speech Signals. Macmillan Publishing, NewYork (1993)
2. Manrique, A.M.B.: Manual de Fonéica Acústica. Hachette, Buenos Aires (1980)
3. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press, Cambrige, Masachussets (1999)
4. Buckow, J., Batliner, A., Huber, R., Nöth, E., Warnke, V., Niemann, H.: Dovetailing of acoustic and prosody in spontaneous speech recognition. In: Proceedings of 5th International Conference on Spoken Language Processing. (1998) Prosody and Emotion 2.
5. Nöth, E., Batliner, A., Kießling, A., Kompe, R., Niemann, H.: Verbmobil: The use of prosody in the linguistic components of a speech understanding system. IEEE Trans. on Speech and Audio Processing **8** (2000) 519–532

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz & D. H. Milone; "Construcción de patrones prosódicos para el reconocimiento automático del habla"
7th Argentine Symposium on Artificial Intelligence, pp. 225-236, 2005.

6. Rajendran, S., Yegnanarayana, B.: Word boundary hypothesization for continuous speech in Hindi based on F0 patterns. Speech Communication **18** (1996) 21–46
7. Hirose, K., Iwano, K.: Accent type recognition and syntactic boundary detection of japanese using statistical modeling of moraic transitions of fundamental frequency contours. In: Proceedings of the IEEE 23rd International Conference on Acoustics, Speech and Signal Processing. Volume 1., Seattle (1998) 25–28
8. Lee, S.W., Hirose, K.: Dynamic beam-search strategy using prosodic-syntactic information. In: Workshop on Automatic Speech Recognition and Understanding. (1999) 189–192
9. Stolcke, A., Shriberg, E., Hakkani-Tür, D., Tür, G.: Modeling the prosody of hidden events for improved word recognition. In: Proceedings of the 7th European Conference on Speech Communication and Technology. Volume 1. (1999) 311–314
10. Milone, D.H., Rubio, A.J., López-Cózar, R.: Modelos de lenguaje variantes en el tiempo. In: Memorias del XXIV Congreso Nacional de Ingeniería Biomédica, Oaxtepec, México, SOMIB (2001)
11. Milone, D.H., Rubio, A.J.: Prosodic and accentual information for automatic speech recognition. IEEE Trans. on Speech and Audio Proc. **11** (2003) 321–333
12. Quilis, A.: Tratado de Fonología y Fonética Españolas. Biblioteca Románica Hispánica. Editorial Gredos, Madrid (1993)
13. Llorach, E.A.: Gramática de la Lengua Española. Real Academia Española. Colección Nebrija y Bello. Editorial Espasa Calpe, Madrid (1999)
14. Almiñana, J.M.G.: Modelización de Patrones Melódicos del Español para la Síntesis y el Reconocimiento del Habla. Servei de Publicacions de la Universitat Autònoma de Barcelona, Facultad de Filosofia i Lletres, Departament de Filologia Espanyola, Barcelona (1991)
15. Rabiner, L.R., Gold, B.: Theory and Application of Digital Signal Processing. Prentice Hall (1975)
16. Rabiner, L.R., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall (1993)
17. Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book (for HTK Version 3.1). Cambridge University Engineering Department., Cambridge, Inglaterra. (2001)
18. Ney, H., Ortmanns, S.: Dynamic programming search for continuous speech recognition. IEEE Signal Processing Magazine **16** (1999) 64–83
19. Noll, A.M.: Cepstrum pitch determination. The Journal of the Acoustical Society of America **41** (1967) 179–195
20. Moreno, A., Poch, D., Bonafonte, A., E.Lleida, J.Llisterri, J.B.Marino, Nadeu, C.: Albayzin speech data base: design of the phonetic corpus. In: Proceedings of the 2th European Conference of Speech Communication and Technology, Berlin (1993) 175–178

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz & D. H. Milone; "Construcción de patrones prosódicos para el reconocimiento automático del habla"
7th Argentine Symposium on Artificial Intelligence, pp. 225-236, 2005.