RECONOCIMIENTO DE LOS NUMEROS CASTELLANOS MEDIANTE SEMISILABAS[1]

J. B. Mariño*, C. Nadeu*, A.Moreno*, E. Lleida*, I. Hernáez**, E. Monte*

*Dpto. Teoría de la Señal y Comunicaciones. [2]
Universidad Politécnica de Catalunya.

** Dpto. de Tecnología Electrónica.
Universidad del País Vasco.

RESUMEN

En esta comunicación se describe el uso de la semisílaba en el reconocimiento de habla continua para una aplicación específica: el reconocimiento de los números enteros castellanos del cero al mil. Tras una breve descripción de la arquitectura del sistema de reconocimiento, se detalla la inferencia de la gramática de estados finitos que representa los números en términos de semisílabas, y se indica el procedimiento seguido para la generación de las referencias de las mismas. Finalmente, se presentan los resultados alcanzados en dos experimentos: en el primero el sistema de reconocimiento es entrenado para un locutor y las referencias utilizadas para las semisílabas son patrones de características frecuenciales; en el segundo, el entrenamiento es multilocutor y las semisílabas son representadas mediante Modelos Ocultos de Markov. En ambos casos la tasa de reconocimiento del sistema es excelente.

---

[1]Este trabajo ha sido financiado parcialmente por el PRONTIC, proy. nº 105/88.

[2]Dirección postal: E.T.S.I. Telecomunicación, Aptdo. 30002, 08080-Barcelona.

I.- INTRODUCCION

El reconocimiento de los números es un experimento de interés en la investigación del reconocimiento del habla continua. Además de su potencial aplicación, el reconocimiento de los números reproduce a escala reducida el problema general del reconocimiento del habla continua: ligeras diferencias en los sonidos pueden corresponderse con grandes diferencias en el significado semántico, los números exhiben una fuerte estructura gramatical y hay una relativamente importante variedad de sonidos en la realización acústica de los números.

Aunque en esta aplicación concreta del reconocimiento del habla, las palabras podrían haber sido una elección posible como unidad fonética para el reconocimiento, el sistema descrito en esta comunicación esta basado en semisílabas [1]. Ello está motivado en nuestro interés de construir un sistema cuya arquitectura fuese adecuada al reconocimiento del habla continua en general, y en la conjetura de que la semisílaba es una unidad fonética para el reconocimiento que se ajusta convenientemente a las características de la Lengua Castellana. Esta conjetura se fundamenta en dos hechos: a) las reglas de silabificación del Castellano están bien establecidas [2]; y b) el inventario de semisílabas en Castellano es relativamente reducido (su número no supera las 750) [3].

A fin de definir el conjunto de semisílabas, cada sílaba fue partida por la vocal fuerte en una semisílaba inicial y una semisílaba final. En nuestra definición, el acento prosódico fue incorporado a la semisílaba final. En consecuencia, distinguimos entre semisílabas finales átonas y tónicas. En Castellano los correlatos básicos del acento prosódico son el tono, la energía y la duración; puesto que las características de frecuencia fundamental y energía no son contempladas en nuestro sistema, la principal diferencia entre las semisílabas finales átonas y tónicas será la longitud de sus refererencias.

II.- DIAGRAMA DE BLOQUES DEL SISTEMA DE RECONOCIMIENTO

En la figura 1 se proporciona una descripción general de la arquitectura del sistema de reconocimiento. La señal es filtrada por un filtro paso bajo de

antialiasing con frecuencia de corte a 3.4 kHz y, posteriormente, muestreada a 8 kHz; a continuación, mediante un algoritmo adecuado se detecta el principio y final de la señal vocal, que pasa a ser parametrizada por un filtro de predición lineal de 8 coeficientes [4]. Tratada de este modo, la señal entra en el algoritmo de reconocimiento; en esencia, este algoritmo realiza programación dinámica para unidades conectadas en un solo paso (descrito para palabras conectadas en la referencia [5]) dirigido por una gramática de estados finitos [6]. Esta gramática suministra de un modo conveniente la transcripción en semisílabas de cada item en el vocabulario o lenguaje que se desea reconocer. El algoritmo de programación dinámica calcula la distancia entre los patrones [7] que representan las semisílabas y la señal vocal a reconocer (o determina la probabilidad de que ésta sea representada por los distintos Modelos Ocultos de Markov [8]), y decide la secuencia de semisílabas (permitida por la gramática) que más se ajusta a la señal oral. Si fuese necesario, un diccionario puede proporcionar el significado semántico de la secuencia decidida

Esta arquitectura general puede ser orientada a un aplicación particular mediante los adecuados diseño de la gramática regular y entrenamiento de las referencias de las semisílabas. En lo que sigue nos referiremos a estas tareas relativas a los números enteros castellanos.

III.- INFERENCIA DE LA GRAMATICA

Consideremos los vocabularios que se indican en la figura 1. El vocabulario D corresponde a los dígitos excluyendo el cero; el vocabulario DC incluye los enteros del diez al diecinueve y los múltiplos de diez; el vocabulario DCI está constituído por los prefijos necesarios para formar los enteros del veintiuno al veintinueve (2.), del treinta y uno al treinta y nueve (3.) y así sucesivamente; el vocabulario C proporciona las centenas y, finalmente, el vocabulario U reune los números especiales cero, cien y mil. El vocabulario IC de los enteros del uno al noventa y nueve puede ser obtenido mediante la siguiente combinación de vocabularios:

$$IC = (D + DC) + (DCI * D)$$

donde el signo + significa unión de vocabularios y el símbolo * representa el producto cartesiano (cada item en DCI puede ser seguido por cada uno de los item en D). Obtenido IC, el vocabulario I de los números enteros del cero al mil puede alcanzarse mediante las operaciones siguientes:

$$I = U + IC + C * IC$$

La gramática G, correspondiente al vocabulario I, puede ser establecida a partir de las gramáticas de los vocabularios D, DC, DCI, C y U mediante el uso de un conjunto de herramientas *software* desarrolladas en nuestro laboratorio [9].

A fin de representar cada elemento de estos vocabularios básicos como cadenas de semisílabas, se llevó a cabo una transcripción fonética standard, incluyendo las más frecuentes variaciones alofónicas; de este modo, un item puede haber sido representado por varias cadenas de semisílabas diferentes. Además, las combinaciones DCI*D y C*IC de vocabularios introducen nuevas combinaciones de sonidos, cuya coarticulación debe ser considerada. Nuestra herramienta para inferir gramáticas regulares resuelve este problema muy eficientemente, dado que puede cambiar una subsecuencia de semisílabas por otra o añadir una nueva subcadena de semisílabas de forma sencilla; la primera operación es necesaria cuando se supone que el efecto de coarticulación que se está considerando, ocurre siempre que aparece la combinación de sonidos en cuestión; la segunda, cuando el efecto coarticulatorio no tiene que producirse con seguridad. En la tabla 2 se ofrecen algunos ejemplos de los efectos de coarticulación más importantes considerados.

La gramática final, correspondiente al vocabulario I de los enteros castellanos del cero al mil (ambos incluídos), requiere un total de 118 estados para representar los diversos contextos en que puden aparecer las 67 semisílabas necesarias en la presente aplicación.

IV.- ENTRENAMIENTO DE LAS REFERENCIAS

El entrenamiento de las referencias se llevó a cabo a partir del material acústico proporcionado por diez locutores, que pronunciaron una sola vez un conjunto (común para todos los locutores) de 44 números enteros. Este conjunto fue diseñado de forma que la aparición de las semisílabas en él

fuese proporcionada a su frecuencia de aparición en los enteros y garantizando que cada una de las 67 semisílabas apareciera al menos dos veces. Estas 440 señales fueron segmentadas mediante un proceso de reconocimiento llevado a cabo con referencias creadas en un experimento previo [10]; esta segmentación fue supervisada y, cuando fue necesario, corregida a mano.

Del conjunto de diez locutores, uno de ellos fue seleccionado para entrenar el sistema de reconocimiento con dependencia del locutor. A partir de los 44 números pronunciados por este locutor, se crearon patrones para representar las semisílabas por el procedimiento que se describe a continuación. Los 44 números proporcionaron un conjunto de muestras para cada semisílaba, cuyo centroide se determinó por un procedimiento de *clustering* [7]; seguidamente, cada muestra fue alineada temporalmente mediante programación dinámica al centroide, para generar, mediante promediado de todas las muestras alineadas, un único patrón representante de la semisílaba.

Después de algunos experimentos de reconocimiento iniciales, se concluyó que en tal procedimiento debían ser introducidas algunas modificaciones menores. Se observó que la longitud de los patrones en general no se correspondía con la duración media de las semisílabas en el material de test. Además, los representantes de algunas semisílabas tenían dificultades en determinados contextos; los casos más importantes fueron los siguientes: a) las semisílabas finales de sílabas abiertas cuando aparecen en el final absoluto del número, y b) las semisílabas iniciales de sílabas que comienzan por una consonante fricativa ([s],[θ]) o nasal ([n]); cuando la sílaba fue la primera del número.

A fin de encarar estos dos problemas, se adoptaron las estrategias siguientes. Primero, la cantidad media de los sonidos castellanos fue establecida a partir de estudios existentes [11,12]; después, la duración de las semisílabas fue determinada sumando la duración individual de cada uno de los sonidos de la misma; por este medio, se normalizó la longitud de los patrones para una articulación natural a razón de 5 ó 6 sílabas por segundo. Este procedimiento de síntesis proporcionó valores para las longitudes de las semisílabas en buen acuerdo con el material de test; adicionalmente se encontró una explicación parcial para el inadecuado comportamiento de las semisílabas en los contextos indicados

anteriormente: en tales situaciones la duración de un sonido determinado en la semisílaba era muy diferente a la que se producía en los demás casos.

La segunda estrategia está relacionada con el promediado de las muestras de las semisílabas para generar los patrones. Las muestras de una semisílaba con problemas de contexto fueron desdobladas en dos conjuntos; en uno de ellos se recogieron las muestras que aparecían en el contexto con problemas; el otro quedó constituído por el resto de las muestras de la semisílaba. Una vez que estos conjuntos fueron establecidos, el procedimiento de entrenamiento de los patrones se prosiguió como se ha descrito previamente, de forma que se generó un patrón para cada contexto. Solamente 7 semisílabas necesitaron este patrón adicional.

En el entrenamiento del sistema de reconocimiento con independencia del locutor, se utilizaron Modelos Ocultos de Markov como referencias para las semisílabas. Cada modelo fue entrenado (siguiendo el algoritmo de Baum-Welch [8]) a partir de las muestras de la semisílaba recogidas de los 10 locutores y distinguiendo los mismos contextos que en el entrenamiento dependiente de locutor. Cuando el número de muestras por locutor para una determinada semisílaba y contexto excedió de cinco, se realizó un proceso de *clustering* para seleccionar cinco representantes; de este modo el número mayor de muestras que se utilizó para entrenar un Modelo fue de cincuenta. El número de estados de cada Modelo fue seleccionado de modo que maximizase la probabilidad de que el Modelo generase las muestras de entrenamiento.

V.- RESULTADOS

Las prestaciones del sistema de reconocimiento fueron analizadas mediante cuatro sesiones en las que se probaron ambas versiones (dependiente e independiente de locutor) del mismo. Las señales vocales fueron obtenidas directamente del locutor por un micrófono vr-230 de Shure (el mismo que fue utilizado para grabar el material de entrenamiento en una sesión previa) en una sala silenciosa; el reconocimiento fue realizado *on line* . En cada sesión la version dependiente de locutor fue probada por el locutor que la entrenó, y la versión independiente de locutor fue probada por dos locutores distintos

de aquellos que proporcionaron el material de entrenamiento y diferentes para cada sesión. Cada locutor pronunció cincuenta números distintos, seleccionados siguiendo dos criterios complementarios: una mitad fue elegida a fin de comprobar los efectos de coarticulación considerados en la fase de entrenamiento, y la otra mitad fue determinada aleatoriamente. El estilo de articulación (juzgado por dos oyentes diferentes en cada sesión) fue considerado equivalente al mantenido en la conversación humana. La velocidad de articulación de los números enteros varió entre 4 a 7 sílabas por segundo, como consecuencia de la diferente longitud de los mismos (cuanto mayor es el número de sílabas de un entero, mayor es la velocidad de articulación).

Los resultados obtenidos fueron excelentes. En la versión dependiente de locutor no se produjeron errores; el único medio para provocar reconocimientos erróneos fue alterar la velocidad de articulación natural: relantizándola excesivamente o articulando a velocidad exagerada. La versión independiente de locutor presentó una tasa de reconocimiento correcto superior al 95%.

VI.- CONCLUSION

Esta comunicación resume un experimento cuyos resultados permiten proponer la utilización de la semisílaba para el reconocimiento del habla continua Castellana, al menos en aplicaciones no excesivamente complejas. Se ha mostrado que el reconocimiento de los números enteros puede ser llevado a cabo mediante esta unidad de reconocimiento con resultados satisfactorios. De todos modos, a fin de establecer conclusiones más generales es preciso mayor experimentación.

VII.- REFERENCIAS

[1] A.E. Rosenberg et al., "Demisyllable based isolated word recognition", IEEE Trans. ASSP-31, pp 713-726: Junio, 1983

[2] J. Alcina and J.M. Blecua, "Gramática española", Ed. Ariel: 1983

[3] J. Romano, "Un sistema automático de síntesis de habla mediante semisílabas", memoria interna Universidad Técnica de Munich (RFA): 1985

[4] J.D. Markel y A.H. Gray, "Linear prediction of Speech", Springer-Verlag: 1976

[5] H. Ney, "The use of an one-stage dynamic programming algorithm for connected word recognition", IEEE Trans ASSP-32, pp 263-271: Abril, 1984

[6] J. B. Mariño et al., "Finite state grammar inference for connected word recognition", Proc. EUSIPCO'88, pp 1035-1038: Septiembre, 1988

[7] L. R. Rabiner y S.E. Levinson, "Isolated and Connected Word Recognition-Theory and selected Applications", IEEE Trans COM-29, pp 621-647: Mayo, 1981

[8] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in Speech Recognition", Proceed. IEEE vol. 77, pp 257-286: Febrero 1989

[9] I. Centeno, "Reconocimiento de los números telefónicos mediante semisílabas", Proyecto Fin de Carrera E.T.S.I. Telecomunicación, U.P.C.: 1988

[10] I. Hernáez, "Desarrollo de un sistema de reconocimiento de palabras conectadas con aplicación al reconocimiento de los números a partir de semisílabas", Proyecto Fin de Carrera E.T.S.I. Telecomunicación, U.P.C.: 1987

[11] T. Navarro Tomás, "Cantidad de las vocales acentuadas", RFE vol-3, pp 387-407: 1916

[12] T. Navarro Tomás, "Diferencias de duración entre las consonantes españolas", RFE vol-4, pp 367-393: 1918

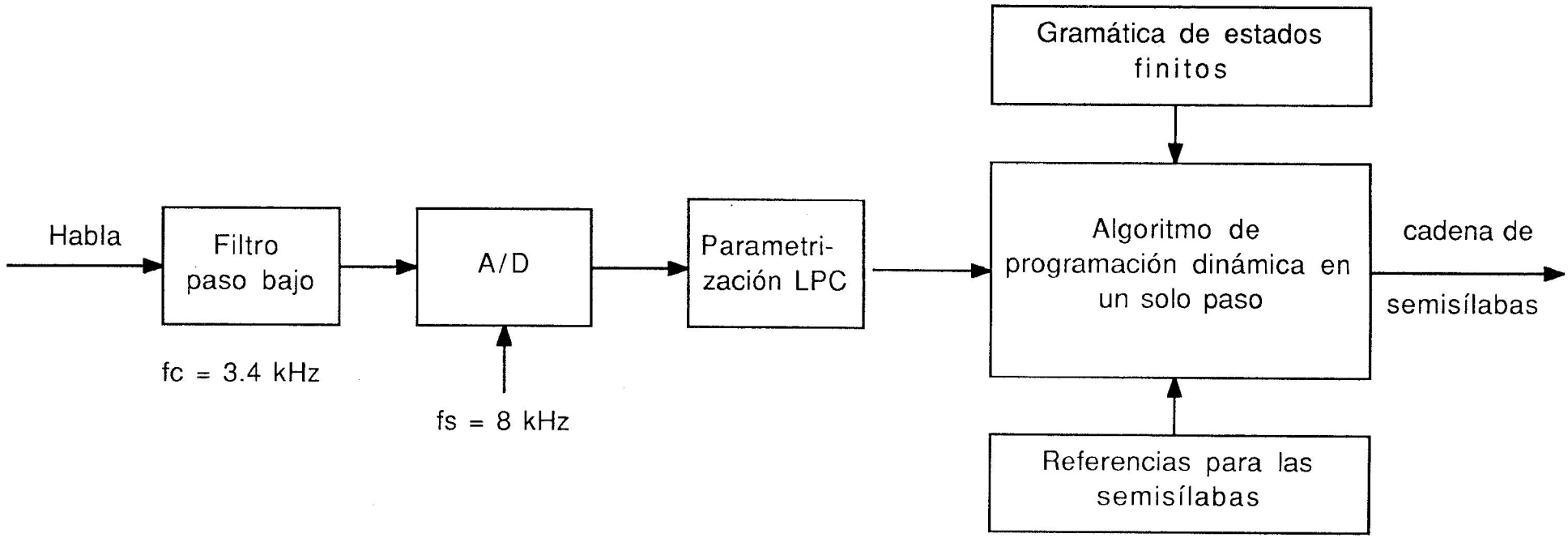

Figura 1.- Arquitectura del sistema de reconocimiento

D = {uno (1), dos (2), tres (3), kwatro (4), θiŋko (5), sejs (6), sjete (7), otʃo (8), nweβe (9)}

DC = {djeθ (10), onθe (11), doθe (12), treθe (13), katorθe (14), kinθe (15), djeθisejs (16), djeθisjete (17), djeθjotʃo (18), djeθinweβe (19), bejnte (20), trejnta (30), kwarenta (40), θiŋkwenta (50), sesenta (60), setenta (70), otʃenta (80), noβenta (90)}

DCI = {bejnti (2.), trejnta i (3.), kwarenta i (4.), θiŋkwenta i (5.), sesenta i (6.), setenta i (7.), otʃenta i (8.), noβeŋta i (9.)}

C = {θjento (1..), dosθjentos (2..), tresθjentos (3..), kwatroθjentos (4..), kinjentos (5..), sejsθjentos (6..), seteθjentos (7..), otʃoθjentos (8..), noβeθjentos (9..)}

U = {θero (0), θjen (100), mil (1000)}

Tabla 1.- Vocabularios básicos para generar los enteros castellanos del cero al mil.

sonorizacion de la s ante consonantes sonoras realización aproximante de oclusivas sonoras	...tos - dos θjen - to - bejn - te	...toz - ðos θjen- to - βejn - te
sinalefa	θjen - to - o - ʃfo bejn - ti - u - no	θjen - to - tʃo bejn - tju - no
resilabificacion	...tos - u - no ...tos - sje - te	...to - su - no ...to - sje - te

Tabla 2.- Ejemplos de los efectos de coarticulación considerados (los guiones separan sílabas).