DECODIFICACION ACUSTICO-FONÉTICA MEDIANTE PLANTILLAS SUBLÉXICAS*

P. Aibar†
F. Casacuberta
E. Vidal

Departamento de Sistemas Informáticos y Computación.
Universidad Politécnica de Valencia,
España

Resumen

El Alineamiento Temporal No Lineal es una de las técnicas más importantes utilizadas en el reconocimiento de palabras aisladas y conectadas. No obstante, su aplicación a la Decodificación Acústico-Fonética no ha sido suficientemente explorada.

En esta aproximación, las unidades léxicas o subléxicas están modelizadas mediante plantillas o secuencias de longitud variable de vectores de características. El aprendizaje de dichos modelos está basado en un proceso de segmentación de las muestras de entrenamiento (palabras o frames), seguido de uno de agrupamiento. El objetivo de esta última fase es el de seleccionar aquellas plantillas más representativas de la unidad correspondiente.

En este trabajo se presenta el problema de la Decodificación Acústico-Fonética de frases castellanas mediante Alineamiento Temporal No Lineal. Los experimentos realizados serán dependientes e independientes del locutor y se estudiará la influencia de un número variable de plantillas por unidad subléxica.

Introducción

El objetivo de la Decodificación Acústico-Fonética (DAF) es la obtención de una interpretación de la señal vocal en función de un cierto conjunto de unidades lingüísticas. El fundamento que comparten todas las técnicas utilizadas en DAF consiste en asociar uno o varios modelos a cada una de esas unidades. Estos modelos deberían ser capaces de representar la variabilidad (inter e intra-locutor) de las características acústicas de cada unidad.

Para diseñar un sistema de DAF es necesario elegir: 1) el tipo de unidad subléxica, y 2) el tipo de modelo con las técnicas que le sean propias.

Las unidades de tipo fonético son quizás las más utilizadas en los últimos años [Lee, 88] [Schwartz, 88], [Nakagawa, 89], aunque los difonemas [Colla, 87], semisílabas [Weigel, 88], sílabas [Watarabe, 86] [Wagner, 87], etc. también son elegidos por algunos investigadores. En la actualidad hay una tendencia a utilizar unidades sin significado lingüístico, basadas únicamente en propiedades acústicas (unidades segmentales acústicas) [Wilpon, 87] [Lee, 89].

Por otra parte, las técnicas y los consiguientes modelos que se utilizan en la DAF pueden clasificarse de forma general en: [Schwartz, 88]

1. Aproximaciones basadas en reglas.
2. Aproximaciones basadas en medidas de distancia.
3. Aproximaciones probabilísticas.
4. Aproximaciones basadas en funciones discriminantes.

* Subvencionado en parte por la Comisión Interministerial de Ciencia y Tecnología (TIC 448/89).

† Becario del Ministerio de Educación y Ciencia.

Las aproximaciones basadas en reglas tienen su fundamento en los Sistemas Basados en el Conocimiento [Haton, 88], y su popularidad ha decaído en los últimos años. El segundo tipo de aproximaciones está basado en técnicas de comparación entre plantillas (secuencias de vectores de características o primitivas), y es una de las menos exploradas en DAF. El tercer tipo parte de los Modelos de Markov Ocultos y constituye una de las más importantes técnicas que se emplean en la actualidad. Los modelos que sustentan el último tipo de aproximación son las Redes Neuronales Artificiales, que hoy en día constituyen un campo de investigación abierto.

La técnica fundamental utilizada en las aproximaciones basadas en distancia es el Alineamiento Temporal No Lineal (ATNL), que originalmente fue introducido para abordar el problema del reconocimiento de palabras aisladas [Sakoe, 78] [Casacuberta, 87]. Esta técnica permite la comparación entre dos secuencias de vectores de características, como consecuencia de la cual se obtiene un camino de alineamiento que define una relación entre pares de vectores (uno de cada secuencia) y una distancia o disimilitud entre las dos secuencias. Posteriormente se extendió esta técnica a la resolución del problema del reconocimiento de palabras conectadas, dando lugar a los algoritmos de Dos Pasos [Sakoe, 79], Constructor de Niveles [Myers, 81] y Un Paso [Ney, 84]. Básicamente, estos algoritmos tratan de comparar una secuencia de vectores de características correspondiente a una pronunciación con un conjunto de secuencias de plantillas-palabras. Un estudio conjunto de estos algoritmos puede encontrarse en [Aibar, 91].

Este artículo trata del desarrollo de procedimientos automáticos para la obtención de uno o varios modelos (prototipos) por unidad subléxica, así como su posterior utilización en DAF en el marco de las aproximaciones basadas en medidas de distancia. Las unidades escogidas son de tipo fonético, lo que no es óbice para que la metodología que se va a describir no sea fácilmente aplicable a otro tipo de unidades. Los modelos que se utilizan son las plantillas. Los algoritmos empleados en la fase de aprendizaje se basan en el ATNL de palabras aisladas, mientras que los empleados en la fase de decodificación se apoyan en el algoritmo de Un Paso. La metodología de aprendizaje ya ha sido utilizada con otros tipos de modelos [Rabiner, 86] [Colla, 90] y problemas [Shiraki, 88]. Trabajos preliminares a éste han sido presentados en [Aibar, 90] y [Castro, 90].

2. Segmentación de la señal vocal en unidades de tipo fonético

Una pronunciación queda representada en esta aproximación mediante una secuencia de longitud variable de vectores de características (Coeficientes Cepstrales, Espectros, Coeficientes LPC, etc.). Cada vector de características es el resultado de aplicar algún tipo de transformación sobre la señal vocal a través de una ventana de análisis de tamaño fijo que se desliza sobre la pronunciación en intervalos de tiempo.

Dado un conjunto de muestras de entrenamiento y sus correspondientes transcripciones fonéticas, la técnica empleada permite segmentar cada muestra de entrenamiento en unidades de tipo fonético mediante un algoritmo convencional de ATNL, obteniendo, una vez segmentadas todas las muestras, un conjunto de plantillas por cada unidad fonética, donde cada plantilla es una secuencia de vectores de características.

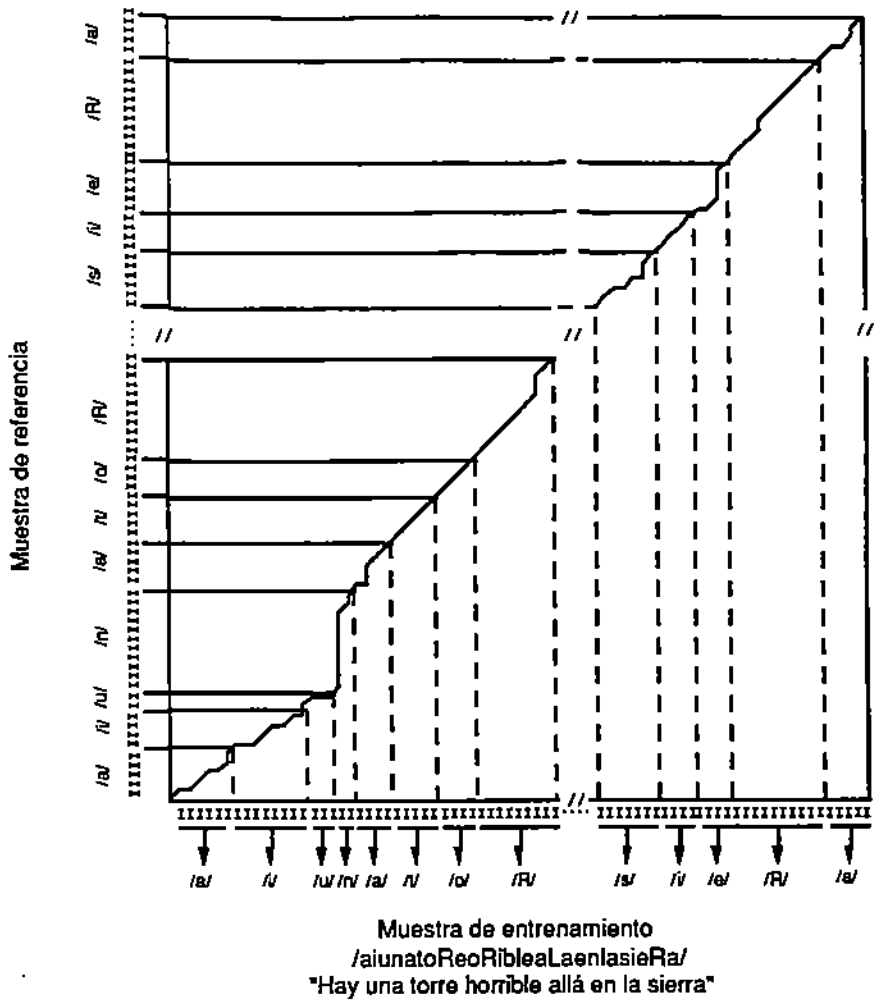

Figura 1. Un ejemplo de segmentación de la muestra de entrenamiento /aiunatoReoRibleaLaenlasieRa/ empleando la técnica de segmentación descrita. Cada símbolo I representa a un vector de características distinto.

El algoritmo desarrollado compara, mediante ATNL, cada muestra de entrenamiento con una muestra de referencia construida mediante la concatenación de los prototipos de cada unidad correspondientes a la transcripción fonética de la muestra de entrenamiento, dando como resultado un camino de alineamiento que relaciona cada elemento acústico de la muestra de entrenamiento con la muestra de referencia obtenida por la concatenación. A partir de este camino, y conocidos los límites de los prototipos concatenados, se puede realizar la segmentación de la muestra de entrenamiento y el etiquetado de los segmentos obtenidos (Figura 1).

En este algoritmo se han introducido restricciones de tipo duracional, forzando en la segmentación a que las plantillas pertenecientes a una unidad fonética tengan una duración comprendida entre un valor mínimo y otro máximo fijados a partir de conocimientos previos sobre las duraciones de las unidades en el corpus empleado.

3. Técnicas iterativas de optimización

En este apartado se describen dos técnicas que permitan obtener, a partir de un conjunto de plantillas de una unidad fonética, un prototipo o conjunto de prototipos de dicha unidad. Estas técnicas se emplean en el siguiente proceso iterativo: en cada iteración, todas las muestras de entrenamiento son segmentadas y etiquetadas mediante la técnica de ATNL descrita anteriormente, obteniéndose, a partir de los conjuntos de plantillas de las unidades fonéticas, el prototipo o prototipos de cada una de ellas. La inicialización está basada en un conjunto de

prototipos obtenidos a partir de un subconjunto reducido de las muestras de entrenamiento que ha sido segmentado manualmente.

En una primera aproximación se obtiene un único prototipo por cada unidad fonética (centroide), calculado a partir del conjunto de plantillas de dicha unidad. Posteriormente, esta técnica se ha refinado para obtener un número variable de prototipos por unidad fonética empleando técnicas de agrupamiento no supervisado ("clustering"), mediante las que se pretende capturar la variabilidad contextual de cada unidad fonética.

3.1. Centroide

Una vez todas las muestras de entrenamiento han sido segmentadas y etiquetadas (a partir de los prototipos obtenidos mediante la inicialización), se obtiene el centroide de cada conjunto de plantillas y se considera éste como el nuevo prototipo de la unidad fonética. Este procedimiento se itera hasta que los prototipos no varían de una iteración a la siguiente. Los prototipos finalmente obtenidos se pueden emplear directamente para la tarea de Decodificación Acústico-Fonética, o bien pueden servir como inicialización para otras técnicas. En esta aproximación, el coste computacional del algoritmo es debido principalmente a la computación del centroide.

3.2. Agrupamiento no supervisado ("Clustering")

Mediante esta técnica, cada unidad fonética está representada por un conjunto de prototipos obtenidos mediante clustering a partir del conjunto de plantillas de dicha unidad. El número total de prototipos es fijado inicialmente, mientras que el número de prototipos que le corresponden a cada unidad fonética es calculado, en cada iteración, a partir de la distorsión local del conjunto de plantillas de la unidad y del peso que esta distorsión tiene con respecto a las distorsiones del resto de unidades. En función del número total de clases, el coste computacional es debido, bien al número de veces que se efectúa el procedimiento de ATNL, bien a la computación de las distancias del algoritmo de clustering.

4. Resultados experimentales

Para probar la viabilidad de las aproximaciones descritas se ha empleado el corpus de *"Frases fonéticamente balanceadas"* [Vidal, 91], del que se han utilizado 50 frases, con una media de 26.5 fonemas por frase. De estas frases, 30 han sido pronunciadas por 4 locutores (2 masculinos y 2 femeninos) y las otras 20 por 8 locutores (4 masculinos y 4 femeninos), con un total de 280 frases. El número de unidades fonéticas consideradas es de 28.

La adquisición y parametrización del corpus se ha realizado en secuencias de Coeficientes Cepstrales compuestos por vectores de características de 10 dimensiones más la energía obtenidos a una velocidad de 100 vectores por segundo. Los elementos acústicos utilizados son los vectores de parámetros. El decodificador acústico-fonético empleado en la fase de reconocimiento esta basado en el algoritmo de Un Paso [Ney, 84].

Para evaluar los resultados de los experimentos se han utilizado los siguientes parámetros:

Porcentaje total = 100 c/c+s+i+b Tasa de sustituciones = 100 s/c+s+i+b

Tasa de inserciones = 100 i/c+s+i+b Tasa de borrados = 100 b/c+s+i+b

donde c es el número de unidades fonéticas correctamente reconocidas, e i, s y b son, respectivamente, el número de inserciones, sustituciones y borrados. Estos parámetros fueron calculados mediante ATNL entre la salida del decodificador y la transcripción fonética correcta de cada frase.

De las 50 frases se han empleado 30 de 4 locutores en el proceso de aprendizaje (un total de 3360 fonemas), utilizándose las restantes en el proceso de reconocimiento (3880 fonemas). Se han diseñado dos tipos de experimentos: multilocutor (ML), en el que las frases de reconocimiento empleadas han sido pronunciadas por los mismos locutores que las de

aprendizaje; e independiente del locutor (IL), en el que las 20 frases de reconocimiento han sido pronunciadas por cuatro locutores distintos de los empleados para aprender. Los resultados de estos experimentos se ofrecen en la tabla siguiente:

Tabla 1: Resultados de los experimentos realizados.

	CENTROIDE				CLUSTERING							
					Sin iteración				Con iteración			
	p	t	t		p	t	t		p	t	t	
	t	i	s	b	t	i	s	b	t	i	s	b
	M	4	5	3	5	8	2		4	9	3	
S	2		6	7	2		9	2	6		3	2
	S	4	4	3	4	7	3		4	8	3	
I	3		7	6	7		2	4	3		6	3

5. Conclusiones y futuros desarrollos

En este trabajo se han presentado varias técnicas para el aprendizaje de unidades fonéticas representadas mediante plantillas. De los resultados se desprende la necesidad de emplear varios prototipos para representar la variabilidad de las unidades fonéticas. Asimismo, puede resultar sorprendente que en la iteración de la técnica de clustering los resultados no mejoren. Sin embargo, esto se puede explicar si tenemos en cuenta, por un lado, que el método de segmentación (óptimo cuando se emplea un único prototipo por unidad fonética) es subóptimo cuando se emplea más de un prototipo, y por otra parte, que en la obtención de los prototipos de cada unidad fonética no se hace ningún tipo de filtrado que permita evaluar la bondad de los seleccionados.

El estudio de los dos temas anteriormente citados, junto con la implementación del algoritmo de Dos Pasos [Sakoe, 79] como método de reconocimiento (lo que puede permitir trabajar con distancias normalizadas [Aibar, 91]), así como la aplicación de estas técnicas al corpus completo de las Frases, son los futuros trabajos a desarrollar.

6. Bibliografía

[Aibar, 90] P. Aibar, M.J. Castro, F. Casacuberta, E. Vidal: "Multiple Template Modeling of Sublexic Units", en "Speech Recognition and Understanding: Recent Advances, Trends and Applications". Ed. P. Laface. Springer-Verlag. NATO Advanced Studies Institute Series, 1990.

[Aibar, 91] P. Aibar, et al: "Alineamiento Temporal Óptimo de Concatenaciones de Modelos y Secuencias Acústicas: una visión unificada y nuevas aportaciones", Informe de investigación, DSIC, Universidad Politécnica de Valencia, 1991.

[Casacuberta, 87] F. Casacuberta, E. Vidal: "Reconocimiento Automático del Habla", Marcombo, 1987.

[Castro, 90] M.J. Castro, P. Aibar, F. Casacuberta, E. Vidal: "Automatic Selection of Sublexic Templates by using Dynamic Time Warping Techniques". Signal Processing V: Theories and Applications. L. Torres, E. Masgrau y M. A. Lagunas (eds.). Elsevier Science Publishers B.V., pp 1351-1354, 1990.

[Colla, 87] A.M. Colla, A.E. Rosenberg: "Unsupervised bootstrapping of diphone-like templates for Connected Speech Recognition", Proc. ICASSP 87, pp 1281-1284, 1987.

[Colla, 90] A.M. Colla: "On training a large vocabulary speech recognition system", Proc. International Conference on Speech Technologies, pp 107-116, Verba 90.

[Haton, 88] J.P. Haton: "Knowledge-Based Approaches in Acoustic-Phonetic Decoding of Speech" in "Recent Advances in Speech Understanding and Dialog Systems", H. Nieman, M. Lang, G. Saguer (eds.), Springer-Verlag, pp. 50-70, 1988.

[Lee, 88] K.F. Lee: "Large-Vocabulary Speaker Independent Continuos Speech Recognition: the SPHINX System", PH. Thesis, Tec. Rep. CMU-CS 88-148. Carnegie Mellon Univ, 1988.

[Lee, 89] C.H. Lee, B.H. Juang, F.K. Soong, L.R. Rabiner: "Word recognition using whole word and subword models". Proc. ICASSP89, pp 683-686, 1989.

[Myers, 81] C.S. Myers, L.R. Rabiner: "A level building Dynamic Time Warping Algorithm for Connected Word Recognition", IEEE Trans. ASSP, vol. 29, pp. 284-297, 1981.

[Nakagawa, 89] S. Nakagawa: "Speaker-independent continuos-speech recognition by phoneme-based word spotting and time-syncronuous context-free parsing", Computer Speech and Language, vol.3, pp 277-299, 1989.

[Ney, 84] H. Ney: "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition", IEEE Trans. ASSP, vol. 32, no. 2, pp 263-271, Apr. 1984.

[Rabiner, 86] L.R. Rabiner, J.G. Wilpon, B.H. Juang: "A Continuous Training Procedure for Connected Digit Recognition", Proc. ICASSP 86, pp 1065-1068, 1986.

[Sakoe, 78] H. Sakoe, S.Chiba: "Dynamic Programming Algorithm Optimization for Spoken Words Recognition". IEEE Trans. ASSP, vol. 26, pp. 43-49, Feb. 1978.

[Sakoe, 79] H. Sakoe: "Two-Level DP matching - A dynamic programming based pattern matching algorithm for connected word recognition", IEEE Trans. ASSP, vol. 27, pp. 588-595, 1979.

[Schwartz, 88] R.M. Schwartz et. al.: "Acoustic-Phonetic Decoding of Speech" in "Recent Advances in Speech Understanding and Dialog Systems", H. Nieman, M. Lang, G. Saguer (eds.), Springer-Verlag, pp. 25-50, 1988.

[Shiraki, 88] Y. Shiraki, M. Honda: "LPC Speech Coding Based on Variable-length Segment Quantization", IEEE Trans. ASSP, vol. 36, no. 9, pp 1437-1444, Sep. 1988.

[Vidal, 91] E. Vidal, et al.: "Construcción de Sistemas de Reconocimiento del Habla mediante Técnicas de Aprendizaje Automático: Objetivos y Estado Actual", Boletín SEPLN, 1991. Pendiente de publicación.

[Wagner, 87] M. Wagner: "A speech recognition experiment with the entire syllable inventory of standar Chinese", Speech Communication, no. 6, pp 363-369, 1987.

[Watarabe, 85] T. Watarabe: "Syllable recognition for continuous Japanese speech recognition", Proc. ICASSP 86, pp 2295-2298, 1986.

[Weigel, 88] W. Weigel: "Recognition of demisyllables based on Dynamic Programming methods", Speech Communication, vol. 7, pp 297-304, 1988.

[Wilpon, 87] J.G. Wilpon, B.H. Juang, L.R. Rabiner: "An Investigation on the use of acoustic sub-word units for automatic speech recognition". ICASSP87, pp 821-824, 1987.