SEGMENTADOR DE FONEMAS EN CATALÁN BASADO EN *DHMM*

Francesc Alías Pujol

Dept. de Comunicaciones y Teoría de la Señal
Enginyeria La Salle. Universitat Ramon Llull
falias@salleURL.edu

Ignasi Iriondo Sanz

Dept. de Comunicaciones y Teoría de la Señal
Enginyeria La Salle. Universitat Ramon Llull
iriondo@salleURL.edu

ABSTRACT

The unit selection based text-to-speech (TTS) systems [1] *work with large speech corpora labelled with a huge amount of data. Recorded speech is time-aligned at phonetic level by segmentation marks (phoneme boundaries). Although manual phonetic alignment is considered more accurate than automatic methods, it is too time consuming to be commonly used for aligning large corpora. The topic of this paper is the development of an automatic system for phoneme segmentation based on Discrete Hidden Markov Models and the evaluation of the system accuracy on two Catalan continuous-speech databases.*

1. INTRODUCCIÓN

El proceso de generación de voz en sistemas de síntesis concatenativa parte de fragmentos o unidades de voz humana real procedentes de un corpus de voz. Este corpus, además de contener las muestras de voz, debe segmentarse y etiquetarse con precisión para obtener la información acústica que permita elegir y procesar las unidades adecuadas para la síntesis del habla.

Una de estas etiquetas son las marcas de segmentación, que se encargan de delimitar la posición de las unidades dentro de la señal de voz. Aunque estas marcas pueden obtenerse manualmente con gran exactitud (etiquetadas por un experto), el coste temporal asociado [2] es excesivo en comparación al necesario para el proceso automático equivalente. Además, si distintos expertos etiquetan la misma frase, las marcas colocadas varían dentro de un cierto margen de desviación [3]. Por estos motivos, en el contexto de los grandes corpus de voz que se utilizan en los sistemas de conversión texto voz basados en selección de unidades [1], es imprescindible disponer de un segmentador automático.

Existen distintas técnicas que permiten la alineación entre la señal de voz y su transcripción fonética (*forced alignment*). Por citar las más utilizadas hasta la actualidad [4], se han utilizado redes neuronales (*NN*), *dynamic time warping* (*DTW*), distintas variantes de los modelos ocultos de Markov (*HMM*) o combinaciones de ellas (*HMM* & *NN*). El objetivo de la implementación del sistema que a continuación se describe es doble: evaluar el funcionamiento de una de estas técnicas sobre un corpus de voz de habla continua en catalán y comparar los resultados obtenidos con los presentados en la literatura para otras técnicas y otras lenguas.

2. DESCRIPCIÓN DEL SISTEMA

De las distintas técnicas que se acaban de comentar, el sistema que se ha desarrollado para obtener las marcas de segmentación está basado en los modelos ocultos de Markov discretos (*DHMM*). Este proceso se divide en dos fases: una primera, en la que se entrenan los modelos de cada uno de los fonemas del catalán y una segunda, donde se realiza la segmentación propiamente dicha, es decir, la fase de explotación del sistema.

2.1. Fase de entrenamiento del sistema

En primer lugar se debe caracterizar acústicamente la señal de voz mediante un ventaneo temporal y la correspondiente parametrización. En este caso se han calculado los parámetros LPC-Cepstrum [5] de las tramas de 20 ms (con 10 ms de solapamiento) de la señal analizada.

Debido a la utilización de *DHMM*, será necesario la generación de un libro de códigos (*codebook*) que permita etiquetar cada una de les tramas parametrizadas de la señal con uno de los *M clusters* de este codificador. Este proceso parte de un conjunto de *F* frases fonéticamente balanceadas en el cual todos los fonemas del catalán estén bien representados. Mediante el algoritmo iterativo de *binary splitting* [5] se genera el *codebook* a partir de la voz parametrizada. Para el sistema se ha escogido $M = 64$, ya que es el primer múltiplo de 2 (en cada iteración se duplica el número de *centroides*) mayor al número de fonemas a codificar.

Este sistema requiere la segmentación manual de estas *F* frases y su posterior cuantificación vectorial (*VQ*) mediante el *codebook*; así el sistema puede conocer la correspondencia entre cada fonema y el conjunto de símbolos que le corresponden.

Con esta información se pasa a entrenar los 37 *DHMMs* (36 fonemas más el silencio) de topología *Left-to-right* y de 5 estados por fonema (fija la duración mínima permitida para un fonema):

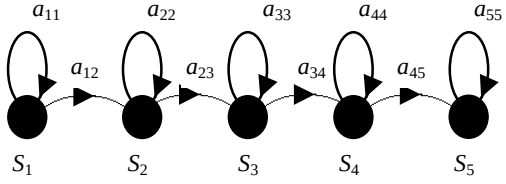

Figura 1. DHMM con topología left-to-right de 5 estados.

A partir de cada una de las secuencias que describen cada una de les *n* representaciones del fonema *i* dentro del conjunto de entrenamiento, se define su modelo (λ=[A,B,π]) correspondiente. Partiendo de una inicialización uniforme de las matrices de probabilidad de transición entre estados (*A*) y de probabilidad de observación de símbolo (*B*), el modelo se refina iterativamente mediante el algoritmo de *Baum and Welch* [5] hasta obtener el modelo definitivo que caracteriza a cada fonema.

2.2. Fase de explotación del sistema

Una vez modelados todos los fonemas, se procede a la segmentación de las frases que componen el corpus de voz. Por un lado, cada frase debe ser parametrizada y cuantificada vectorialmente mediante el *codebook*, obteniéndose la secuencia de símbolos que le corresponde. Por otro lado, se construye un macromodelo de Markov concatenando los modelos de los fonemas que contiene la frase, según su transcripción fonética.

A continuación, se procede al óptimo alineamiento entre el macromodelo (λ') y la secuencia de símbolos de la frase (*O*) mediante el algoritmo de Viterbi, que es el encargado de determinar la secuencia de estados óptima (*Q*) (maximiza la $P[Q|O, \lambda']$). Para evitar posibles errores de precisión se escoge la variante LogViterbi [5] que aplica logaritmos en los cálculos.

A partir de los índices de la secuencia óptima de estados (*Q*) se obtienen las tramas asignadas a cada fonema sabiendo que 5 estados corresponden a un fonema. Las marcas se colocarán en la mitad de la primera ventana que corresponda al estado de cambio de fonema, exceptuando la inicial y final que se colocan en la primera y última muestras de la señal, respectivamente. La posición de la marca, en muestras absolutas, se obtiene teniendo en cuenta el tamaño de la ventana de análisis y su solapamiento. Este proceso se repite para todas las frases del corpus; los resultados se almacenan en un fichero asociado a cada frase.

3. TESTS Y EVALUACIÓN DE RESULTADOS

Para poder evaluar el funcionamiento del sistema descrito, se han utilizado dos corpus de habla continua en catalán. El primero (C1) de 5.500 fonemas, proviene de la voz en *off* de un reportaje televisivo. Su mayor inconveniente es que, como no fue grabado para síntesis, presenta una deficiente vocalización El segundo (C2), de tamaño inferior (1.200 fonemas), presenta una mejor vocalización. Ambos corpus contienen voz del mismo locutor profesional y han sido etiquetados manualmente como marco para la verificación del sistema automático.

Se han realizado dos tests, uno sobre cada corpus. A partir de *N* frases escogidas adecuadamente (20 para C1 y 10 para C2), se entrena al sistema y se segmenta todo el corpus asociado. Los resultados (tabla 1) se presentan en tanto por ciento de acierto respecto a una determinada desviación entre la marca manual y la marca automática (referencia). Estadísticamente se ha podido comprobar que si un mismo experto repite el proceso de delimitar los fonemas de una misma frase, las marcas obtenidas pueden variar dentro de un margen de ±20 ms. Este criterio [4] es el escogido para evaluar los resultados del sistema automático.

Desviación [ms]	≤ 15	≤ 20	≤ 25	≤ 30
% Acierto C1	68	78	83	88
% Acierto C2	72	85	87	92

Tabla 1. Resultados de la segmentación de los dos corpus.

Si se observan los resultados para la desviación de 20 ms en la tabla 1, se puede comprobar que: la calidad en la vocalización de la señal grabada influye en el funcionamiento del sistema (C2 mejora en un 7% los resultados de C1), y que los resultados obtenidos se mueven dentro del margen referido en la literatura [4] (del 77% al 90% de acierto, para 20ms, en varios sistemas). También se ha segmentado C2 con los modelos de C1, con un 5% menos de acierto, resultado lógico por las características de C1.

4. CONCLUSIONES Y POSIBLES MEJORAS

Se ha desarrollado y evaluado un sistema de segmentación automática sobre habla continua en catalán que aporta unos buenos resultados de etiquetado. Sus puntos críticos son: la transcripción fonética, la elección adecuada del conjunto de entrenamiento y la calidad (vocalización) del corpus a segmentar.

El sistema se podría mejorar trabajando con una estructura más flexible en los modelos de fonemas (no fijar 5 estados/fonema), usando *HMM* continuos (se evita el proceso de discretización), sin utilizar una segmentación manual previa (segmentación jerárquica [6]) o añadiendo información acústico-fonética [4] en el proceso de alineación. Se prevé mejorar el funcionamiento del sistema con estas ideas, estudiando el coste computacional añadido respecto a la mejora en la tasa de acierto.

AGRADECIMIENTOS

Este trabajo se ha realizado con el apoyo del Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya mediante la beca 2000FI-00679 del *DOGC 07/02/01*

5. REFERENCIAS

[1] Black, A. and Taylor, P., "Automatically clustering similar units for unit selection in speech synthesis", European Conference on Speech Communication and Technology, pp.601-604, Rodes, Greece, 1997.

[2] Kvale, K., "On the connection between manual segmentation conventions and "errors" made by automatic segmentation" Proc. of ICLSP'94, Yokohama, Japan, vol. 3, pp. 1667-1670, 1994.

[3] Wesenick, M.-B. and Kipp, A., "Estimating the quality of phonetic transcriptions and segmentations of speech signals" Proc. of ICLSP'96, Philadelphia, USA, pp.129-132, 1996.

[4] Hosom, J.-P., "Automatic time Alignment of phonemes using acoustic-phonetic information", Ph. D. Thesis, 2000.

[5] Rabiner, L.R. and Juang, B.-H. "Fundamentals of speech recognition", Prentice Hall, 1993.

[6] Paws, S., Kamp, Y., and Willems, L. "A hierarchical method of automatic speech segmentation for synthesis applications". Speech Communication, 19, pp. 207-220, 1996.