Análisis de la Segmentación Automática de Fonemas para la Síntesis de Voz.

Jordi Adell, Antonio Bonafonte

Dpto. de Teoría de la Senyal y Comunicación
Centro de Investigación TALP
Universidad Politécnica de Catalunya (UPC)
www.talp.upc.es

Jon Ander Gómez, María José Castro

Dpto. de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
www.dsic.upv.es

Resumen

En este artículo se presentan dos nuevos sistemas para las segmentación de voz en fonemas. Uno basado en un *clustering* acústico previo a un alineado por programación dinámica y el segundo basado en una corrección específica de las fronteras mediante un árbol de regresión. Se discute el uso de medidas objetivas o perceptuales para la evaluación de estos sistemas. Los sistemas presentados claramente mejoran los resultados del sistema de partida basado en HMM y obtienen resultados similares a la concordancia entre diferentes segmentaciones manuales. Se muestra como las características fonéticas pueden ser utilizadas satisfactoriamente, junto con los HMM, para la detección de las fronteras. Finalmente, se enfatiza la necesidad de utilizar tests perceptuales para evaluar la segmentación de las bases de datos para síntesis de voz.

1. Introducción

Actualmente, los sistemas de síntesis de voz por concatenación son los más ampliamente utilizados y lideran la calidad de los sistemas de conversión texto-voz. Aún así, esta aproximación necesita grandes bases de datos para contener la máxima variabilidad posible. En muchos casos, el éxito de los sistemas recae en el tratamiento correcto de la base de datos.

Al trabajar con sistemas de síntesis por concatenación, debemos invertir una parte importante del esfuerzo en preparar la base de datos. Muchas veces se requieren nuevas bases de datos, ya sea para crear nuevos locutores del mismo idioma, para tareas multilingüísticas, para crear diferentes estilos de habla o incluso para adaptar nuestro sistema a nuevos dominios. Por lo tanto, es importante reducir el contexto necesario para crear nuevas bases de datos.

Algunas partes de este proceso de crear nuevas bases de datos deben ser, al menos por el momento, realizadas de forma manual. Este tipo de tareas manuales tienen un alto coste, requieren mucho tiempo y además pueden introducir inconsistencias. La segmentación en fonemas de la señal de voz es una de estas tareas y su posible resolución de forma automática reduciría, en mucho, el esfuerzo necesario para crear estas nuevas bases de datos.

Antes de afrontar el problema debemos acotar el tipo de problema al que nos enfrentamos. La segmentación automática se puede atacar desde tres puntos de vista, en función de las restricciones que tengamos: *ninguna*, *acústicas* o *lingüísticas* [1]. El trabajo que aquí se presenta se enmarca dentro de unas restricciones lingüísticas. Hemos considerado que se conoce la transcripción fonética de la base de datos; la cual se puede obtener, por ejemplo, mediante un supervisado manual de un etiquetado automático.

A pesar de que algunos investigadores afirman que los sistema actuales de segmentación obtienen resultados parecidos a los conseguidos mediante segmentaciones manuales [2], muchos otros siguen llegando a sus mejores resultados utilizando supervisión manual de la base de datos. Por eso, en la literatura reciente podemos encontrar estudios sobre diferentes métodos de segmentado automático de la voz, como por ejemplo:

- Modelos Ocultos de Markov [3, 2]
- Redes Neuronales [4]
- Alineado temporal mediante programación dinámica [5, 6]

En este artículo se escogen algunas de estas técnicas y se aplican a una misma base de datos. Se ha considerado que hay una falta de marcos reconocidos que permitan la comparación entre diferentes sistemas de segmentación. Por eso se ha comparado esta serie de sistemas aplicándolos a la misma base de datos, bajo las mismas condiciones. Además, se proponen dos nuevos métodos para la segmentación de la voz en fonemas; se describen los métodos, se discute sobre su evaluación y se presentan los resultados analizando las conclusiones.

2. Descripción de los métodos

A continuación se describirán los métodos utilizados en este trabajo. Se presenta su entorno teórico y características distintivas de cada uno de ellos.

2.1. Modelos Ocultos de Markov

Estos modelos han sido utilizados frecuentemente y desde hace tiempo para la segmentación de voz [3]. Se trata de realizar un proceso de reconocimiento sobre la base de datos de voz que se ha grabado. Sin embargo aquí, como conocemos la transcripción fonética de la base de datos, solo se realizará un alineado forzado de la señal con los modelos correspondientes. Ello se llevará a cabo mediante el algoritmo de Viterbi.

Los modelos acústicos usados son semifonemas dependientes del contexto [7]. Para realizar el alineado forzado se ha utilizado nuestro sistema de reconocimiento: RAMSES, que es el sistema de reconocimiento del habla de la UPC. Para ello se han entrenado los modelos independientes del contexto con 12 iteraciones, más 6 iteraciones dependientes del contexto.

Este trabajo has sido parcialmente subvencionado por la Unión Europea mediante la beca FP6-506738 (proyecto TC-STAR, http://www.tc-star.org) y por el gobierno español mediante la subvención TIC2002-04447-C02 (proyecto ALIADO, http://gps-tsc.upc.es/veu/aliado).

2.2. Alineado acústico mediante Programación Dinámica

Este método se basa en el alineado de una señal que ya está segmentada con otra que no lo esté. Lo que se propone es alinear la señal que queremos sintetizar con una señal sintetizada con los mismos fonemas. Como la base de datos utilizada para la síntesis está etiquetada, conocemos las fronteras de los fonemas sintetizados. Estas fronteras se mapean a la vez en la señal grabada obteniendo así la segmentación de la señal deseada.

Se ha usado una parte del corpus que está manualmente segmentado para construir un locutor para nuestro sistema de síntesis,el sistema UPC-MLTTS, el sistema de síntesis del TALP. Luego, con esta voz, se ha sintetizado el resto del corpus. Después de esto, se ha alineado la voz grabada con la voz generada, utilizando la herramienta DTW-Festvox [8].

2.3. *Clustering* acústico + Alineado Temporal No lineal

A continuación presentamos la técnica de segmentación automática *Clustering* acústico + Alineado Temporal No lineal (*Acustic Clustering-Dynamic Time Warping*, AC-DTW), basada en una fase inicial de *clustering* realizado sobre el espacio definido por los vectores acústicos para obtener clases acústicas, y la posterior asociación, en una segunda fase, de estas clases con unidades fonéticas mediante probabilidades condicionales.

Las fronteras entre unidades fonéticas son establecidas mediante un algoritmo de programación dinámica que utiliza las probabilidades *a posteriori* de que cada unidad fonética haya sido pronunciada dado un vector acústico. Estas probabilidades *a posteriori* son calculadas combinando las probabilidades de nivel acústico, que se estiman a partir de la mezcla de Gaussianas fruto del *clustering* realizado sobre el espacio definido por los vectores acústicos, y las probabilidades condicionales de cada clase acústica con respecto de cada unidad fonética. [9].

Para realizar el *clustering* asumimos que cada clase acústica puede modelarse mediante una distribución normal o de Gauss. Los parámetros que definen cada distribución normal: media y matriz de varianzas-covarianzas, son estimados mediante la versión no supervisada del proceso de estimación por máxima verosimilitud [10]. Por tanto, es posible estimar la probabilidad de cada clase acústica a_c dado un vector acústico x_t, $\Pr(a_c|x_t)$, a partir de la mixtura de Gaussinas. Pero necesitamos la probabilidad de cada unidad fonética u_f dado un vector acústico x_t, $\Pr(u_f|x_t)$. Para ello se utiliza un conjunto de probabilidades condicionales que permite calcular las probabilidades de nivel fonético a partir de las de nivel acústico.

El uso de las probabilidades condicionales nos permite calcular las densidades de probabilidad condicional de observar un vector acústico x_t cuando ha sido pronunciado un fonema u_f como sigue [9]:

$$p(x_t|u_f) = \sum_{c=1}^{C} p(x_t|a_c) \cdot \Pr(a_c|u_f) \qquad (1)$$

donde C es el número de clases acústicas, $p(x_t|a_c)$ es la densidad de probabilidad condicional a nivel acústico estimada según la fórmula de Gauss para las distribuciones normales, y $\Pr(a_c|u_f)$ es la probabilidad condicional de que se manifieste la clase acústica a_c cuando ha sido pronunciado el fonema u_f. Entonces, aplicando la regla de Bayes obtenemos las probabilidades de nivel fonético según la siguiente ecuación:

$$\Pr(u_f|x_t) = \frac{\sum_{c=1}^{C} p(x_t|a_c) \cdot \Pr(a_c|u_f)}{\sum_{j=1}^{F} \Big(\sum_{c=1}^{C} p(x_t|a_c) \cdot \Pr(a_c|u_j) \Big)} \qquad (2)$$

donde F es el número de unidades fonéticas.

El conjunto de probabilidades condicionales $\Pr(a_c|u_f)$ fue calculado a partir de frases segmentadas y etiquetadas manualmente.

En realidad, estos modelos acústicos tienen grandes similitudes con los HMMs semicontinuos de un estado, donde las probabilidades de emisión se identificarían con las probabilidades a posteriori estimadas según (2) y las probabilidades de transición, es decir, el modelado de la duración, estaría implícito por el hecho de normalizar las densidades de probabilidad a nivel fonético. Esta normalización enfatiza las transiciones de un fonema a otro, lo que facilita el modelado en etapas posteriores.

2.4. Corrección Específica de Fronteras mediante un Árbol de Regresión

Aquí se presenta un nuevo sistema basado en la Corrección Específica de Fronteras (BSC - Boundary Specific Correction) [11]. La BSC consiste en aplicar una corrección dependiente del tipo de transición, es decir, del tipo de fonemas que toman parte en ella.

El método propuesto consta de 2 pasos. En el primero se realiza una segmentación inicial que se refina en el segundo paso. La técnica de realizar la segmentación en dos pasos ha sido ya utilizada con anterioridad como vemos en [12].

Habitualmente en la literatura encontramos sistemas basados en medidas acústicas locales [13]. Estos sistemas se basan en la detección de fronteras mediante medidas acústicas locales. Sin embargo, experimentos previos muestran como las características fonéticas nos permiten una mejor corrección de las fronteras de los HMM que las medidas acústicas [14]. Por lo tanto, en el sistema que presentamos se introduce, en el segundo paso, el uso de información fonética. Las fronteras se refinan en base a las características fonéticas (modo y punto de articulación, sonoridad, etc.) de los dos fonemas involucrados en la transición.

Primero se realiza una segmentación inicial con base acústica mediante HMM (se han utilizado los resultados de aplicar 2.1). El árbol de regresión nos permite hacer una regresión del error de las fronteras. Una parte pequeña del corpus segmentado manualmente se utiliza para entrenar el árbol. Por lo tanto, este árbol nos permitirá estimar el error cometido por los HMM en el resto del corpus y poder corregirlo.

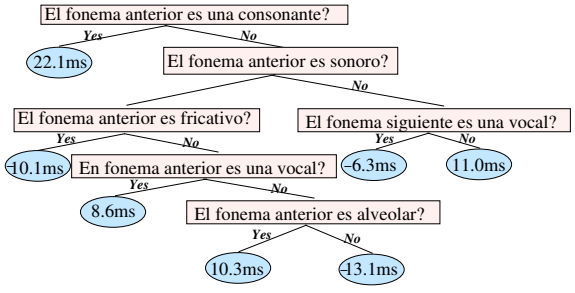

Figura 1: Ejemplo de Árbol de Regresión

La regresión se lleva a cabo en base a información fonética. Para ello, los criterios de decisión en los nodos del árbol son preguntas binarias sobre las características fonéticas de los fonemas anterior y posterior a la frontera. En la Figura 1 se presenta un ejemplo.

3. Evaluación de la segmentación

El sistema de evaluación que más se ha utilizado en la literatura es el que trata de evaluar el grado de parecido entre la segmentación automática y una segmentación manual, llevada a cabo por una o más personas. Habitualmente se mide como el porcentaje de fronteras cuyo error es inferior a una cierta tolerancia y se calcula para un cierto rango de tolerancias. En [4] también se propone calcular la media de estos porcentajes, obteniendo así un solo valor que permita comparar más fácilmente diferentes sistemas, a este valor lo llamaremos *MeanTol*.

En el marco de este tipo de evaluaciones objetivas, algunos investigadores se han preguntado si una segmentación manual es o no una referencia válida [4, 15]. Para evaluarlo han hecho que diferentes personas segmentaran la misma base de datos. Luego han medido las diferencias entre ellas. Como resultado de este experimento en [4] el 97 % de las fronteras mostraron una diferencia entre ellas inferior a 20ms y en [15] el 93 %. Interpretamos estos valores como un límite para sistemas que sean evaluados mediante este método, debido a que en el mejor de los casos cuando podamos obtener una segmentación perfecta con tolerancias del 100 %, si cambiamos la segmentación de referencia entonces pasaremos a tener tolerancias de aproximadamente el 95 % para la misma base de datos.

Por otro lado, en [2] proponen una evaluación perceptual para poder comparar diferentes segmentaciones. Una evaluación perceptual del sistema de síntesis puede medir el objetivo final de la aplicación que estamos desarrollando, además de que nos ayudaría a darnos cuenta de cual es el impacto real de los posibles errores en la segmentación, en la percepción de la voz sintetizada. Por estas razones, lo que aquí se propone es hacer una primera evaluación objetiva de los sistemas para eliminar aquellos que obtengan resultados bajos. Por otro lado, cuando los resultados son comparables a los manuales, se puede hacer una evaluación perceptual de los mejores sistemas para evaluar la mejora real que pueden aportar.

4. Resultados Experimentales

4.1. Corpus

Para realizar los experimentos se ha utilizado un corpus grabado en el centro de investigación TALP (Tècniques i Aplicacions del Llenguatge i la Parla). Consiste de 516 oraciones segmentadas manualmente, lo que supone alrededor de media hora de voz. El locutor es una mujer profesional y el estilo es neutro. 40 de estas oraciones se han seleccionado al azar para ser el corpus de entrenamiento para los sistemas AC-DTW y RT-BSC y utilizadas para crear el locutor para el DTW. Por lo tanto los resultados presentados están evaluados sobre el resto del corpus.

Se ha utilizado una parametrización basada en Cepstrums (MFPC - Mel Frequency Power Cepstrums), con su primera y segunda derivadas y la primera derivada de la energía. La ventana para el cálculo de los parámetros ha sido de 20ms con una cadencia de 4ms.

4.2. Evaluación Objetiva

Se han calculado los porcentajes para las siguientes tolerancias: 5, 10, 15, 20 y 25ms. Los resultados se pueden ver en la Tabla 1.

Sistema	< 5	< 10	< 15	< 20	< 25	*MT*
HMM	41 %	67 %	85 %	92 %	94 %	76
DTW	30 %	50 %	62 %	69 %	73 %	58
AC-DTW	**52 %**	**78 %**	**89 %**	**93 %**	**96 %**	**82**
RT-BSC	**58 %**	**82 %**	**91 %**	**94 %**	**96 %**	**84**

Tabla 1: Porcentajes de fronteras con un error inferior a ciertas tolerancias. También se muestra el valor *MeanTol* para cada uno de los sistemas estudiados. (*Tolerancias en ms*)

En la Tabla 1 se puede observar que los porcentajes más bajos corresponden al DTW y que AC-DTW y el RT-BSC claramente mejoran los resultados obtenidos por los HMM.

El algoritmo de programación dinámica se ha considerado más preciso en media que los HMM, a pesar que su problema es que comete errores esporádicos pero grandes [5]. En la Tabla 2 vemos que si únicamente consideramos los errores inferiores a 20ms para las medidas, los HMM siguen siendo mejores que el algoritmo de programación dinámica. Por lo tanto no podemos decir que el algoritmo de programación dinámica sea superior a los HMM en ningún sentido.

También se han realizado algunos experimentos más en relación con el sistema DTW. Se han utilizado más oraciones manualmente segmentadas para construir el locutor necesario para el sistema de síntesis. La oraciones han sido escogidas mediante un algoritmo voraz para asegurar la máxima cobertura fonética. Los resultados se muestran en la Tabla 3.

Sistema	< 5	< 10	< 15	MT
DTW	44 %	74 %	90 %	69
HMM	44 %	77 %	93 %	**71**

Tabla 2: Resultados cuando solo se tiene en cuenta errores inferiores a 20ms. (*Tolerancias en ms*)

En esta tabla podemos observar como los resultados del sistema de programación dinámica mejoran significativamente al utilizar más datos segmentados manualmente. Sin embargo, incluso utilizando 400 oraciones los resultados no consiguen llegar a los niveles de los otros métodos. Estas observaciones nos permitirán descartar este método para ser utilizado en segmentación automática de voz. Aunque no se puede descartar que con bases de datos mayores el sistema obtuviera resultados más satisfactorios.

DTW Porcentajes						
Oraciones	< 5	< 10	< 15	< 20	< 25	MT
40	30 %	50 %	62 %	69 %	73 %	57
200	37 %	61 %	72 %	80 %	85 %	67
300	39 %	59 %	72 %	80 %	84 %	67
400	**40 %**	**62 %**	**77 %**	**85 %**	**88 %**	**70**

Tabla 3: Resultados para el DTW utilizando diferentes grupos de oraciones manualmente segmentadas. (*Tolerancias en ms*).

4.3. Evaluación Perceptual

El algoritmo DTW no se ha contemplado en el test perceptual debido a los resultados obtenidos en el test objetivo. Aunque sí los otros tres sistemas ya que obtienen resultados comparables a las discrepancias entre segmentaciones manuales descritas en el apartado 3.

Para este test se han utilizado 50 oraciones que se han sintetizado utilizando cada sistema de segmentación y nuestro sistema de síntesis (UPC-MLTTS), el cual se basa en selección de unidades y TD-PSOLA se utiliza sólo en los caso en que las duraciones difieran en más de 15ms y la frecuencia fundamental en más de 20Hz. Las 476 oraciones de la base de datos de test se usaron como catálogo de unidades y la prosodia fue extraída de las oraciones grabadas para eliminar los efectos que podrían incorporar los modelos de prosodia.

Las oraciones se han presentado a un grupo de 10 participantes en bloques de 20, escogidas al azar entre las 50. Cada oración ha sido sintetizada utilizando dos sistemas diferentes escogidos al azar, y los participantes fueron preguntados sobre cual de las dos era más natural. Las respuestas permitidas fueron: *igual*, *más natural* o *mucho más natural*. Esto nos permitió evaluar cada sistema frente a cada uno de los otros.

Los resultados del test se presentan en las Figuras 2 y 3. Se puede observar como los HMMs se han preferido siempre a cualquier otro sistema y se han preferido igual a la segmentación manual. El sistema RT-BSC sólo se prefiere ante el AC-DTW.

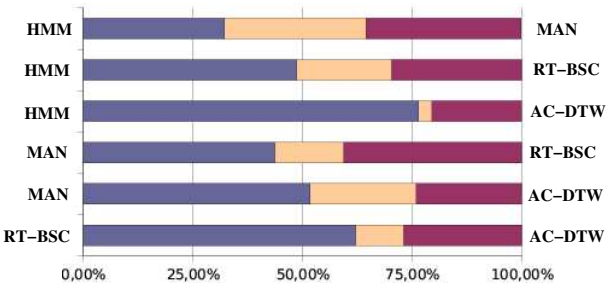

Figura 2: Porcentaje de veces que un sistema se prefirió a cada uno de los demás. Los colores oscuros muestran la veces que los sistemas indicados en los extremos fueron preferidos. EL color claro muestra las veces que fueron considerados igual de naturales.

En la Figura 3 se presenta la distribución de las respuestas. Se observa como HMM vs. MAN presenta una distribución plana, lo que indica que son sistemas comparables. El resto de gráficas muestra un claro sesgo hacia los sistemas HMM y MAN.

Al analizar los resultados perceptuales se debe tener en cuenta que para comparar cuatro sistemas como se ha hecho aquí, se debería disponer de una elevado número de participantes en el test. Para suplir esta carencia se ha evaluado la consistencia de las respuestas entre participantes, y la varianza de las respuestas para una misma pregunta, es baja; lo cual nos permite confiar en los resultados obtenidos.

A pesar de que los resultados obtenidos por los sistemas MAN y HMM son parecidos en media, los HMM realizan errores más graves que no se darían en una segmentación manual. Para ello se podría eliminar aquellas unidades de la base de datos segmentada con HMM que difirieran mucho de unas estadísticas medias de cada tipo de unidad. Ello nos permitiría

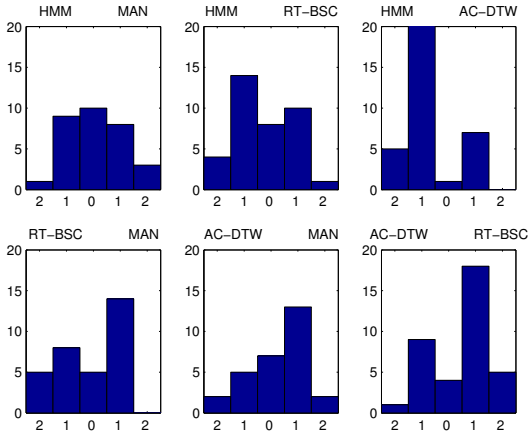

Figura 3: Distribuciones de las respuestas para cada par de sistemas. Los números del eje horizontal significan: 0-*igual*, 1-*más natural* y 2-*mucho más natural*. Sobre las figuras aparecen el nombre de los sistemas comparados.

eliminar unidades inservibles [16].

Los nuevos sistemas que hemos presentado aquí han conseguido unos muy buenos resultados objetivos. Aunque estos no se han visto reflejados en la evaluación perceptual.

5. Conclusiones

En el presente articulo se han revisado las técnicas existentes destinadas a la segmentación de voz aplicadas a la síntesis. También se ha discutido sobre los métodos posibles para la evaluación de estas técnicas y sobre la conveniencia de realizar o no evaluaciones objetivas y/o perceptuales. El hecho de aplicar una serie de sistemas a una misma base de datos y bajo las mismas condiciones nos ha permitido comparar de forma adecuada los diversos sistemas.

Se han presentado dos nuevos métodos para la segmentación que han mejorado los resultados del sistema de partida basado en HMM. Los resultados muestran que es posible conseguir resultados similares a las discrepancias entre segmentaciones manuales, simplemente refinando las fronteras en base a características fonéticas usando un árbol de decisión.

La mayoría de los métodos presentes en la literatura se han comparado mayoritariamente con segmentaciones manuales. Aquí se ha mostrado que esto no es suficiente para garantizar una mejora real en el funcionamiento del sistema. Por ello, recomendamos, tal y como se propone en [2], el uso de tests perceptuales para la evaluación de como nuevos sistemas de segmentación consiguen influenciar el funcionamiento final del sistema global.

6. Referencias

[1] A. Marzal and E. Vidal, “A review and new approaches for automatic segmentation of speech signals,” in *EUSIPCO*, 1990, pp. 43–53, Barcelona, España.

[2] Matthew J. Makashay, Colin W. Wightman, Ann K. Syrdal, and Aliasir Conkie, “Preceptual evaluation of automatic segmentation in Text-to-Speech synthesis,” in *ICSLP*, Octubre 2000, Beijin, China.

[3] P.A. Taylor and S.D. Isard, “Automatic phone segmeta-

tion," in *Eurospeech*, Septiembre 1991, pp. 709–711, Genova, Italia.

[4] D. T. Toledano, "Segmentación y etiquetado fonéticos automáticos.," Tesis Doctoral, Febrero 2001, Universidad Politécnica de Madrid.

[5] John Kominek, Christina Bennet, and Alan W. Black, "Evaluating and correcting phoneme segmentation for unit slection synthesis," in *Eurospeech*, Septiembre 2003, pp. 313–316, Ginebra, Suiza.

[6] Sérgio Paulo and Luís C. Oliveira, "DTW-based Phonetic alignement using multiple acoustics features," in *Eurospeech*, Septiembre 2003, pp. 309–312, Ginebra, Suiza.

[7] José B. Mariño, A.Ñogueiras, P. Pachès, and A. Bonafonte, "The demiphone: an efficient contextual subword unit for continuous speech recognition.," *Speech Comunication*, vol. 32, no. 3, pp. 187–197, Octubre 2000.

[8] Alan W. Black and Kevin Lenzo, "Building voices in the festival speech synthesis," http://www.festvox.org/bsv.

[9] J.A. Gómez and M.J. Castro, "Automatic Segmentation of Speech at the Phonetic Level," in *Structural, Syntactic, and Statistical Pattern Recognition*, T. Caelli et al., Ed., vol. 2396, pp. 672–680. Springer-Verlag, 2002.

[10] R. O. Duda, P. E. Hart, and D. G. Stork., *Pattern Classification*, John Wiley and Sons, 2 edition, 2001.

[11] Jindrick Matousek, Daniel Tihelka, and Josef Psutka, "Automatic segmentation for czech concatenative speech synthesis using statistical approach with Boundary-Specific Correction," in *Eurospeech*, Septiembre 2003, pp. 301–304, Ginebra, Suiza.

[12] D. T. Toledano, A. Hernández Gómez, and Luis Villarrubia Grande, "Automatic phone segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 6, pp. 617–625, Noviembre 2003.

[13] Antonio Bonafonte, Albino Nogueiras, and Adrián Rodríguez-Garrido, "Explicit segmentation of speech using gaussian models," in *ICSLP*, Octubre 1996.

[14] Jordi Adell and Antonio Bonafonte, "Towards phone segmentation for concatenative TTS," in *the 5th ISCA Workshop on Speech Synthesis*, Julio 2004, pp. 139–144, Pittsburgh, Pennsylvania.

[15] Andreas Kipp, Maria Barbare Wesenick, and Florian Schiel, "Pronunciation modeling applied to automatic segmentation of spontaneous speech," in *Eurospeech*, 1997, Rodas, Grecia.

[16] John Kominek and Alan W. Black, "Impact of durational outlier removal from unit selection catalogs," in *the 5th ISCA Workshop on Speech Synthesis*, Julio 2004, pp. 155–160, Pittsburgh, Pennsylvania.