Desarrollo de un Segmentador Automático de Voz mediante Modelos Ocultos de Markov.

Luis Fernando D´Haro
Docente Universidad Autónoma de Occidente
lfdharo@cuao.edu.co

Resumen

Hoy por hoy la mayoría de los sistemas de síntesis y reconocimiento automático de voz se fundamentan en las técnicas de aprendizaje estocástico a partir de bases de datos muy extensas; sin embargo, no basta solamente con disponer únicamente de la alocución, sino que se necesitan también su trascripción y las ubicaciones temporales (segmentación) de los diferentes sonidos dentro de la misma. El problema es que al ser un trabajo muy largo y hecho por expertos, los costos y tiempos se elevan. Para solventar estos inconvenientes, y para mejorar la calidad de los sistemas, generalmente se utilizan los modelos ocultos de Markov, ya que permiten obtener muy buenos resultados de forma automática.

El artículo presenta un breve estudio del estado del arte en el desarrollo de segmenatdores automáticos y la utilidad de los mismos. Luego se describe el modelo de Markov empleado, su número de estados y transiciones, la base de datos utilizada, los vectores de características, el software usado, así como las estrategias de entrenamiento y reconocimiento. Luego se presentan los diversos experimentos realizados y los resultados alentadores obtenidos (un 79,81% de las marcas en la banda de error menor a 20 ms). Finalmente se dan algunas conclusiones y líneas futuras

.

1. Introducción

Una de las más grandes necesidades al trabajar en las tecnologías del habla es la de contar con bases de datos etiquetadas y segmentadas, ya que estas son indispensables para realizar el entrenamiento de la mayor parte de los modelos estocásticos que se trabajan actualmente. Sin embargo, al ser este un trabajo hecho de forma manual por expertos (fonetistas), de requerir de grandes periodos de tiempo para ser completados y al ser, muchas veces, de carácter subjetivo, los costos de producción, así como los errores humanos asociados, se incrementan; De allí la necesidad de buscar formas automáticas de realizar este trabajo, que al menos permitan dar una primera aproximación para que luego pueda ser "retocada" manualmente, o bien que genere aproximaciones muy cercanas al de las personas.

El presente trabajo se divide en 5 secciones principales, la primera presenta algunas de las aplicaciones en el campo de las tecnologías del habla en las que se requiere de bases de datos segmentadas, así como algunos conceptos básicos necesarios para enmarcar este trabajo. En la segunda sección se presenta un breve estudio del estado del arte, en el que se describen trabajos relevantes que sirvieron de base para el presente proyecto. En la tercera parte se hace una descripción básica de los modelos de Markov, la base de datos empleada, la parametrización de la voz y la herramienta de software utilizada; en la cuarta sección se presentan los diversos experimentos realizados, así como los resultados alcanzados. Finalmente se dan algunas conclusiones y líneas futuras del proyecto.

2. Conceptos básicos y justificación del proyecto

2.1 Definición de Segmentación y Etiquetado automático

Antes de comenzar es menester diferenciar entre los términos segmentación automática y etiquetado automático, pues aunque son fácilmente confundibles, especialmente en la literatura relacionada, son planteamientos distintos por lo que requieren ser definidos, máxime cuando se espera delimitar los alcances y metodología del presente proyecto. Las siguientes definiciones han sido extraídas de [1].

"Se puede definir el problema genérico de la segmentación automática de voz como el problema de determinar con la mayor precisión posible y de forma automática, sin intervención humana, las fronteras temporales entre sonidos correspondientes a unidades de cierto tipo y que forman por concatenación un determinado fragmento de voz humana. Para ello se tomarán como datos de partida dicho fragmento de voz, y a veces, otras informaciones"

"Se puede definir el problema genérico del etiquetado automático de voz como el problema de determinar de forma automáticas, es decir, sin intervención humana, la secuencia de etiquetas de las unidades de cierto tipo a las que corresponden los sonidos que forman, por concatenación, un determinado fragmento de la voz

humana. Para ello se tomarán como datos de partida dicho fragmento de voz y, a veces, otras informaciones".

A partir de estas definiciones, se entiende que el objetivo de este trabajo no es el etiquetado automático de la voz, sino por el contrario se parte de él para determinar la ubicación temporal de los segmentos fonéticos de la voz, mediante el alineamiento forzado entre una alocución grabada y su trascripción fonética.

2.2 Utilidad de las bases de datos segmentadas

Sumado a lo dicho en la introducción, la segmentación de una base de datos es un proceso al que le surgen otros problemas: Se requiere de mucha habilidad y conocimiento específico para identificar correctamente la porción de señal que se corresponde con cierto símbolo en el texto, específicamente en habla continua, y aún más si se hace de forma rápida o relajada. Dado que es un trabajo arduo, dispendioso y largo, siempre se presentan errores humanos asociados con tareas tediosas o con decisiones subjetivas. De allí que se requieran sistemas fiables automáticos que disminuyan estos problemas, o al menos que sirvan de base o comparación con los trabajos hechos por los expertos.

Los campos de aplicación en los que se requieren bases de datos, también son un justificante para emprender este tipo de proyectos, a continuación se mencionan algunos cuantos.

1.Síntesis de voz: Los sistemas de conversión texto a voz, emplean básicamente dos técnicas complementarias, la primera es la del modelado acústico fonético y la segunda el modelado prosódico. La primera busca que las características acústicas de los sonidos sintetizados puedan ser identificados con la secuencia de fonemas deseados. En la segunda, se procura que dichos fonemas se sinteticen de forma que la amplitud, duración y frecuencia fundamental reproduzcan las variaciones que se producen en la voz natural. Ambos procedimientos requieren bases de datos segmentadas y etiquetadas. Los sintetizadores convencionales se basan en la concatenación de unidades, para ello requieren seleccionar las unidades que mejor representen una determinada realización sonora; entre más exacta la segmentación mejor es la selección. Por el otro lado, para realizar modificaciones prosódicas automáticas es preciso utilizar modelos entrenados con una gran cantidad de información proveniente de las bases de datos segmentadas y etiquetadas.

2. Reconocimiento de voz: En este campo, se impone, prácticamente, las técnicas de reconocimiento estadístico de patrones. Dado que se pueden plantear sistemas de reconocimiento basados en palabras, fonemas, trifonemas u otros, se requiere tener segmentación y etiquetado en diferentes niveles lo que obligaría a realizar diferentes trabajos según el caso. Además, el emplear la segmentación se garantiza un entrenamiento más preciso y consistente.

3. Identificación y verificación del locutor: Si bien este campo requiere muchos datos, y no necesariamente estos segmentados, puede resultar interesante la información de la segmentación y el etiquetado, para basar la identificación del locutor en ciertas clases de fonemas, o bien para tratar de forma diferenciada las distintas clases de fonemas.

3. Resumen del estado del arte consultado en segmentación

Antes de proceder a la descripción del estado del arte, es necesario advertir dos cosas: Primero que la mayor parte de los trabajos en esta área son para idiomas distintos al español, por lo que hay que ser cuidados con las conclusiones a las que se pueden llegar; segundo, que la gran mayoría presentan algoritmos posteriores de corrección a la segmentación automática, para así obtener mejores resultados, este proceso aún no se ha realizado para el presente proyecto y se presenta como una línea futura de trabajo.

En [2] se utilizan HMM independientes del contexto, entrenados con voz segmentada y etiquetada fonéticamente de forma manual, y que considera como datos de partida la voz y la trascripción ortográfica. Compara los resultados obtenidos al permitir o no transcripciones fonéticas alternativas y pausas entre las palabras en la gramática del reconocedor, y observa que la precisión de la segmentación aumenta al permitir transcripciones alternativas y pausas opcionales. También observan que al considerarse conocida la trascripción fonética manual se obtienen los mejores resultados para la segmentación. Emplean la base de datos TIMIT (habla continua) para el inglés y obtienen un 86,2% de fronteras fonéticas con errores menores a los 20 ms de la posición de las fronteras manuales de referencia. Para el italiano obtienen un 90,9% a menos de 20 ms.

En [3] se utilizan HMM independientes del contexto, entrenados con voz segmentada y etiquetada fonéticamente de forma manual. Se comparan resultados de segmentación con modelos independientes y dependientes de locutor. Se obtienen resultados de 90,5% con una tolerancia de 20 ms, aunque baja para otros locutores alcanzando un 88,6% para una locutora femenina.

En [4] Usa HMM independientes del contexto, compara con modelos independientes y dependientes del locutor, utiliza transcripciones fonéticas manuales, pronunciaciones alternativas y silencios opcionales entre palabras. Obtiene mejores resultados con pronunciaciones alternativas y con la trascripción fonética manual, sin considerar silencios opcionales y sin adaptación al locutor. Obtiene un 93,42% con un 10% de las fronteras fonéticas automáticas conteniendo un error mayor que 34.4 ms.

En [5] Se emplea HMM independientes del contexto entrenados con la voz del mismo locutor sobre el que se

realizan pruebas. Supone conocida la trascripción fonética y obtienen un resultado de un 70% de las marcas automáticas con errores inferiores a 20 ms. Este es uno de los enfoques mas parecidos al presentado en este trabajo.

En [6] Se emplean HMM independientes del contexto y del locutor. Suponen conocida la trascripción ortográfica y consideran que a partir de ella pueden generar la fonética. La precisión es del 90% de las marcas con errores menores a 35 ms, empleando la base de datos TIMIT.

En [1] Propone un sistema para el español, que emplea HMM dependientes e independientes del contexto, con diversas variaciones en el número de mezclas de Gaussianas, así como modelos dependientes e independientes del locutor, empleando diversos algoritmos de adaptación de locutor, además de cancelación de errores sistemáticos de segmentación. Emplea diversas bases de datos, en su mayoría con habla aislada y algunas con habla continua, obteniendo así un 91% de marcas con errores menores a los 20 ms. Luego intentan probar redes neuronales, algoritmos difusos y estadísticos de características acústicas de la voz, para disminuir los errores en las fronteras llegando a obtener un 96,2 % de marcas con errores menores a 20 ms. Además propone como figura de mérito, para comparar los diversos sistemas, usar el promedio de marcas entre 0 y 100 ms, asignando igual peso a todos los porcentajes.

Existen otra serie e trabajos empleando redes neuronales tanto para la segmentación, como para el postproceso, así como el uso de alineamientos temporales, los cuales no se mencionan, ya que en general permiten obtener resultados inferiores a los conseguidos con los HMM.

4. Descripción del método de segmentación automática

4.1 Base de datos empleada

Se empleó la base de datos denominada Natvox, que pertenece al Grupo de Tecnología del Habla (GTH) de la Universidad Politécnica de Madrid. Esta presenta las siguientes características:

1.Voz femenina, que ha sido ampliamente utilizada en estudios sobre modificaciones de frecuencia fundamental y de duraciones.

2. Frases de habla continua con tres tipos de texto: El primero es en discurso, dos coloquiales en forma de entrevista, incluyendo frases interrogativas y exclamativas. Finalmente, un grupo de frases de laboratorio que incluyen muestras de cada uno de los esquemas simples posibles con una o dos sílabas tónicas, en estructuras independientes de hasta 8 sílabas.

3. Un total de 732 frases (con 15.141 fonemas) de los cuales se utilizaron 1.072 para el conjunto de prueba y de 4.252 para el entrenamiento. Las frases han sido transcritas a nivel grafémico, y fonético. Aunque en este proyecto se trabajó con el grafémico, para hacer aún más robustas las conclusiones.

4.Grabaciones en formato PCM de un solo canal, con frecuencia de muestreo de 32 KHz (remuestreadas luego a 16 KHz), sin compresión y con 16 bits de resolución.

4.2Modelos Ocultos de Markov

Un modelo de Markov es un modelo paramétrico que permite describir hechos acústicos del habla, y que está caracterizado por una serie de variables estadísticas. Un modelo de Markov estará formado por:

Un número determinado de estados, cada uno caracterizada por una determinadas probabilidades tanto de emisión, como de transición entre estados. Se produce un cambio de estado cada cierto tiempo T, según la probabilidad de transición que depende del estado del cual parte. Después de cada transición se produce una observación de salida que depende del estado en que se produce y no de cómo se ha llegado hasta él. Esta observación se produce siguiendo una función de densidad de probabilidad, que generalmente se asume como Gaussiana.

Un modelo de Markov está formado por dos procesos estocásticos, uno conocido, que es la producción de los símbolos o salidas, y otro oculto, que es el paso de unos estados a otros. El hecho de que ni la observación, ni la probabilidad de transición dependan de cómo se ha llegado al estado emisor, limita la memoria del modelo, pero reduce la complejidad del mismo. Para intentar resolver este inconveniente se utilizan la primera y segunda derivada de los vectores de características que se obtienen del habla.

En este proyecto se han escogido dos tipos de HMM, el primero es de 5 estados emisores, con posibilidad de transiciones al mismo estado, simples y dobles de izquierda a derecha, y sin transición hacia atrás, este es el tipo de modelo para todos los fonemas, figura 1. Para modelar los silencios o pausas, se escogió un modelo de 3 estados, con transiciones a sí mismo, simples, dobles y hacia atrás (para compensar duraciones muy altas de silencios).

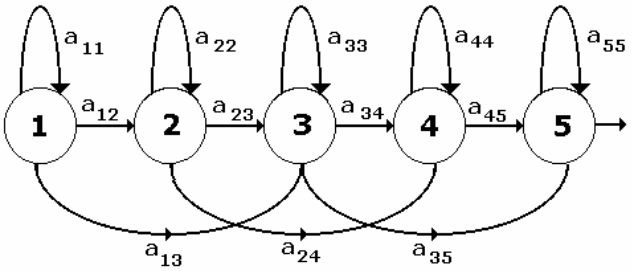

Figura 1. Modelo de Markov utilizado para los todos los fonemas, excepto el de silencio.

4.3 Parametrización de la voz.

Uno de los problemas que presentan los HMM en la detección de fronteras fonéticas, que al requerir trabajar en el dominio de la frecuencia para hallar los vectores de características, la ventana de análisis suele ser mayor a los 20 ms, por lo que se pierde resolución temporal. Para remediar esto, se probó una técnica que [1] menciona, y es la de emplear desplazamientos cortos de la ventana, típicamente entre 2 y 5 ms. Además algunos autores mencionan que al usar modelos dependientes se tiene menos precisión, hipótesis que otros desmienten, por lo que se optó con probar con ambos tipos. En este contexto, se trabajó con la siguiente parametrización:

Formato de archivo: WAV
Longitud de la ventana de análisis: 25 ms
Desplazamiento de la ventana de análisis: 3 ms
Parámetros: 12 PLP + Energía (normalizada), parámetros de velocidad y aceleración (total 39)
Ventana de Hamming
Preénfasis con factor de 0,97.
Número de bandas en la escala MEL: 24.
Frecuencia de corte inferior de 125 Hz, frecuencia de corte superior de 3800
Eliminación de nivel DC y normalización cepstral (CMN) y (CVN).

4.4 Software empleado

Se trabajó con el paquete HTK v.3.1 desarrollado por Cambridge University Engieneering Department. Esta herramienta corre bajo Unís, presentando una gran cantidad de funciones necesarias para desarrollar un sistema de reconocimiento de voz, especialmente modelos continuos; permite trabajar con habla aislada, continua, HMM de distintas topologías, modelos dependientes e independientes del contexto, cálculo de vectores de parametrización. Adicionalmente el programa permite realizar aumento del número de Gaussianas o mezclas, calcular CMN y CVN, inicialización con Viterbi o Baum-Welch, realizar árboles de decisión, enlazado (tying), clonación de modelos, lo cual permite partir de modelos de monofonemas hacia trifonemas, incluido clustering de los mismos, y finalmente realizar la evaluación del reconocimiento del sistema.

4.5 Experimentos y resultados

El primer experimento realizado consistió en emplear un modelo de 24 fonemas independientes del contexto (incluidos el modelo de silencio y el de pausa corta), que empleaba como semilla de inicialización, la segmentación manual de toda la base de datos. El modelo se re-estima 4 veces por cada mezcla adicional que se le agrega al modelo inicial. El número final de mezclas es de 10. La figura 2, muestra el histograma acumulado del porcentaje de marcas con diferencias inferiores a 100 ms. Generalmente se considera un buen indicador el porcentaje obtenido en la banda de los 20 ms. La gráfica muestra el efecto de aumentar el número de Gaussianas a los CHMM. Se observa que al aumentar el número de mezclas los resultados son peores, esto se debe al bajo número de unidades de entrenamiento. La tabla 1 permite observar que, en general, el empleo de 2 mezclas ayuda a aumentar el porcentaje de marcas.

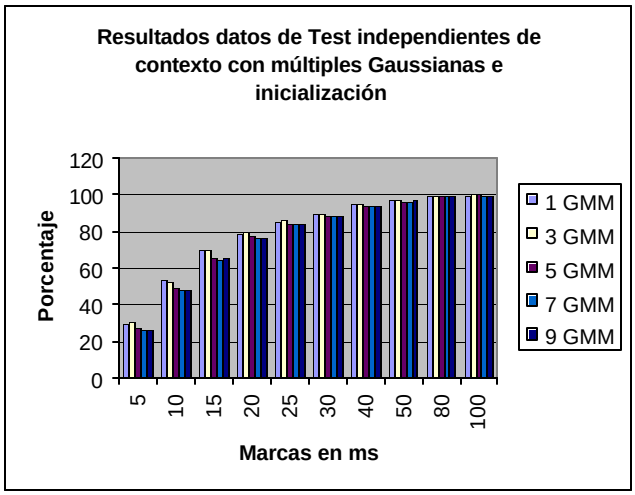

Figura 2. Resultados de los modelos independientes del contexto con inicialización

No. de Mezclas	Porcentaje en 20 ms
1 GMM	78,83
2 GMM	**79,81**
3 GMM	79,72
5 GMM	77,45
7 GMM	75,94
9 GMM	76,16

Tabla 1. Porcentaje de marcas con errores inferiores a 20 ms en función del número de mezclas, para modelos independientes con inicialización.

El segundo experimento, consistió en emplear los mismos 24 fonemas, pero sin emplear la semilla para inicializar, sino que al algoritmo de Viterbi se le indicó que segmentara automática de forma equidistante. Como en el caso anterior, se probó aumentando el número de Gaussianas. La gráfica 3 muestra que, en general, se obtienen mejores resultados con un número bajo de mezclas. Sin embargo, la tabla 2 muestra una leve disminución en el porcentaje de marcas con errores por debajo de los 20 ms. Este resultado es importante porque

muestra que es casi innecesario tener una segmentación manual de la base de datos. Es necesario confirmar estos datos mediante el empleo de porcentajes variables de inicialización para confirmar este resultado.

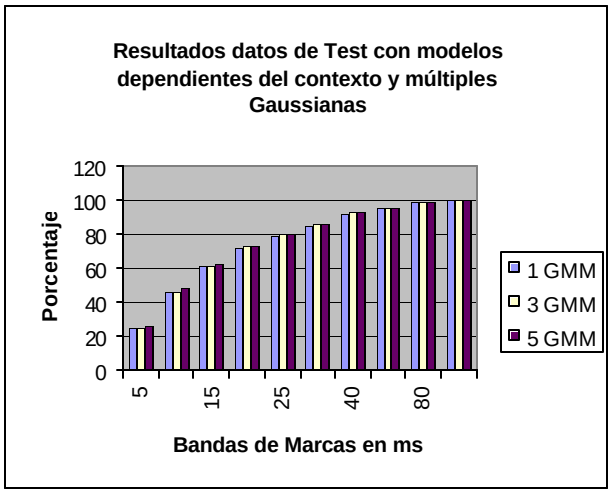

Figura 3. Resultados de los modelos independientes del contexto sin inicialización.

No. Mezclas	Porcentaje en 20 ms
1 GMM	75,47
3 GMM	75,37
5 GMM	**76,40**
7 GMM	75,84
9 GMM	76,40
10 GMM	75,65

Tabla 2. Porcentaje de marcas con errores inferiores a 20 ms en función del número de mezclas, para modelos independientes sin inicialización.

Finalmente, se probó con modelos dependientes del contexto, en este caso trifonemas, sin inicialización. Para ello se partió de los modelos de monofonemas (excepto el de silencio), los cuales fueron clonados para inicializar los modelos de los trifonemas, luego dichos modelos son re-estimados una vez, para posteriormente realizar un clustering mediante un árbol de decisión binario top-down, con el objetivo de agrupar los modelos que fueran similares; dado el bajo número de datos de entrada, se consiguió un total de 244 modelos de trifonemas., a los cuales se les aumentó hasta 6 mezclas.

La figura 4, muestra los resultados obtenidos para el modelo de trifonemas. Se puede observar que hay una mayor uniformidad en los resultados, con una ligera preferencia por un valor medio de Gaussianas (3 o 4). La tabla 3 muestra los resultados en la banda de 20 ms. Una causa por la cual este experimento no muestra mejores resultados fue el que no se empleó el modelo de silencio, de haberlo hecho es muy probable que se hubieran obtenido mejores resultados.

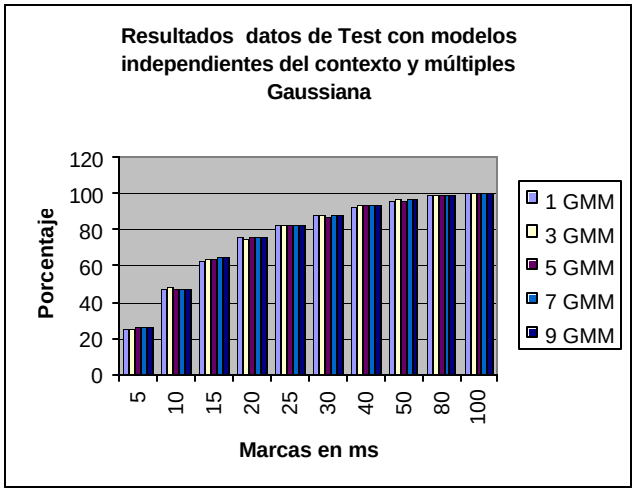

Figura 4. Resultados de los modelos dependientes del contexto sin inicialización.

No. de mezclas	Porcentaje en 20 ms
1 GMM	71,92
2 GMM	72,29
3 GMM	72,67
4 GMM	73,41
5 GMM	**73,51**
6 GMM	73,51

Tabla 3. Porcentaje de marcas con errores inferiores a 20 ms en función del número de mezclas, para modelos dependientes sin inicialización.

5. Conclusiones y líneas futuras de desarrollo

De los resultados obtenidos se puede concluir lo siguiente:

1. Es necesario emplear una mayor cantidad de datos, empleando si es posible toda la base de datos Natvox.
2. Se requiere adicionar una mayor cantidad de modelos a entrenar, entre los que se sugieren el modelo de vocal acentuada, vocal nasalizada y consonantes fricativas aproximantes. Las cuales no se implementaron por la poca cantidad de datos empleados.
3. Igual que para otros trabajos, el empleo de modelos independientes del contexto parecen ser mejores para la segmentación. Sin embargo, dado el número distintos de modelos de

trifonemas obtenidos y la poca cantidad de datos, esta conclusión puede ser precipitada.

4. Al comparar los resultados de hacer inicialización de los modelos, con los que no la empleaban, tanto para el modelo de monofonemas, como el de trifonemas, se observa que no hay una gran diferencia en los resultados. Lo anterior, como ya se ha mencionado, es prometedor, pues se probó inicializando con el 100% de la base de datos segmentada manualmente, y sin ningún tipo. Sin embargo, al igual que para las anteriores conclusiones, el bajo número de datos hace que sea un poco premeditado el concluirlo, haciéndose necesario probar con porcentajes variables de inicialización en busca de un compromiso.
5. Que el modelo de silencio o de pausa hace que los resultados aumenten considerablemente, pues aunque no se muestran en este documento, los resultados obtenidos sin usarlo son por lo menos 10 puntos por debajo de los que aquí se presentan. De haberse usado este modelo en el uso de los trifonemas es muy probable que los resultados hubiesen sido mas altos. Adicionalmente, el empleo de transcripciones alternativas también mejora los resultados del alineamiento.
6. Finalmente, que el uso de modelos con una gran cantidad de mezclas no mejora los resultados, siempre y cuando el número de datos de entrenamiento no sea alto, ya que se generan mezclas demasiado pegadas a los datos de entrenamiento, perdiéndose en la generalización del modelo.

Los posibles trabajos futuros que se proponen, y que ya están en proceso son los siguientes:

1. Aumentar la base de datos de entrenamiento y test.
2. Adición de nuevos modelos de fonemas, como los que ya se mencionó antes.
3. Probar con distintos tipos de HMM y desplazamientos de la ventana, pues por lo que se alcanzó a percibir, en este caso se hacia casi innecesario el trabajar con tramas tan cortas, pues la parametrización era muy redundante. Se observó que los saltos dobles en los HMM, eran los que presentaban más alta probabilidad, lo cual confirma el manejar desplazamientos más grandes o reducir el número de estados del modelo.
4. Realizar postprocesado con redes neuronales, para reducir los errores en las fronteras.
5. Inicializar los distintos modelos con un porcentaje variable de la base de datos segmentada y etiqueta manualmente.
6. Finalmente, probar con un diccionario de etiquetado automático y con distintos niveles de trascripciones alternativas.

6. Bibliografía

[1] [TOR01] TORRE, Toledano Doroteo. Segmentación y Etiquetado Fonéticos Automáticos: Un enfoque basado en Modelos Ocultos de Markov y Refinamiento posterior de las Fronteras Fonéticas.

[2] [ABFGG093] ANGELINI, B; BRUGNARA, FALAVIGNA, F; GIULIANI, D; GRETTER D y OMOLOGO M. Automatic segmentation and labelling of English and Italian Speech Databases, In Proceedings EUROSPEECH 1993, pp 653 – 656.

[3] [ABFOS97] ANGELINI, B; BRUGNARA, FALAVIGNA, F; OMOLOGO M y SANDRI, S. Automatic diphone extraction for an Italian Text-To-Speech System, In Proceedings EUROSPEECH 1997, vol II, pp 581 – 584.

[4] [CBJ98] COX, S; BRADY, R y JACKSON, P. Techniques for accurate automatic annotation of speech waveforms, In Proceedings of the International Conference on Spoken Language Processing, 1998 Sydney (Australia), Vol V, pp 1947 – 1950.

[5] [LHS97] LJOLJE, A; HIRSCHBERG, J y VAN SANTEN, JPH. Automatic speech segmentation for concatenative inventory selection, In Van Santen JPH et al (eds), Progress in Speech Synthesis, Springer 1997, pp 305 – 311.

[6] [WT97] WIGHTMAN, CW y TALKIN, DT. The Aligner: Text-To-Speech alignment using Markov Models, In Van Santen JPH et al (eds). Progress in Speech Synthesis, Springer, 1997, pp 313 – 323. [RAB89]

[7] RABINER, LR. A Tutorial on Hidden Markov Models and Selected Applications in Speech Reconigtion. In Proceedings of the IEEE 77 (2). Feb 1989.