Decodificador Eficiente para Normalización del Tracto Vocal en Reconocimiento Automático del Habla en Tiempo Real

*Antonio Miguel, * Richard Rose, Eduardo Lleida,*
Luis Buera, Alfonso Ortega, Oscar Saz

Departamento de Electrónica y Comunicaciones
Universidad de Zaragoza
* Department of Electrical and Computer Engineering
McGill University, Canada
{amiguel, lbuera, lleida, ortega, oskarsaz}@unizar.es, * rose@ece.mcgill.ca

Resumen

La normalización del tracto vocal es un proceso utilizado con éxito hasta la fecha para aumentar las prestaciones en sistemas de reconocimiento automático del habla con el objetivo de reducir la variabilidad inter-locutor. En el presente artículo se describe un sistema que evita el tener que hacer varios procesados de las frases para poder hacer la tarea en tiempo real. El sistema presentado se basa en la decodificación simultánea de los estados del modelo de Markov convencional con un conjunto de posibles transformaciones lineales frecuenciales. La flexibilidad temporal de la decodificación conjunta y de la normalización del tracto vocal trama a trama proporciona una mejora de modelado sobre el sistema original sin un excesivo costo computacional adicional.

1. Introducción

Una de las fuentes de variabilidad en reconocimiento automático del habla por medio de Modelos Ocultos de Markov es la variabilidad entre distintos hablantes debido a la forma y tamaño de su tracto vocal. La normalización del tracto vocal es una técnica bien conocida y desarrollada por diversos autores [1, 2, 3, 4, 5, 6], que intenta disminuir ese efecto.

La motivación común de las técnicas normalización del tracto vocal reside en el modelo de producción de la voz, y en concreto, en la formación de los sonidos llamados sonoros, cuya envolvente del espectro viene marcada por las resonancias del tubo que forma el tracto vocal al ser excitado por una onda de presión que vibra según una frecuencia, llamada *pitch* o tono fundamental, que le imprimen las cuerdas vocales. Como se intentará mostrar más adelante, los intentos de modelado para mejorar la independencia del locutor de los sistemas basándose en esta descripción no son tan adecuados para sonidos sordos o transiciones entre vocales por medio de consonantes.

Veamos una breve descripción de las aproximaciones que se han ido tomando para resolver el problema. En los primeros intentos [1, 5], se trataba el problema de la normalización desde un punto de vista de identificación de vocales y posterior medición de las diferencias entre sus formantes como medio para llegar a una estimación de la longitud del tracto vocal, con el posterior intento de compensar estas variaciones en las posiciones de los formantes. Las técnicas que se basan en la estimación de los formantes como medio para estimar el tracto vocal, suelen tener problemas, ya que es una tarea difícil y suele conllevar poca robustez al ruido.

Posteriormente, en [4] aparece un intento de normalización del tracto vocal con un método de máxima verosimilitud, donde ya no se trabaja a nivel de formantes directamente, sino de parámetros acústicos, de una manera más parecida a como funcionan los sistemas de reconocimiento habituales. En concreto, los vectores acústicos se obtenían gracias a una transformación lineal del eje de frecuencias que se aplicaba haciendo un remuestreo de la señal de audio original, lo cual era bastante costoso computacionalmente.

Este método se sofisticó en [6], por medio de uno mucho más eficiente, que presentaba un esquema de entrenamiento de modelos y de reconocimiento adaptados a la técnica de la normalización del tracto vocal. La forma en que se aplica la transformación frecuencial en este método es más eficiente y es la que se ha empleado en este trabajo, consiste, como se verá posteriormente, en la modificación del banco de filtros utilizado normalmente en muchas de las parametrizaciones más habituales.

El método propuesto en este artículo intenta evitar algunos de los problemas o restricciones de estos métodos, que suelen hacer una estimación media del factor de transformación para una frase completa. Se trata de mejorar dos problemas, el primero es que se necesita mucha señal para tener la estimación, es decir se trabaja con la estimación de la frase anterior o se introduce un gran retardo y el segundo es que al calcular la verosimilitud de una frase no se hace distinción entre los fonemas o sonidos que la componen o la cantidad de silencio que hay, ya que, como se intentará mostrar con el caso concreto del silencio, el hecho de no tenerlo en cuenta puede introducir un sesgo en la estimación del factor de transformación frecuencial dependiente de la proporción de voz y silencio de cada frase.

Para solucionar en parte estos problemas se propone una modificación de los modelos y del algoritmo de búsqueda añadiendo un nuevo grado de libertad para la transformación frecuencial, de manera que se optimice conjuntamente la verosimilitud para los estados del modelo de Markov y los factores de transformación. Con esta modificación se consigue una estimación y reconocimiento simultáneos y se modela de una manera más detallada las transiciones entre vocales, con una serie de factores estimados que varía con el tiempo, no un promedio para toda la frase y todos los estados. Además, se tratará de evitar el problema del sesgo para el silencio. Este tipo de sistemas ha aparecido con anterioridad en [7], pero como veremos el modelo propuesto difiere en algunos aspectos que

veremos posteriormente.

El artículo está organizado de la siguiente manera: en la sección 2 se describe los métodos anteriores de normalización del tracto vocal, en la sección 3 se describe el método propuesto y en la sección 4 se presentan algunos resultados de reconocimiento. Por último se extraen una conclusiones sobre el trabajo realizado.

2. Transformación frecuencial lineal

En esta sección se describe la forma en que se ha implementado la transformación lineal básica y algunas de las primeras consecuencias que se pueden observar de su aplicación.

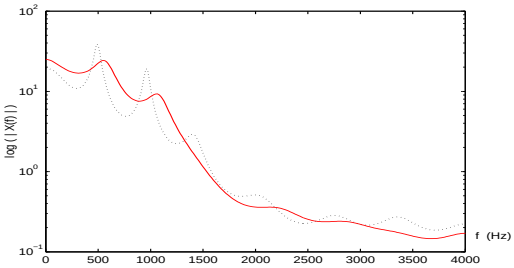

Figura 1: Comparación de las envolventes de los espectros de dos fragmentos de dos realizaciones por dos locutores de la vocal u en la palabra ”two”(obtenido por lpc orden 16).

En la figura 1, podemos observar que en una serie de realizaciones de la misma vocal (en este caso la u), los formantes principales de la envolvente no coinciden exactamente, causando en los modelos un desajuste inter-locutor debido principalmente al tracto vocal.

Así pues, para modelar estas posibles variaciones tenemos que añadir algún grado de flexibilidad en el proceso de aprendizaje-reconocimiento que compense esa variabilidad. Centramos nuestra atención en el proceso de parametrización y aplicaremos la transformación en el banco de filtros descrita en [6], ya que es muy eficiente computacionalmente y será más conveniente que otras técnicas para aplicar en el método propuesto.

En primer lugar definimos la función de transformación frecuencial, que es la función por la cual cada componente espectral del espectro original tiene asociada su imagen en el espectro transformado. En este caso es una función lineal por partes $g^{\alpha}(f)$, definida de la siguiente manera:

$$g^{\alpha}(f) = \begin{cases} \alpha f & \text{si } f <= f_0, \\ \alpha f_0 + \frac{f_{max} - \alpha f_0}{f_{max} - f_0}(f - f_0) & \text{si } f > f_0. \end{cases} \quad (1)$$

Donde f_{max} es el valor de $N_{FFT}/2$, con N_{FFT} el número de puntos de la transformación de frecuencia. El valor del punto de inflexión de las dos rectas:

$$f_0 = \begin{cases} \alpha f_{max} \frac{7}{8} & \text{si } \alpha <= 1, \\ f_{max} \frac{7}{8\alpha} & \text{si } \alpha > 1. \end{cases}$$

La transformación del banco de filtros la realizamos aplicando la función de transformación de frecuencia, $g^{\alpha}(f)$, en la función que utilizamos para calcular los centros de los filtros promediadores del espectro del banco de filtros, que en nuestro caso es:

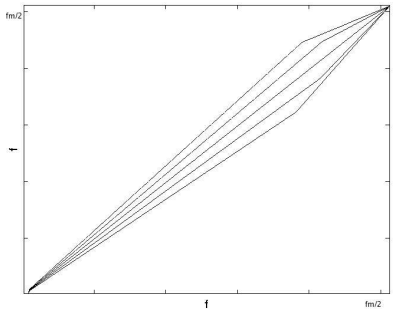

Figura 2: Conjunto de funciones de transformación $g^{\alpha}(f)$.

$$f(n) = \begin{cases} 100 \cdot n & \text{si } n <= 10, \\ 100 \cdot 10 \cdot 1{,}1^{n-10} & \text{si } n > 10. \end{cases} \quad (2)$$

Con esta función que proporciona la posición de los centros para los filtros, generamos el banco de filtros como una matriz B con los filtros triangulares organizados por filas, así, el tamaño de la matriz es $(b \times f_{max})$, con b el número de filtros y f_{max} la mitad del número de puntos de la transformación a frecuencia.

Una vez definida la función de transformación, aplicamos una serie de N pendientes, $\mathcal{A} = \{\alpha_i\}_{i=1}^{N}$,en la función de transformación, obtenemos un conjunto de funciones de transformación (ver figura 2) $\mathcal{G} = \{g^{\alpha_i}\}_{i=1}^{N}$ y con él podemos generar una serie de bancos de filtros diferentes $\mathcal{B} = \{B^{\alpha_i}\}_{i=1}^{N}$, que aplicaremos a la señal para obtener N distintas parametrizaciones de la misma.

Los métodos que se presentan a continuación se basan en la maximización de la verosimilitud como medio para reducir el desajuste debido al tracto vocal. Al comparar los vectores acústicos normalizados con los mismos modelos que utilizamos para reconocer, tenemos una medida del grado de ajuste de los vectores acústicos obtenidos para cada función de transformación del conjunto $\mathcal{A}$ de una manera coherente con la escala de medidas que se utilizan en el sistema de reconocimiento.

A continuación vemos cómo estiman las transformaciones óptimas y se utilizan para normalizar la variabilidad del tracto vocal.

2.1. Estimación del factor de la transformación lineal

El factor de la transformación indica, en cierta medida, la relación que existe entre las longitudes de los tractos vocales de los distintos locutores, a través de la relación que existe entre los formantes principales de ciertas realizaciones vocálicas. Los métodos orientados a conseguir estimaciones precisas de dichos formantes son costosos y poco robustos. En el método revisado en este trabajo la comparación entre las distintas realizaciones se realiza de forma implícita al utilizar modelos estadísticos como una medida del grado de desajuste. Una de las ventajas consiste en que son los mismos que en el sistema de reconocimiento.

Puesto que el proceso se basa en la maximización de la verosimilitud de unos vectores acústicos, describimos brevemente el proceso para obtener esos vectores. Dada una secuencia de muestras de audio, se utiliza una ventana de Hamming deslizante para obtener una secuencia de tramas. A continuación se obtiene el módulo de la transformación discreta de frecuencia para cada trama. A esta serie de vectores

la llamamos $X_F = \{x_{F_1}, x_{F_2} ... x_{F_L}\}$, con L el número de tramas. A esta serie de tramas de frecuencia le aplicamos las transformaciones de frecuencia por medio del conjunto de bancos de filtros antes descrito $\mathcal{B}$, para después de aplicar la transformada coseno DCT(*Discrete Cosine Transform*), obtener varias versiones parametrizadas de la señal original, $\mathcal{C} = \{C^{\alpha_i}\}_{i=1}^{N}$,según los factores de transformación $\mathcal{A}$, expresamos el proceso en forma matricial de la siguiente manera:

$$\underset{(c \times L)}{C^{\alpha_i}} = \underset{(c \times b)}{DCT} \cdot \underset{(b \times f)}{B^{\alpha_i}} \cdot \underset{(f \times L)}{X_F} \quad (3)$$

Donde:

- c: número de coeficientes cepstrales
- b: número de salidas del banco filtros
- f: número de componentes frecuenciales, $N_{FFT}/2$
- L: longitud en tramas de la frase

Entonces para una frase, una vez tenemos las versiones transformadas de la señal $\mathcal{C}$, y un modelo de Markov entrenado de forma genérica λ y la transcripción T, estimamos el factor de transformación α más apropiado, y para ello procedemos a maximizar, dentro del conjunto de vectores acústicos previamente definido, de la siguiente manera:

$$\hat{\alpha} = \underset{\alpha_i \in \mathcal{A}}{\arg\max} \{P(C^{\alpha_i}|\lambda, T)\} \quad (4)$$

De esta manera se obtiene una estimación del factor de trasformación o *warping* como el de más verosimilitud para un modelo genérico dentro del rango que definimos en $\mathcal{A}$. Así pues, el valor de la estimación obtenida se obtendrá mediante un barrido dentro de un conjunto discreto de valores posibles, normalmente se suele tomar $0{,}88 < \alpha < 1{,}12$, tomando más de 11 pasos.

Posteriormente en [8] se ha determinado soluciones para obtener el fáctor de transformación como una adaptación MLLR con restricciones (*Maximum Likelihood Linear Regression*). Pero, como veremos, el método dependiente del tiempo propuesto no trata el factor de transformación como uno único a lo largo de toda una secuencia del mismo locutor. Ya que se pretende trabajar en tiempo real, además de trama a trama, no podemos usar estos últimos métodos, será más conveniente la extensión del método de Lee-Rose [6].

2.2. Procedimiento de entrenamiento

El proceso de entrenamiento de los modelos se realiza de manera que los nuevos modelos se adapten mejor a los vectores de características transformados. Así, si con el proceso de normalización conseguimos una aproximación entre las realizaciones de los mismos sonidos para distintos locutores, con el entrenamiento apropiado se obtienen unos modelos en los que se ha eliminado esa fuente de variabilidad, podemos decir más independientes del locutor.

El entrenamiento se hace de una manera iterativa, ya que una vez transformamos los parámetros con el factor óptimo, reentrenamos los modelos. Después, volveremos a estimar las transformaciones con los nuevos modelos y volverlos a entrenar. Este proceso iterativo se puede comprobar que es convergente [6].

2.3. Procedimiento de reconocimiento

En esta sección se describe el proceso de reconocimiento para los modelos normalizados. En primer lugar se necesita una estimación del factor de transformación que se acerca más a los modelos, para lo cual según la ecuación 4 necesitamos una transcripción. Se puede ver de esta manera que los métodos de estimación basados en la maximización de la verosimilitud de la frase completa necesitan de una hipótesis de transcripción y por lo tanto varios procesados de la frase.

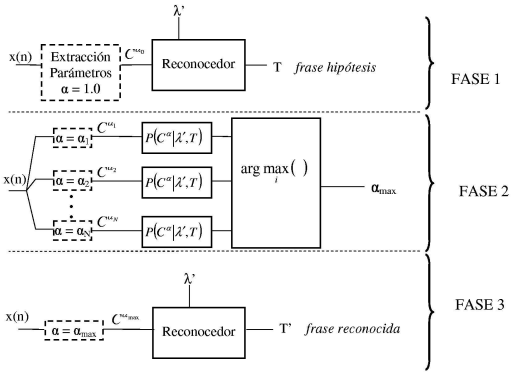

Figura 3: Esquema del proceso de reconocimiento en tres pasos.

Uno de los métodos utilizado en [6] hace los siguientes pasos:

1. Una pasada de reconocimiento para extraer la hipótesis de transcripción. Para ello utiliza un reconocedor convencional.
2. La búsqueda del α_i cuya verosimilitud sea la máxima entre el conjunto de factores de transformación $\mathcal{A}$, según la ecuación 4.
3. El reconocimiento utilizando ahora la versión parametrizada por el factor α_i, esto es C^{α_i} y los modelos normalizados.

2.4. Problemas de la transformación lineal de frecuencia

En trabajos anteriores se ha considerado el factor de normalización como una descripción, que no varía para un mismo locutor, para normalizar el tracto vocal. Si bien es cierto que la longitud del tracto vocal no cambia en el trancurso de un breve espacio de tiempo, la transformación, que aplicamos por igual a todos los sonidos de una realización, no tiene razones fisiológicas para mejorar la verosimilitud de los vectores acústicos más asociados a las consonantes, ya que su producción depende, en general, de la disposión de la boca, labios y dientes, y no tanto del tracto vocal.

2.4.1. El problema del modelo de silencio

Además del problema de los sonidos no vocálicos, díficil de aislar con el método original, existe otro problema que proviene del hecho de que se computen las verosimilitudes con todos los vectores acústicos de una frase completa. Este nuevo problema, no previsto teóricamente, es el de la transformación del silencio. Observando la verosimilitud media de las tramas de silencio (con la segmentación hecha a partir de nuestro modelo original λ) en función de las transformaciones, hemos comprobado que no es independiente del factor de transformación, por el contrario tiene una tendencia hacia los valores altos. Esta tendencia introduce un sesgo en nuestras estimaciones del factor de transformación, según el método propuesto.

Se ha realizado la estimación del factor de transformación óptimo de todo el conjunto de entrenamiento de la base de datos Aurora 2 (ver sección de resultados para descripción), de dos manera: teniendo en cuenta los modelos correspondientes al silencio y descartándolos. Esto se aplica durante el cómputo de la verosimilitud en 4.

El resultado de la media del factor estimado ha sido 1.0116 con silencio y 1.0053 sin silencio. La desviación típica del error producido es 0.0182, la cual es comparable a la diferencia entre factores consecutivos en $\mathcal{A}$, que para $N = 11$ factores es 0.0225. En la figura 4 se representa el histograma de la distribución del error de la estimación calculado frase por frase teniendo en cuenta el silencio y sin tenerlo.

Por lo tanto, se comprueba la existencia de cierto sesgo y de error de estimación que puede deberse a que el ruido en las tramas de silencio no es blanco sino que tiene distorsiones ya sea por el micrófono o la acústica de la sala. Este sesgo perturba la estimación de las tramas que nos interesan, las asociadas a sonidos vocálicos.

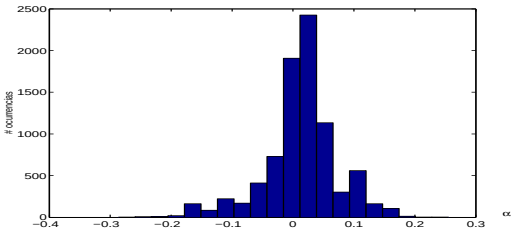

Figura 4: Histograma del error del estimación del parámetro α al incluir el modelo de silencio en la verosimilutud respecto a no incluirlo

3. El normalizado dependiente del tiempo

En el método de normalización propuesto la normalización se realiza trama a trama, con lo cual se pretende que solucione parcialmente los problemas citados anteriormente, tanto realizando una transformación adecuada en los sonidos no vocálicos y no transformando las tramas que se consideren correspondientes al modelo de silencio.

Para conseguir estos objetivos se presentan unos modelos de Markov que añaden un grado de libertad añadido para el rango de factores de normalización, que como hemos visto anteriormente tiene un conjunto finito de valores posibles. Así, este grado extra de libertad se concreta de la siguiente manera, cada estado del modelo de Markov $q_i \in \{q_j\}_{j=1}^{M}$, lo repetimos N veces, siendo N el número de transformaciones posibles para α en $\mathcal{A}$, así que, de un modelo inicial de M estados pasamos a un modelo aumentado de $N' = N \cdot M$. Aunque el número de estados ha crecido, como veremos, la complejidad de la búsqueda no crece en la misma medida, ya que, en primer lugar acotaremos las transiciones posibles, y la búsqueda de Viterbi en haz en este nuevo espacio no activa muchos estados más que en el modelo original.

Este proceso lo hacemos para todos los estados del modelo original excepto para el modelo de silencio, que tendrá el mismo número de estados que en modelo el original porque para él no añadimos este grado de libertad. Se pretende que la única transformación disponible para una trama asociada al modelo de silencio sea $\alpha = 1{,}0$.

En la figura 5 se hace una comparación gráfica del modelo propuesto con una solución intermedia que opera también en tiempo real, pero manteniendo el factor de transformación fijo

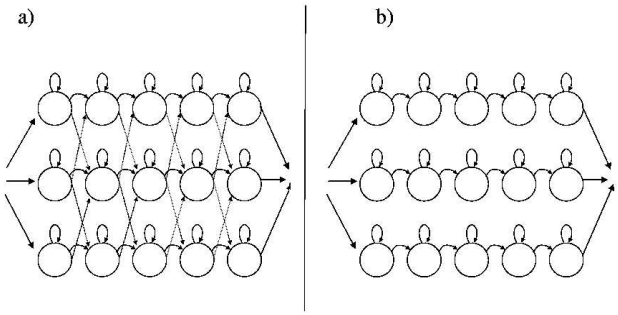

Figura 5: Comparación del método propuesto (b) con un método de normalización en tiempo real con factor de transformación fijo.

a lo largo de la frase. Nos referimos al modelo situado a la derecha en la figura 5, que puede describirse como la operación de N reconocedores en paralelo, tantos como transformaciones posibles en $\mathcal{A}$, con una posterior recombinación por medio de la selección de la solución de mayor verosimilitud. Esta solución es mucho más compleja como demuestran algunas medidas realizadas, ya que obliga a activar siempre todas las transformaciones posibles, cuando en el método propuesto la búsqueda en haz permite descartar los estados menos probables. Además, de la misma manera que sucede en el método original, el factor de transformación permanecería fijo en toda la frase, incluidos los modelos de silencio.

Una ventaja adicional es que el algoritmo de entrenamiento no difiere del estándar utilizado para entenar modelos de Markov.

3.1. Procedimiento de reconocimiento

Una vez definido el nuevo modelo, describimos la aplicación del mismo para estimar y reconocer simultáneamente por medio de una modificación de la recursión del algoritmo de Viterbi original, en las siguientes expresiones podemos ver la notación de la recursión original 5 y la modificada 6:

$$\phi_j(t) = \max_{i \in \mathcal{I}} \{\phi_i(t-1) \cdot a_{i,j}\} \cdot b_j(c_t) \qquad (5)$$

Donde:

- j: Es el índice del estado dentro del modelo λ, tal que $j \in \mathcal{I}$, con $\mathcal{I} = \{i\}_{i=1}^{M}$.
- t: Es el instante de tiempo de la iteración.
- $\phi_j(t)$: es la variable de estado que acumula la verosimilitud del mejor camino que llega al estado j en el instante t.
- $a_{i,j}$: Es la probabilidad de la transición del estado i al j.
- $b_j()$: Es la función de densidad de probabilidad para el estado j. Compuesta en este trabajo por una mezcla de gaussianas.
- c_t: Es el vector acústico en el instante t, utilizando como hemos visto en la sección anterior, parámetros cepstrales y siguiendo la notación anterior, c_t sería la columna número t de la matriz $C^{\alpha}(c \times L)$, con $\alpha = 1{,}0$.

La ecuación de Viterbi modificada es:

$$\phi_j^{\alpha_n}(t) = \max_{i \in \mathcal{I}, \alpha_m \in \mathcal{A}} \phi_i^{\alpha_m}(t-1) \cdot a_{i,j}^{m,n} \cdot b_j(c_t^{\alpha_n}) \qquad (6)$$

Donde:

- (j, n): Son los índices del estado dentro del modelo extendido λ', tal que $j \in \mathcal{I}$, y $\alpha_n \in \mathcal{A}$.
- $\phi_j^{\alpha_n}$ (): Es la nueva variable de estado de acumulación de verosimilitud.
- $a_{i,j}^{m,n}$: Es la probabilidad de la transición del estado (i, m) al (j, n).
- b_j (): Es la función de densidad de probabilidad para el estado j, como destacaremos más adelante se ha hecho independiente del factor de normalización, de manera que el número de parámetros del modelo λ' no sean mucho mayor que el de λ.
- $c_t^{\alpha_n}$: Es el vector acústico número t en la secuencia de tramas parametrizadas con la transformación α_n , es decir la columna número t de la matriz C^{α_n}.

Normalmente con la finalidad de acotar el espacio de búsqueda y de hacer las transiciones entre factores de transformación más lentas hacemos que:

$$a_{i,j}^{m,n} = 0, \text{ si } |m - n| > 1$$

A continuación, en la figura 6 se muestra un ejemplo del resultado de la búsqueda en el espacio de estados aumentado, se trata de la realización de la palabra inglesa "two". En primer lugar se ha hecho una búsqueda con el algoritmo de Viterbi modificado 5, obteniéndose la mejor secuencia de estados en el modelo aumentado recorriendo hacia atrás (*backtraking*) la secuencia de estados. Entonces, para cada estado de la secuencia óptima se han representado todas las verosimilitudes (en el eje vertical) de los estados que tienen en común el origen en un mismo estado del modelo original λ, es decir para un estado (j, n) en la secuencia óptima, se representan en el eje vertical las verosimilitudes de todos los estados de la forma $(j, m), \forall m = 1, \ldots, N$. De esta forma podemos ver el camino que ha seguido el algoritmo de Viterbi dentro del espacio de búsqueda para el grado de libertad añadido de la transformación de frecuencia.

Con este simple ejemplo de una palabra aislada se puede ver una de las ventajas en cuanto a modelado del método, ya que, como podemos ver, en la parte inicial de la palabra correspondiente a la explosión de la "t", que es un sonido de alta frecuencia y no vocálico, el algoritmo a elige valores cercanos a no hacer transformación o un factor $\alpha \simeq 1{,}0$. Sin embargo, en la parte vocálica de la palabra, la probabilidad de los factores de transformación más altos crece y se sitúa en torno a la estimación que se hizo anteriormente con el método original de transformación por frases, que daba una trasnformación óptima para toda la frase de $\alpha = 1{,}13$.

El método descrito en [7], se basaba en la misma idea de la ampliación del espacio de búsqueda por medio del factor de transformación. Pero en su modelo, la matriz de transformación no es como la definida anteriormente $a_{i,j}^{m,n}$, si no que se ha hecho una suposición de independencia entre las transiciones entre estados dentro de la misma transformación y entre transformaciones. La suposición de independencia es la siguiente:

$$a_{i,j}^{m,n} = a_{i,j} \cdot b_{m,n}$$

Donde $a_{i,j}$ es la probabilidad de transición original y $b_{m,n}$ es la probabilidad de transición entre factores de transformación que sólo toma los valores 0 ó 1. Esta manera de modelar las transiciones se podría pensar menos controlada, ya que al aprender la probabilidad conjunta, aprendemos en cierta medida si un estado tiene tendencia a tener algún tipo de factores de transformación en el conjunto de entrenamiento, como sería deseable para no aumentar la confusión entre las consonantes por ejemplo.

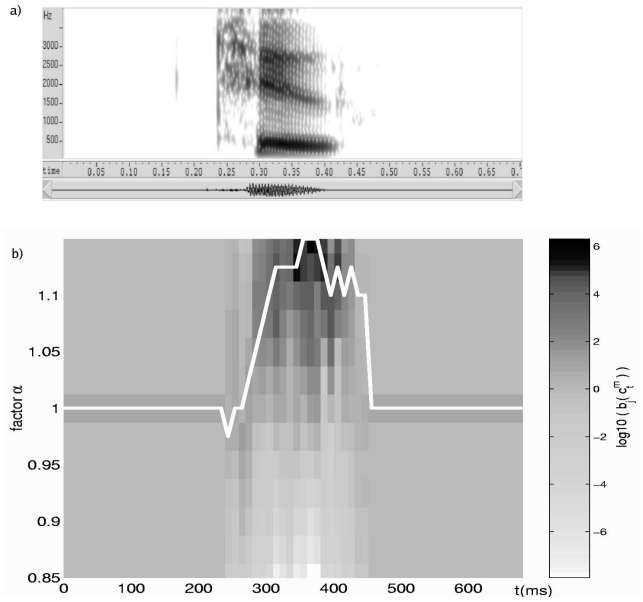

Figura 6: a) Espectrograma de una realización de la palabra two. b) Probabilidad de observación de cada factor de transformación para los estados del mejor camino encontrado por el algoritmo de Viterbi

Otro factor de diferencia es que en los modelos propuestos en el citado trabajo, no hay ninguna referencia explícita a un tratamiento diferenciado para el modelo de silencio, así que es de suponer que se permite, como un estado más, que se asignen tramas transformadas al modelo de silencio, lo cual, como hemos visto, puede ser una fuente de error para el sistema.

4. Resultados

	exactidud $\frac{Sus.+Borr.+Ins.}{NWords}$	% Mejora
MFCC-D-A	98.74	-
$g^{\alpha}(f)$	98.85	+8.7 %
$g^{\alpha}(f,t)$	98.89	+11.9 %

Tabla 1: Tabla de resultados para la base de datos Aurora2

Una vez vista la descripción del método y algún ejemplo de su aplicación, procedemos a explicar los experimentos realizados para comprobar su capacidad de modelado frente a la parametrización original sin transformar y a la transformación por frases de métodos anteriores. Los experimentos se han desarrollado con una base de datos de dígitos conectados, Aurora 2, de la que se ha utilizado el entrenamiento y test de señales limpias. En total 8440 frases de entrenamiento (27727 dígitos en inglés) y de test, 4004 frases (en total 13159 dígitos en inglés).

Los modelos utilizados fueron modelos de palabra con 16 estados y 3 gaussianas por estado para los dígitos, 3 estados y 6 gaussianas por estado para el modelo de silencio incial y final y 1 estado con 6 gaussianas para el modelo de silencio entre

palabras. Los parámetros utilizados fueron los descritos en el proceso de parametrización en la sección 2.1 con un tamaño de ventana de 25 ms y un deslizamiento de 10 ms.

En la tabla 4, se puede ver la media de la exactitud de reconocimiento en varios casos:

- MFCC-D-A: En el caso del modelo original y la parametrización de referencia utilizada (los parámetros MFFC, *Mel Filter Bank Cepstrum*, con las derivadas y aceleraciones)
- $g^{\alpha}(f)$: Para la transformación lineal de máxima verosimilitud, con el sistema de tres pasos de procesado de la frase.
- $g^{\alpha}(f,t)$: Para la transformación dependiente del tiempo, estimada en tiempo real.

De esta manera comprobamos que el modelo dependiente del tiempo consigue unos resultados comparables, ligeramente superiores, sin necesidad de una hipótesis previa o varios procesados de la frase, operando en tiempo real y menos costo computacional, (en los experimentos, aproximadamente un 30 % menos de cálculo).

Estos resultados permiten pensar que hay más trabajo que realizar en el campo de la normalización del tracto vocal, en especial en su aplicación especializada trama a trama o para modelos diferentes, ya que se ha intentado mostrar que las transformaciones genéricas mejoran pero quedan lejos de modelar por completo la compleja dinámica de la producción de voz. Por otra parte, con modelos que tengan en cuenta los detalles, como el presente, hay que tener cuidado en construir modelos con excesiva libertad, ya que esto podría causar más errores que evitarlos.

5. Discusión

La técnica presentada aporta un tipo de modelado muy interesante para conseguir la independiente del locutor, gracias al grado de flexibilidad que permite normalizaciones del tracto vocal variables en el tiempo, sin necesidad de aumentar el número de gaussianas.

Otra de las posibles aplicaciones en la que sería interesante estudiar la aplicación de este sistema, es en el modelado del efecto Lombard [9, 10], como se ha estudiado anteriormente este tipo de situaciones en las que el locutor está sujeto a algún tipo de estrés modifican la forma en la que se pronuncia. En [11] se estudió la relación la la normalización del tracto vocal sugiriendo algo muy interensante, y es que los efectos del estrés sobre el tracto vocal no son uniformes a lo largo de una frase, con lo que el sistema propuesto debería modelar bien esta situación.

Para finalizar, queda por destacar otro tema de desarrollo futuro en relación a esta técnica, que consistiría en el estudio de la robustez del método frente a ruido, y de su mejora, mediante la combinación con algunas de las técnicas existentes como sustracción espectral, normalizado y compensación de parámetros o técnicas de análisis multibanda y de *missing-data*, (reconocimiento con descarte de parámetros acústicos en función del ruido).

6. Conclusiones

Se ha presentado una técnica de modelado que permite la estimación y la decodificación simultánea del factor de transformación para la normalización del tracto vocal y de la secuencia de estados óptima por medio de un modelo con un grado de libertad añadido en la búsqueda de Viterbi. Esto permite la operación en tiempo real del sistema sin un excesivo costo computacional. Anteriormente ha aparecido otro método con la misma filosofía aunque en el presente se incorporan algunas mejoras como el tratamiento especializado del modelo de silencio y la matriz de transiciones conjunta para todos los grados de libertad, así como el proceso de entrenamiento asociado. Además se el método propuesto trata de solucionar el problema de estimación del factor de transformación debido a la transformación de las tramas asociadas al modelo de silencio. Con todo esto se consiguen unos resultados comparables a los métodos de varios procesados e hipótesis de transcripción pero operando en tiempo real.

7. Referencias

[1] H. Wakita, "Normalization of vowels by vocal tract length and its application to vowel identification," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, April 1977, pp. 25:183–192.

[2] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[3] A. Acero and R. M.Stern, "Robust speech recognition by normalization of the acoustic space," in *Proc. of EUROSPEECH'91.*, 1991.

[4] T. Kamm A. Andreou and J. Cohen, "A parametric approach to vocal tract length normalization," in *In Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[5] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *In Proceedings of ICASSP'96*, Atlanta, USA, 1996.

[6] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions SAP*, vol. 1, no. 6, pp. 49–60, 1998.

[7] T. Fukada and Y. Sagisaka, "Speaker normalized acoustic modeling based on 3-d viterbi decoding," in *In Proceedings of ICASSP'98*, Seattle, Washington, USA, 1998, vol. 1, pp. 437–440.

[8] M. Pitz, S. Molau, R. Schluter, and H.Ñey, "Vocal tract normalization equals linear transformation in cepstral space," in *Proc. of the EUROSPEECH'01*, Aalborg, Denmark, 2001.

[9] Y. Chen, "Cepstral domain talker stress compensation for robust speech recognition," *IEEE Transactions on Signal Processing*, vol. 36, pp. 433–439, Apr. 1988.

[10] J.-C. Junqua, "The influence of psychoacoustic and psycholinguistic factors on listener judgments of intelligibility of normal and lombard speech," in *In Proceedings of ICASSP'91*, Toronto, Canada, 1991, vol. 1, pp. 361–364.

[11] J. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (mce-acc) for speech recognition in noise and lombard effect," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 598–614, Oct. 1994.