LA FRECUENCIA FUNDAMENTAL DE LA VOZ Y SUS EFECTOS EN RECONOCIMIENTO DE HABLA CONTINUA

C. García, D. Tapias
División de Tecnología del Habla
Telefónica Investigación y Desarrollo, S.A. Unipersonal
C/ Emilio Vargas, 6
28043 - Madrid (España)

RESUMEN

El objetivo del estudio que aquí se presenta es el de analizar el efecto de la variación de la frecuencia fundamental en las características de la señal de voz, estudiar los efectos de este parámetro sobre el comportamiento de los sistemas de reconocimiento de habla continua y evaluar técnicas de compensación de los efectos observados. Para llevarlo a cabo se han grabado varias bases de datos específicas que también se describen. Las pruebas realizadas indican que la tasa de acierto del sistema se ve afectada por el valor medio de la frecuencia fundamental, llegando a experimentar una reducción relativa del 38.8% para algunos valores extremos de este parámetro. La aplicación de la técnica MLLR permite compensar en parte esta degradación, experimentándose reducciones de la tasa de error de hasta el 51.9% para los locutores más problemáticos con sólo 10 frases de adaptación.

1. INTRODUCCIÓN

Los sistemas de reconocimiento de habla continua actuales, presentan una tasa de acierto elevada cuando las condiciones de evaluación son similares a las de entrenamiento. Sin embargo cuando estas difieren entre sí, la tasa de acierto se reduce drásticamente. Hay muchos factores que afectan a las condiciones de evaluación, como el ruido de fondo, el canal y la variabilidad inter/intralocutor. En este artículo nos centramos en la variabilidad inter/intralocutor, y en particular en el efecto de la variación de la frecuencia fundamental en los reconocedores de habla continua.

La frecuencia fundamental (F0) es uno de los parámetros que caracterizan la voz de un locutor. A diferencia de lo que sucede con otros parámetros, los efectos de la frecuencia fundamental de la voz sobre el reconocimiento del habla apenas han sido estudiados hasta el momento. Tan solo existe algún estudio previo, como el descrito en [1] y [2], que analiza los efectos de la frecuencia fundamental, entre otros parámetros, sobre un sistema de reconocimiento de palabras aisladas dependiente del locutor. Los efectos sobre sistemas de reconocimiento de habla continua en general, y sobre sistemas independientes del locutor en particular apenas han sido, pues, objeto de estudio. Este trabajo pretende contribuir a un mejor conocimiento de este parámetro de la voz y de estos efectos, así como a evaluar técnicas de compensación del mismo.

El artículo está organizado de la siguiente forma: En el apartado 2 se describen las bases de datos utilizadas en el estudio y en los experimentos. El apartado 3 recoge los resultados del estudio del grado de variabilidad de la frecuencia fundamental. En el apartado 4 se presenta el análisis de las características del espectro en función del valor de frecuencia fundamental. El apartado 5 se dedica al estudio de los efectos de la frecuencia fundamental sobre el reconocimiento de habla continua, y el apartado 6 presenta los resultados obtenidos al aplicar la técnica MLLR para compensar estos efectos. Por último, el apartado 7 recoge las conclusiones y las futuras líneas de investigación que quedan abiertas tras este estudio.

2. BASES DE DATOS

2.1 VOCASTEL

Se trata de una base de datos de palabras aisladas, cadenas de números y frases fonéticamente balanceadas en castellano, que contiene grabaciones a 8 KHz realizadas en 12.493 llamadas telefónicas procedentes de todas las provincias españolas.

Esta base de datos ha sido empleada para estudiar el grado de variabilidad de la frecuencia fundamental. El número de locutores empleados en el estudio asciende a 4.509 (2.193 masculinos y 2.316 femeninos), lo que se traduce en 104.855 ficheros (25.171 palabras aisladas, 42.603 cadenas de números y 37.081 frases enunciativas). Esto es, aproximadamente 63 horas de voz (29 horas de voz masculina y 34 de femenina).

2.2 Base de Datos de Pitch

Esta base de datos tiene el objetivo de proporcionar una cantidad suficiente de datos en todo el rango de variación de la frecuencia fundamental media, incluso en los valores más extremos. Asimismo, tiene el fin de aislar, en la medida de lo posible, los efectos de la frecuencia fundamental de la voz de otros efectos, como la longitud del tracto vocal de los locutores, su acento o dialecto, la velocidad del habla o el ruido de fondo. Para ello se ha seguido un cuidadoso proceso de selección de los locutores y de las condiciones de grabación [3].

La base de datos contiene un total de 2.490 frases del dominio de ATIS [4] en castellano, divididas en 6 rangos de frecuencia fundamental media, cada uno de los cuales contiene un número mínimo de 200 frases. La base de datos contiene grabaciones de 56 locutores adultos (37 masculinos y 19 femeninos) y cada rango de frecuencia fundamental media contiene voz de al menos 5 locutores.

2.3 Base de Datos de Pitch Artificial

El objetivo de esta base de datos es la de contrastar los resultados obtenidos con la Base de Datos de Pitch descrita en el apartado 2.2. Esta base de datos está compuesta por voz generada mediante un conversor texto-voz, cuyas características se describen en [5].

Está dividida en 10 grupos de frases, cada uno de ellos con un valor diferente de frecuencia fundamental media. El número de frases de cada uno de estos grupos es de 2.270, pertenecientes al dominio de ATIS en castellano.

Las muestras de voz están generadas a una frecuencia de muestreo es de 8 KHz.

3. VARIABILIDAD DE F0

El estudio de la variabilidad de la frecuencia fundamental se ha llevado a cabo sobre la base de datos VOCASTEL.

En cada uno de los ficheros de voz de estos locutores se ha realizado una medida de frecuencia fundamental cada 10 ms. Posteriormente, se ha calculado la frecuencia fundamental media, así como el margen dinámico y la desviación típica de F0 de cada fichero. Asimismo, se ha obtenido, para todas las frases enunciativas, la pendiente resultante de ajustar los valores de F0 obtenidos por el método de mínimos cuadrados, con el fin de analizar la tendencia de estos valores a lo largo de la frase.

Los resultados obtenidos revelan la existencia de importantes diferencias inter e intralocutor de todos los parámetros analizados. Cabe destacar también las diferencias existentes en función del sexo de los locutores. Así, el margen dinámico de F0 es, en media, superior para las mujeres que para los hombres (189.8 Hz de media frente a 152.3 Hz), y lo mismo sucede con la desviación típica (44.2 Hz frente a 32.3 Hz) y con la pendiente con que tiende a decaer a lo largo de la frase (32.7 Hz/s frente a 10.1 Hz/s).

En cuanto a la frecuencia fundamental media, también es superior, para los locutores femeninos (180 Hz frente a 125 Hz). La función densidad de probabilidad de este parámetro se muestra en la figura 1.

Se puede apreciar que la función presenta dos máximos situados en los 112 Hz y los 180 Hz, dado que la mayoría de los hombres tienen un valor de F0 medio cercano a los 110 Hz y la mayor parte de las mujeres lo tienen comprendido entre los 175 y los 200 Hz, aproximadamente. Sin embargo, como se puede observar, la frecuencia fundamental media para un locutor dado, puede adquirir cualquier valor dentro de un rango de variación que va, aproximadamente, de los 75 a los 300 Hz, aunque se han observado algunos casos aislados con frecuencias medias superiores.

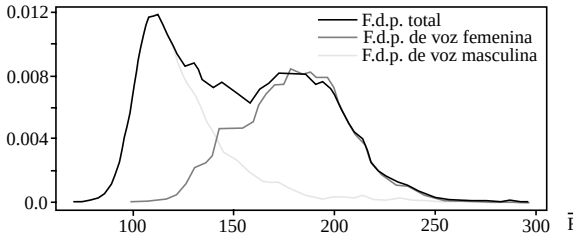

Figura 1

Por lo tanto, todo sistema de reconocimiento robusto frente a la frecuencia fundamental de la voz, debe ser capaz de hacer frente a voz con

valores de frecuencia fundamental media comprendidos en todo este rango de variación.

4. EFECTOS DE F0 SOBRE EL ESPECTRO

Para analizar qué efectos tiene el valor de frecuencia fundamental sobre las características del espectro, se ha comparado el espectro de un mismo fonema dentro de una misma frase pronunciada por locutores con distintos valores de frecuencia fundamental media. El proceso se ha repetido para los distintos tipos de fonemas, observándose varios ejemplos de cada uno de ellos.

Las frases utilizadas proceden de la base de datos VOCASTEL. También se han generado frases con el conversor texto-voz utilizado para grabar la Base de Datos de Pitch Artificial para generar frases que difieran tan sólo en su valor de F0 medio.

El estudio indica que las características del espectro de las vocales y las consonantes sonoras son dependientes del valor medio de frecuencia fundamental de la frase en la que han sido pronunciados. En concreto, se aprecia cómo la mayor o menor separación de los armónicos de la frecuencia fundamental provoca que varíe mucho la amplitud pico-valle de los lóbulos espectrales en función del valor de F0.

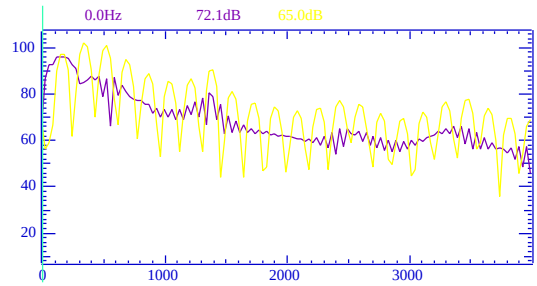

Figura 2

La figura 2 ilustra lo descrito anteriormente. En ella se muestran los espectros de la vocal ‘a’ en dos frases con frecuencia fundamental media 90 y 250 Hz (trazo más oscuro y más claro respectivamente). Las frases han sido generadas por el conversor texto-voz, de manera que la única diferencia entre ellas reside en el valor de frecuencia fundamental media. Como se puede apreciar, las características de ambos espectros son diferentes. Por tanto, es razonable pensar que este hecho afectará a los vectores de características empleados en reconocimiento de voz.

El grado en que estas diferencias espectrales afectan al comportamiento de los sistemas de reconocimiento de habla continua se discute en el siguiente apartado.

5. EFECTOS DE F0 EN RECONOCIMIENTO

Para analizar estos efectos se ha estudiado si la tasa de acierto de un sistema de reconocimiento se ve afectada por el valor de frecuencia fundamental media de la voz que ha de reconocer o si, por el contrario, esta tasa apenas varía con el valor de F0 medio.

5.1 Sistema de Reconocimiento

El sistema utilizado es un reconocedor de habla continua independiente del locutor basado en modelos ocultos de Markov continuos. Emplea unidades dependientes del contexto (trifonemas) y un modelo estadístico del lenguaje basado en trigramas. El vocabulario que maneja es de unas 5.000 palabras. La tasa media de error de palabra del sistema, obtenida al reconocer 2270 frases del dominio de ATIS en castellano, es del 8%.

5.2 Resultados Experimentales con la Base de Datos de Pitch

Los resultados experimentales se resumen en la figura 3, que muestra el valor de la tasa de acierto en función de la frecuencia fundamental media.

Según se puede apreciar, la tasa de acierto supera el valor promedio (92%) para los valores de frecuencia fundamental media cercanos a los 110 Hz y los comprendidos entre 175 y 200 Hz. Esto se debe a que estas frecuencias son las más frecuentes en la base de datos de entrenamiento y, por ello, son las mejor representadas por los modelos acústicos.

Por otro lado, el valor de la tasa de acierto para los valores menos comunes en los datos de entrenamiento es inferior a la tasa media global. Este es el caso de las frecuencias fundamentales medias comprendidas entre los 130 Hz y 175 Hz y, sobre todo, el de las inferiores a los 100 Hz y las superiores a los 220 Hz. Para algunos de estos valores de F0 medio, fundamentalmente los que superan los 220 Hz, la tasa de acierto es especialmente baja. Así, la tasa de acierto para las frases con una frecuencia fundamental media mayor de 260 Hz es de tan solo un 56.3%, lo que supone una reducción relativa del 38.8%

con respecto a la tasa global del sistema.

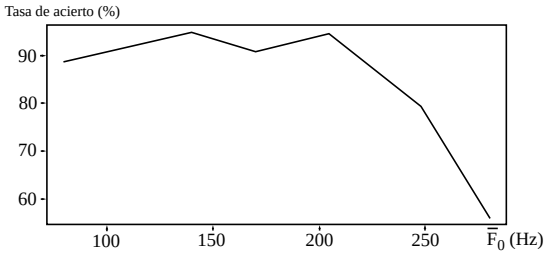

Figura 3

Un análisis más detallado de los resultados indica que la mayor parte de los errores de reconocimiento son errores de sustitución, lo que sugiere que estos errores se deben, principalmente, a un inadecuado modelado acústico. En el caso de los valores de F0 medio más problemáticos, la contribución de las sustituciones al error total es incluso más importante que para los demás casos, lo que induce a pensar que los modelos acústicos son especialmente inadecuados para representar la voz con estos valores de frecuencia fundamental media.

5.3 Base de Datos de Pitch Artificial

Los resultados obtenidos con la Base de Datos de Pitch han sido contrastados utilizando la Base de Datos de Pitch Artificial. Esta base de datos tiene la limitación de estar grabada con voz generada artificialmente. Sin embargo tiene las ventajas de disponer de un gran número de frases por cada valor de frecuencia fundamental media y de que las frases generadas para cada uno de los 10 grupos, se diferencian por su valor de F0 medio, ya que los otros parámetros permanecen constantes (tracto vocal del locutor, velocidad del habla, volumen, ruido y modelo de producción de la voz).

La figura 4 resume los resultados obtenidos. En esta figura aparece la tasa de acierto del sistema en función del valor de F0 medio para las frases de esta base de datos.

Puede comprobarse que, en este caso, la tasa de acierto para los valores de frecuencia fundamental media más altos y más bajos no se degrada tanto como en el caso de la voz natural. Sin embargo, puede apreciarse cómo esta tasa es claramente inferior a la obtenida para los valores de F0 medio mejor representados en la base de datos de entrenamiento.

Por lo tanto, las pruebas realizadas con esta base de datos confirman el hecho de que la frecuencia fundamental de la voz es un parámetro que afecta al comportamiento del sistema de reconocimiento, haciendo que la tasa de acierto del mismo se degrade para los valores de F0 medio menos frecuentes en los datos de entrenamiento.

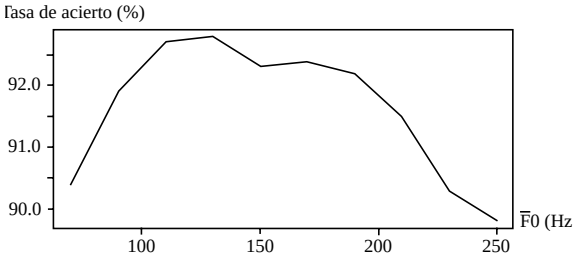

Figura 4

6. COMPENSACIÓN DE LOS EFECTOS DE F0

Los experimentos de compensación se han realizado con el algoritmo Maximum Likelihood Linear Regression (MLLR) [7], al que se le ha impuesto la restricción de que se precise una cantidad reducida de datos de adaptación para conseguir las mejoras. El objeto de este planteamiento es que el sistema resultante pueda emplearse en algunos tipos de servicios telefónicos en los que si bien no es posible solicitar que el usuario repita una serie de frases para reentrenar el sistema, se puede disponer de datos de forma incremental cada vez que el usuario accede al servicio.

Los experimentos han evaluado el funcionamiento de este algoritmo con dos enfoques: generación de modelos independientes del locutor pero dependientes del F0 medio y generación de modelos dependientes del locutor.

6.1 Modelos independientes del locutor y dependientes del F0 medio.

La ventaja fundamental de este enfoque es que una vez generados los modelos para cada grupo de frecuencias, la adaptación se realiza de forma instantánea, dado que una vez conocida la frecuencia fundamental media del nuevo usuario, sólo hay que reconocer aplicando los modelos adaptados a la misma. Por otro lado, el principal problema es la dificultad de encontrar un número significativo de locutores para valores extremos de frecuencia fundamental media. En el caso que nos ocupa, sólo se pudieron emplear 3 locutores para adaptar los modelos, utilizándose las frases de los restantes locutores para evaluación.

Los resultados experimentales obtenidos mostraron que el número de locutores empleado para realizar la adaptación es insuficiente, dado que los modelos se adaptaban demasiado a sus características particulares y, por tanto, las tasas de error se incrementan para el resto de los locutores que no han sido empleados para adaptar el sistema pero que están en el mismo grupo de frecuencia fundamental..

Así pues, la aplicación de esta técnica no resulta satisfactoria para compensar los efectos de la variabilidad de la frecuencia fundamental en estas condiciones, aunque en un futuro se intentará disponer de más datos de adaptación para probar de nuevo este enfoque.

6.2 Modelos dependientes del locutor

En este caso se han adaptado los modelos acústicos al locutor empleando el menor número posible de frases de adaptación, con el fin de no retardar demasiado el acceso a servicios de información telefónica basados en el sistema de reconocimiento.

En este caso los resultados obtenidos han sido bastante satisfactorios, comprobándose que se puede mejorar sustancialmente la tasa de acierto para los locutores más problemáticos empleando tan solo 5 ó 10 frases de adaptación.

TABLA 1

Locutor	F0 medio	Sin adapt.	Adaptados
mf	93 Hz	93.8	92.3
mb	97 Hz	93.3	91.0
me	152 Hz	92.5	91.5
fa	251 Hz	84.6	85.1
fs	259 Hz	65.5	91.8
fx	278 Hz	52.8	71.0

La tabla 1 muestra la tasa de acierto para 6 locutores de la Base de Datos de Pitch, utilizando los modelos iniciales (sin adaptar) y los adaptados mediante MLLR utilizando 10 frases de adaptación y 3 matrices de transformación.

Según se puede observar, la tasa de acierto para los locutores menos problemáticos tiende a empeorar ligeramente cuando se emplean los modelos adaptados. Sin embargo, la degradación de la tasa es pequeña, por lo que el valor de la misma sigue siendo bastante alto en el caso de utilizar los modelos adaptados. Por otra parte, esta degradación podría evitarse sin demasiada dificultad estimando la frecuencia fundamental media de la voz de entrada al reconocedor, para adaptar los modelos sólo para aquellos valores de F0 medio que resulten problemáticos.

Por contra, la tasa de acierto para los locutores que plantean una mayor dificultad experimenta notables mejoras cuando se emplean los modelos adaptados al locutor. En el caso del locutor *fs* se obtiene una mejora relativa del 40.2%.

Las mejoras obtenidas cuando se emplean 5 frases de adaptación son algo menores, aunque podrían resultar suficientes para comenzar a prestar ciertos servicios. Cuando se utilizan 20 frases de adaptación, la tasa de acierto aumenta algo más, llegándose a obtener para algún locutor (*fx*) una mejora relativa del 46.4%.

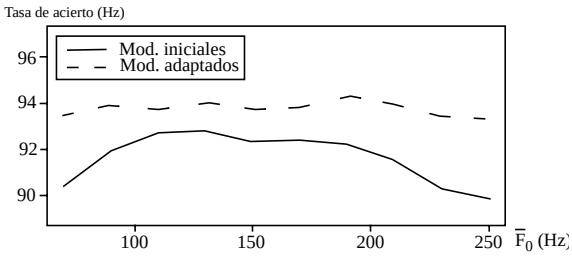

Figura 5

La robustez de la técnica frente al valor de frecuencia fundamental se ha probado con ayuda de la Base de Datos de Pitch Artificial. La figura 5 presenta el valor de la tasa de acierto del sistema utilizando los modelos iniciales y los modelos adaptados utilizando 10 frases de adaptación. Puede comprobarse cómo las mayores mejoras se consiguen precisamente para los valores de F0 medio más problemáticos, por lo que la aplicación de la técnica hace que las tasas de acierto para todos los valores de frecuencia fundamental media tiendan a igualarse.

Así pues, la adaptación de los modelos acústicos al locutor mediante MLLR permite, con tan solo 5 ó 10 frases de adaptación, compensar los efectos de la variabilidad de F0 sobre el sistema de reconocimiento, al mejorar la tasa de acierto para aquellos locutores que resultan más problemáticos debido al valor de frecuencia fundamental media de su voz.

7. CONCLUSIONES

En este artículo se ha demostrado que la frecuencia fundamental es un parámetro que presenta una gran variabilidad inter e intralocutor. En concreto, la frecuencia fundamental media

de un locutor puede adquirir cualquier valor en un rango de variación que va aproximadamente de los 75 Hz a los 300 Hz. Esto se traduce en diferencias en los vectores de características empleados por los sistemas de reconocimiento y, por tanto, en aumentos de la tasa de error para aquellas frecuencias que no están bien representadas en la base de datos de entrenamiento.

Por último, se concluye que la adaptación al locutor de los modelos acústicos del sistema de reconocimiento mediante MLLR permite, con pocos datos de adaptación, compensar en gran medida los efectos de la gran variabilidad de la frecuencia fundamental de la voz, aunque no permite la generación de modelos dependientes de la frecuencia e independientes del locutor, debido a la escasez de datos de adaptación.

REFERENCIAS

[1] Thomas T.J., Peckham J., Frangoulis E. *A Determination of the Sensitivity of Speech Recognisers to Speaker Variability.* Proceedings of ICASSP, Glasgow (1989), pp. 544-547.

[2] Thomas T.J., Peckham J., Frangoulis E., Cove J. *The Sensitivity of Speech Recognisers to Speaker Variability and Speaker Variation.* Proc. of Eurospeech, París (1989), pp. 408-411.

[3] García C. *Estudio de la Frecuencia Fundamental de la Voz y de sus Efectos en el Reconocimiento de Habla Continua*. Proyecto Fin de Carrera en E.T.S.I. Telecomunicación (U. Politécnica de Madrid), (2000).

[4] Ward W. *The CMU Air Travel Information Service: Understanding Spontaneous Speech*. Proceedings of the DARPA Speech and Natural Language Workshop, Pennsylvania (1990), pp.845-848.

[5] Rodriguez M. A., Escalada J. G., Macarrón A., Monzón L. *AMIGO: Un Conversor Texto-Voz para el Español*. Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN'92, Vol. 13, (1992), pp. 389-400.

[6] Talkin D. *A Robust Algorithm for Pitch Tracking (RAPT)*. En *Speech Coding and Synthesis* (ed. por Kleijn W.B., Paliwal K.K.). Elsevier Science, (1995), pp. 495-518.

[7] Leggeter C.J., Woodland P.C. *Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models*. Computer Speech and Language, vol. 9, (1995),pp.171-185.