¿Podemos imitar la voz de una persona? Técnicas de conversión de hablante

Juana Mª Gutiérrez Arriola, Juan Manuel Montero Martínez**, Ricardo de Córdoba Herralde**, Jose Manuel Pardo Muñoz***

**Departamento de Ingeniería de Circuitos y Sistemas*
***Grupo de Tecnología del Habla. Departamento de Ingeniería Electrónica*
Universidad Politécnica de Madrid
e-mail: jmga@ics.upm.es

RESUMEN

En este artículo presentamos dos técnicas de conversión de hablante y su aplicación a la síntesis multilocutor. La primera de las técnicas se basa en la transformación lineal de los formantes y los anchos de banda. La segunda trata de convertir pronunciaciones codificadas mediante LPC utilizando correspondencia de codebooks. En ambos casos se transforma la curva de tono y se aplican los parámetros de fuente glotal del locutor deseado. Los resultados obtenidos son satisfactorios en ambos casos de forma que la voz transformada se identifica claramente como la del locutor deseado, el problema de estas técnicas es que se necesitan muchos datos (tantos como son necesarios para generar una nueva voz en el sintetizador) para que la transformación sea aceptable.

ABSTRACT

We will present two speaker-conversion techniques and their application to a multi-speaker synthesizer. The first technique linearly transforms formants and bandwidths. The second one converts LP coded speech using codebook matching. In both cases F0 curve is transformed and the target speaker glottal pulse parameters are applied. Results are good for both synthesizer and the transformed speech is easily identified as being uttered by the target speaker. The main problem is that a lot of data is required (the same amount as to generate a new speaker from scratch).

1. INTRODUCCIÓN

La forma de transmitir la información cuando nos comunicamos con otro ser humano no es única ni uniforme. Además del significado de las palabras o frases que pronunciamos existe una serie de factores adicionales que son igual de importantes, como son, la voz de la persona que habla, su timbre, la entonación, los gestos que hace cuando transmite el mensaje, las palabras que escoge para hacerlo, el acento,.... Y así podríamos seguir enumerando factores. Gracias a todas esas características podemos percibir si la persona que nos está hablando está triste, alegre, enfadada, de dónde es,... Y si no estamos viendo a la persona que habla podemos identificarla por su voz. La identidad del locutor es importante no sólo porque nos ayuda a identificar a la persona con la que estamos hablando, sino porque enriquece nuestra vida diaria con variedad.

La conversión de hablante, llamada en algunos textos conversión de voz o conversión de locutor, es una técnica usada para cambiar o modificar la identidad del hablante; es decir, lo pronunciado por un locutor es transformado para que suene como si lo hubiera articulado otro.

Las aplicaciones potenciales de esta técnica son numerosas. En los sistemas de conversión texto-voz esta técnica puede añadir variedad además de permitir la adaptación de la

voz en prótesis para personas con problemas de fonación. En los sistemas basados en selección y concatenación de unidades acústicas se reconoce que la calidad del sintetizador depende, entre otros factores, de la cantidad de datos de que se dispone. Parece lógico pensar que cuántos más datos tengamos más posibilidades tenemos de encontrar la unidad que nos hace falta con el mínimo de modificación, por lo tanto se utiliza gran cantidad de datos. Estos datos son difíciles de conseguir y de manejar (segmentación, almacenamiento, acceso en tiempo real), desarrollar una nueva voz puede ser un proceso extremadamente costoso. La conversión de voz podría ser una alternativa.

Los sistemas de traducción automática también se beneficiarán de esta técnica. Pensemos, por ejemplo, en una comunicación telefónica entre dos personas que hablan distintos idiomas y utilizan un sistema de traducción automática, sería muy importante que la voz traducida mantuviera las características de la persona que ha dicho la frase.

En los sistemas telefónicos de información, donde cada vez parece más clara la tendencia a sustituir al operador humano por un sistema automático basado en tecnologías de habla, también parece importante dotar al sistema de variedad en dos aspectos, primero en cuanto a la identidad del operador y segundo en cuanto al estilo del habla producida por el operador cuando se entablan conversaciones pregunta-respuesta.

Otra posible aplicación es el diseño de ayudas auditivas para problemas auditivos específicos.

1.1 Características acústicas de la identidad del locutor

Los factores que son relevantes para caracterizar la identidad del locutor se pueden categorizar en términos socio-psicológicos y en términos fisiológicos [30], [3]. Las propiedades anatómicas de los órganos de fonación son las responsables de la frecuencia fundamental y de las características espectrales de la voz. Mientras que nuestro entorno, educación o estado de ánimo caracterizan el acento, la curva entonativa, la duración de las palabras, ritmo, pausado, niveles de potencia,...

Las investigaciones sobre la identidad del locutor tienen una historia relativamente larga, que abarca desde los primeros estudios psicológicos y fonéticos que demuestran las relaciones entre los parámetros acústicos y las características del locutor como son: edad, sexo, altura, peso y otras propiedades físicas, hasta los estudios más recientes que intentan aclarar qué parámetros acústicos son relevantes en lo que a la identidad del locutor se refiere [5],[7],[8],[12],[16], [21],[32],[40],[41],[42],[45]. Matsumoto et al investigaron la contribución del tono (F0), la frecuencia de los formantes, la envolvente espectral y otros parámetros acústicos para varios locutores masculinos pronunciando vocales [33]. Concluyeron que, en cuanto a la identidad del locutor se refiere, el parámetro más importante era el tono o F0 seguido de la frecuencia de los formantes, la fluctuación de F0 y la inclinación espectral de la fuente glotal. Karlsson estudió las variaciones presentes en la voz femenina y como se pueden cambiar las características de la voz sintetizada variando la fuente glotal [26],[27],[28],[29]. Furui estudió la relación entre distancias psicológicas y físicas entre locutores [15], e informó que la correlación más alta se daba con el espectro medio a largo plazo, seguido del tono medio. En particular, el rango de frecuencias de 2.5-3.5KHz parecía ser el que más contribuía a definir la identidad del locutor.

Más recientemente, Hanson estudia las características glotales de hombres y mujeres en [19] y [20]. Bachorowski en [5] concluye que los factores determinantes a la hora de distinguir el género del locutor son la frecuencia fundamental y la longitud del tracto vocal, calculada a partir de las frecuencias de los formantes. Sin embargo, dentro de un mismo género los parámetros más significativos para distinguir locutores son los relacionados con el tracto vocal.

Teniendo en cuenta estos resultados y los obtenidos en [6],[31],[34] podemos concluir que los parámetros más influyentes en la identidad del locutor son:

Fuente vocal:

- El tono medio
- El patrón tiempo-frecuencia del tono (curva de tono).
- La fluctuación de la frecuencia fundamental.
- La fluctuación de la energía de la fuente glotal.
- La forma de onda glotal.

Tracto vocal:

- La longitud del tracto vocal.
- La forma de la envolvente espectral y la inclinación espectral.
- Las diferencias relativas en la amplitud de los formantes.
- Los valores absolutos de las frecuencias de los formantes.
- El patrón tiempo-frecuencia de las frecuencias de los formantes (trayectorias de los formantes).
- El ancho de banda de los formantes.

No vamos a asumir, en principio, que ningún parámetro acústico solo conlleva la información sobre la identidad del locutor que está hablando. Asumiremos, por tanto, que las cualidades de la voz están representadas en varios parámetros cuyo orden de importancia puede variar de hablante a hablante e incluso dependiendo del tipo de habla que se esté analizando.

1.2 Cambio de estilo

Aunque el objeto de este trabajo es, fundamentalmente, el cambio de la identidad de la voz que percibe un oyente, hay una aplicación directa de las técnicas de transformación de hablante al cambio de estilo del habla de un locutor.

El objetivo en este caso es dotar de matices o emociones a la voz que se sintetiza. Por lo tanto parece lógico pensar que debido a que siempre es el mismo locutor el que habla las características propias del tracto vocal permanecerán inalterables. Y, por lo tanto, las distintas emociones o matices de la voz vendrán determinadas por cambios en las características prosódicas y de fuente glotal (tono, duración, pausas, inclinación espectral de la fuente,...). Esta suposición se ve confirmada en [10],[11]. En estos artículos se plantean dos modelos de fuente glotal:

a) Modelo polinómico. Con el sintetizador GELP [9] se hace el siguiente experimento. El modelo de fuente glotal se basa en aproximar, mediante un polinomio de 6º orden, la derivada de la velocidad del flujo glotal o lo que es lo mismo la integral de la señal residual. Con todos los polinomios se genera un codebook de 32 palabras que son las que se utilizan en síntesis. También hay un codebook estocástico de 256 palabras que se usa como excitación de los sonidos sordos.

Se analizan 7 tipos de voz (se graban vocales sostenidas y la frase "we were away a year ago" por un locutor profesional) normal, con carraspera, aérea, áspera, falsete, susurro y ronca. Para cada tipo de voz se genera el codebook de 32 palabras y luego poniendo los LPC de la voz normal se sintetiza con la fuente de otro tipo de voz. Aunque no hay una evaluación formal

concluye que los resultados son indistinguibles del original, y asume, por tanto, que las características de las voces estudiadas dependen, básicamente, de la información contenida en la fuente glotal.

b) Modelo LF. Para tres tipos de voz: normal, con carraspera y aérea, se analizan 4 parámetros del modelo de Fant: anchura del pulso glotal, pendiente del pulso, brusquedad del cierre del pulso glotal y la inclinación espectral del pulso glotal. Los datos se obtienen de 9 locutores cuya voz se define en uno de los tres tipos mencionados anteriormente: tres locutores con voz normal, tres con voz entrecortada y tres con voz aérea.

Se concluye que los cuatro parámetros de la fuente glotal son significativamente diferentes, de forma que usando los de la voz aérea con los formantes de la normal se percibe al locutor normal con voz aérea.

En la misma línea Higuchi, Hirai y Sagisaka [22] consiguen sintetizar voz "con prisa" y "educada" en japonés modificando exclusivamente el tono y la duración de los fonemas, sin embargo concluye que para la voz "enfadada" el cambio prosódico no es suficiente y debería intentarse el cambio en el dominio espectral. Similar resultado obtiene Montero en [36],[37] donde pretende sintetizar voz alegre, triste, enfadada y sorprendida cambiando sólo parámetros prosódicos. Todas las emociones se sintetizan con éxito menos la voz enfadada para la que se sugiere que deberían cambiarse las características de la fuente glotal.

Abe va un paso más allá introduciendo el cambio de formantes para conseguir el cambio de estilo [2]. Analiza 3 tipos de habla del mismo locutor en japonés: lectura de una novela, anuncio y lectura de una enciclopedia. Se analiza F0, duración, energía, tres primeros formantes e inclinación espectral. Todos los factores presentan diferencias significativas entre los tres estilos. En la etapa de síntesis se toma como base el estilo enciclopedia y modificando los parámetros se sintetizan los otros dos estilos. Los resultados son esperanzadores. En cuanto a la conversión de la prosodia, Abe propone convertir sólo el factor global que caracteriza la curva entonativa de la frase y dejar fija la componente local que es la que define los acentos y la entonación de cada palabra en la frase.

1.3 Conversión de hablante

Siguiendo a [30] distinguiremos dos tipos de técnicas en lo que se refiere a la conversión del locutor: llamaremos técnicas paramétricas a aquellas que tratan de modificar parámetros acústicos y técnicas no-paramétricas a aquellas que intentan modificar la identidad del locutor sin hacer un cambio explícito de los parámetros acústicos. Teniendo en cuenta esta definición la transformación de un modelo polinómico de fuente glotal o de los coeficientes LPC del filtro de síntesis serían técnicas no-paramétricas mientras que la conversión de los parámetros del modelo de Fant o de los formantes y anchos de banda serían técnicas paramétricas.

Ya hemos visto en el apartado 1.1."Características acústicas de la identidad del locutor" que no hay un único parámetro responsable de la identidad del locutor. Por lo tanto cualquier sistema que pretenda acometer la tarea de cambiar la identidad de una voz debe transformar, implícita o explícitamente varios parámetros acústicos. Se plantearán, por tanto, dificultades no sólo en cuanto a la transformación en sí, sino también en cuanto a las interrelaciones que existen entre parámetros que deberán ser tenidas en cuenta.

Se han propuesto diversas soluciones dependiendo, sobre todo, de la aplicación o el sistema en el que se va a usar la transformación [13],[14],[17],[23],[24],[25],[35],[38],[39],[43], [44].

En el Grupo de Tecnología del Habla hemos escogido una técnica paramétrica y una no paramétrica y las hemos aplicado a sendos sintetizadores para generar síntesis multilocutor.

2. CONVERSIÓN DE HABLANTE APLICADA A UN SINTETIZADOR DE FORMANTES

El esquema de la conversión se muestra en la figura 1.

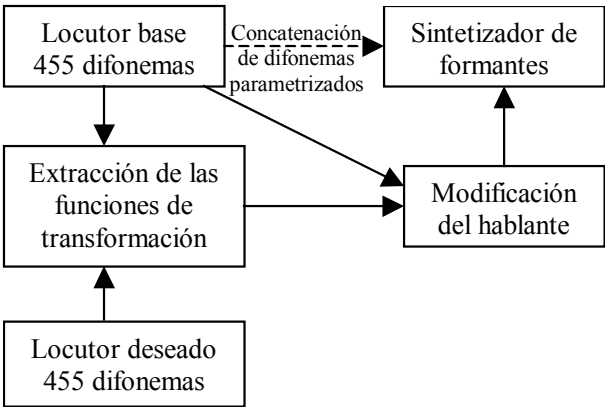

Figura 1. Esquema de la conversión de hablante

La conversión se realiza únicamente para los segmentos sonoros del habla. Como se puede observar en la figura anterior tenemos un sintetizador de formantes que concatena unidades. El esquema del sintetizador es el que se muestra en la figura 2.

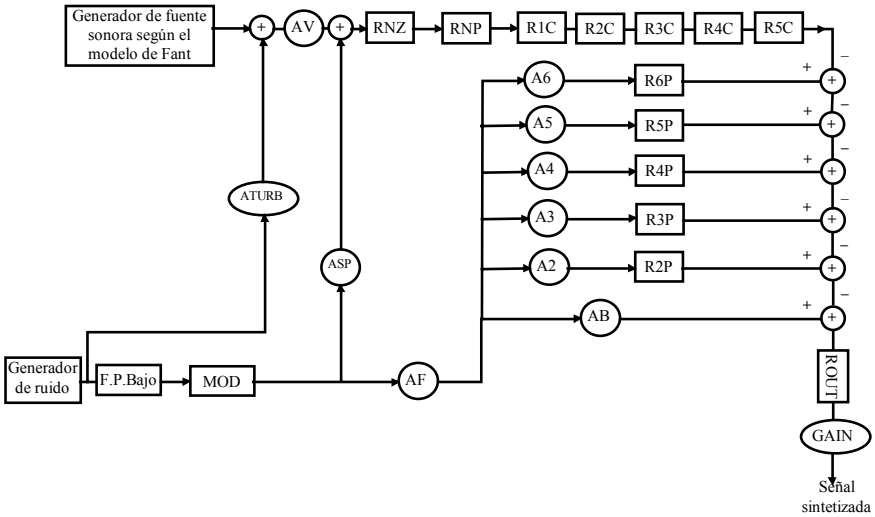

Figura 2. Esquema del sintetizador de formantes

Utilizamos 455 difonemas parametrizados que se concatenan para producir la secuencia deseada. Obtenemos una trama de parámetros por marca de tono fundamental. Cada trama consta de 39 parámetros que se describen en la tabla 1.

Parámetros de la fuente	F0	Tono fundamental
	AV	Ganancia de la excitación sonora
	ASP	Ganancia de aspiración
	ATURB	Ganancia del ruido de turbulencia
	AF	Ganancia del ruido de excitación de los sonidos sordos
	KOPEN	Coeficiente de apertura glotal
	TILT	Inclinación espectral de la señal glotal
	SKEW	Tiempo de cierre del periodo glotal
	FLUTT	Oscilación del tono fundamental
	VELO	Coeficiente de velocidad
	E0	Parámetros de amplitud del pulso glotal
	Ee	
Parámetros del tracto vocal	F1, F2, F3, F4, F5, F6	Frecuencias de los seis principales formantes
	B1, B2, B3, B4, B5	Anchos de banda de los cinco principales formantes de la rama en cascada
	FNZ, BNZ	Frecuencia y ancho de banda del filtro antirresonante RNZ que simula el cero nasal
	FNP, BNP	Frecuencia y ancho de banda del filtro resonante RNP que simula el polo nasal
	A2, A3, A4, A5, A6	Ganancias de la rama paralelo
	B2P, B3P, B4P, B5P, B6P	Anchos de banda de la rama paralelo
	AB	Ganancia de bypass
Ganancia total	GAIN	Ganancia

Tabla 1. Parámetros del sintetizador de formantes

2.1 Conversión de la fuente

En la figura 3 se muestra un periodo de flujo glotal. Basándonos en el modelo de fuente glotal de Liljencrants-Fant (modelo LF) hemos seleccionado cuatro parámetros para definir la onda glotal:

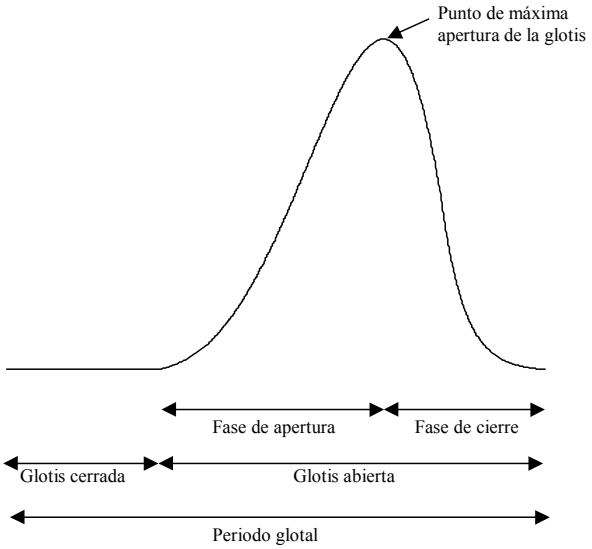

Figura 3. Periodo de flujo glotal

1. Coeficiente de apertura (KOPEN): relaciona el tiempo que permanece abierta la glotis con el tiempo que dura un periodo glotal. Se expresa en tanto por ciento y su rango de variación va del 0%, la glotis no se abre nunca, al 100%, la glotis no se cierra.
2. Coeficiente de velocidad (VELO): Relaciona el tiempo durante el que la glotis está abriéndose con el tiempo durante el que la glotis está cerrándose. Los valores del coeficiente de velocidad pueden ir desde 0 (no hay fase de apertura) hasta infinito (no hay fase de cierre). La inclinación del pulso glotal se representa normalmente por el coeficiente de velocidad.
3. Grado de brusquedad del cierre glotal (SKEW): refleja la relación entre el instante en el que el flujo glotal es máximo y el instante de máxima excitación en su derivada. Valores pequeños de este parámetro indican cierres glotales abruptos mientras que valores grandes indican cierres glotales más lentos.
4. Inclinación espectral (TILT): indica la inclinación espectral del flujo glotal.

Para 8 locutores extraídos de la base de datos EUROM1 se ha realizado el análisis de los parámetros de fuente y hemos llegado a la conclusión de que hay pocas variaciones dentro de un locutor por lo que hemos decidido dejar los 4 parámetros antes mencionados fijos.

En la tabla 2 se muestran los datos obtenidos durante el análisis. Todas las frases fueron resintetizadas con los valores extraídos y dejando los parámetros a un valor constante igual a la media y el resultado era indistinguible.

	OQ		SKEW		TILT		VELO	
	Media	Desv. T.	Media	Desv. T.	Media	Desv. T.	Media	Desv. T.
BA	62.8	14.5	0.181	0.060	21.4	4.9	2.31	0.92
EB	69.5	14.8	0.220	0.073	24.5	3.3	1.81	0.66
GB	68.2	12.3	0.230	0.074	23.8	3.5	1.87	0.69
MB	64.7	12.8	0.243	0.075	23.8	4.1	1.81	0.60
NA	57.8	12.2	0.238	0.074	21.4	4.8	2.16	0.60
RA	68.0	13.1	0.194	0.063	23.7	4.2	1.93	0.83
TA	55.4	13.6	0.236	0.078	22.7	4.5	1.92	0.77
VA	53.6	7.3	0.283	0.084	20.3	4.6	2.35	0.57

Tabla 2. Análisis de los parámetros de fuente de 8 locutores

2.2 Conversión de los parámetros que describen el tracto vocal

Los parámetros que vamos a convertir en este caso son las frecuencias de los formantes y sus anchos de banda. Las pronunciaciones se dividen en segmentos espectralmente estables utilizando una distancia espectral y un umbral.

Los segmentos se alinean mediante DTW (Dynamic Time Warping) y para cada parámetro se define una función mediante regresión lineal. Se obtienen así los coeficientes A y B de la fórmula:

Par_locutor_deseado=A*Par_locutor_de_partida + B

Durante el proceso de síntesis se modifican los formantes y los anchos de banda segmento a segmento y se realiza un suavizado en las transiciones entre segmentos y en las transiciones entre difonemas. En la figura 4 mostramos un ejemplo de conversión de formantes.

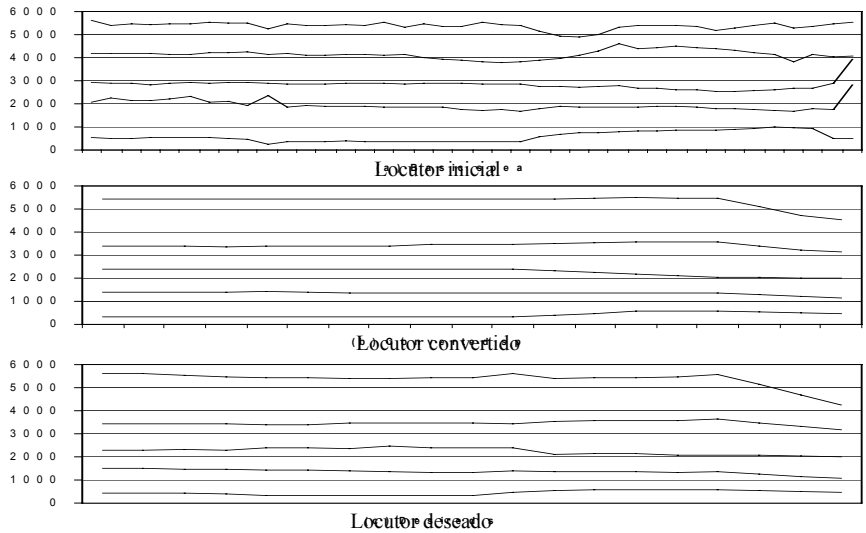

Figura 4. Conversión de los cinco primeros formantes

2.3 Resultados y evaluación

Hemos analizado dos voces masculinas y una femenina. En la etapa de análisis se extrajeron los formantes, la curva de tono y los cuatro parámetros que definen la fuente glotal. Las trayectorias de los formantes se revisaron manualmente y se hicieron las correcciones necesarias.

Hemos empleado dos frases para realizar pruebas: "la bodega del avión" y "mi mamá me mima". Todos los sonidos son sonoros ya que a los sonidos sordos no se les aplica ninguna técnica de conversión de hablante.

Todos los locutores resultaron buenos cuando los utilizamos de locutor inicial. Al aplicar la conversión de hablante se consigue reconocer el locutor deseado sin problemas.

Para evaluar las características de la conversión hemos realizado todas las transformaciones posibles entre los tres locutores analizados y hemos sintetizado la frase "la bodega del avión". A continuación hemos comparado los formantes del locutor convertido con los formantes del locutor deseado para calcular el error cometido. Los resultados se muestran en la tabla 3. Los números en negrita representan la media de cada formante para cada locutor. El resto de los números representan el error medio entre el formante i del locutor inicial convertido y el formante i del locutor deseado en valor absoluto y en porcentaje respecto a la media.

			Locutor inicial		
			HOMBRE1	HOMBRE2	MUJER1
Locutor deseado	HOMBRE1	F1	**399**	53 13%	64 16%
		F2	**1451**	153 11%	119 8%
		F3	**2448**	199 8%	170 7%
		F4	**3707**	202 5%	200 5%
		F5	**5511**	252 5%	678 12%
	HOMBRE2	F1	43 9%	**499**	105 21%
		F2	140 9%	**1592**	246 15%
		F3	148 6%	**2505**	163 7%
		F4	158 5%	**3350**	162 5%
		F5	285 7%	**4193**	388 9%
	MUJER1	F1	72 14%	84 17%	**507**
		F2	163 9%	201 11%	**1825**
		F3	170 6%	194 7%	**2898**
		F4	178 4%	239 6%	**4135**
		F5	230 4%	299 5%	**5449**

Tabla 3. Error cometido al realizar la conversión

Los resultados muestran que los formantes altos se convierten mejor que los bajos, esto puede deberse a que los formantes altos son más estables durante la pronunciación. Creemos que la mayor parte de los errores vienen de la transición entre unidades que es el punto más débil del sistema.

3. CONVERSIÓN DE HABLANTE APLICADA A UN SINTETIZADOR LP-PSOLA

Este trabajo se encuentra en fase inicial. En este caso partimos de un sintetizador LP-PSOLA con 455 unidades de síntesis codificadas CELP. El modelo de fuente es sustituido por una aproximación polinómica de sexto orden de la integral del residuo LP que se corresponde con la derivada del flujo glotal. La conversión se realiza exclusivamente para los segmentos sonoros del habla y se aplica sólo a los parámetros del tracto vocal, es decir los coeficientes del filtro LP.

Para realizar la conversión transformamos los coeficientes LP en LSP más adecuados para la codificación. A continuación se genera un codebook por fonema y se realiza una correspondencia de codebooks entre el locutor inicial y el locutor deseado. El número de palabras que componen el codebook es cuatro.

La correspondencia se realiza alineando los difonemas en el tiempo y anotando la palabra del codebook que se ha asignado al locutor inicial y al locutor deseado.

Los resultados iniciales son esperanzadores.

REFERENCIAS

[1] M. Abe. "A Study on Speaker Individuality Control". PhD Dissertation, Waseda University. 1992

[2] M. Abe. "Speaking Styles: Statistical Analysis and Synthesis by a Text-to-Speech System". Progress in Speech Synthesis, Cap. 39. Ed. Springer. 1997

[3] J. A. Argente. "From speech to speaking styles". Speech Communication, vol. 11, pp 325-335. 1992

[4] L.M. Arslan, D. Talking. "Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum" Proc. Eurospeech'97, vol. 3, pp 1347-1350. Rodhes, 1997

[5] J.A. Bachorowski, M.J. Owren. "Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech". J. Acoust. Soc. Amer., vol 106 (2), pp 1054-1063. 1999

[6] D.G. Childers, D.M. Hicks, B. Yegnanarayana. "Voice Conversion". Speech Communication, vol. 8, pp 147-158. 1989

[7] D.G. Childers, C.K. Lee. "Vocal quality factors; Analysis, synthesis and perception". J. Acoust. Soc. Amer., vol.90 (5), pp 2394-2410. 1991

[8] D.G. Childers, K. Wu. "Gender recognition from speech, Part II; Fine analysis". J. Acoust. Soc. Amer., vol.90 (4), pp 1841-1856. 1991

[9] D.G. Childers, H.T. Hu. "Speech Synthesis by glottal excited linear prediction". J. Acoust. Soc. Amer., vol.96, pp 2026-2036. 1994

[10] D.G. Childers, C. Ahn. "Modeling the glottal volume-velocity waveform for three voice types". J. Acoust. Soc. Amer., vol. 97 (1), pp 505-519. 1995

[11][Childers 95b] D.G. Childers, . "Glottal Source Modeling for Voice Conversion". Speech Communication, vol. 16, pp 127-138. 1995

[12] T. Cleveland, J. Sundberg. "Acoustic analysis of three male voices of different quality". STL-QPSR 4/1983, pp 27-38. 1983

[13] C. d'Alessandro, B. Doval. "Experiments in voice quality modification of natural speech signals: the spectral approach". The Thrid ESCA /COCOSDA Workshop on Speech Synthesis. Jenolan, 1998

[14] V. Darsinos, D. Galanis, G. Kokkinakis. "Designing a speaker adaptable formant-based tex-to-speech system". Eurospeech'97. Rodhes, 1997

[15] S. Furui. "Research on individuality features in speech waves and automatic speaker recognition techniques". Speech Communication, vol. 5, no 2, pp 183-197.1986

[16] C. Gobl, A. N. Chasaide. “Acoustic characteristics of voice quality”. Speech Communication, vol. 11, pp 481-490. 1992

[17] B. Granström. “The use of speech synthesis in exploring different speaking styles”. Speech Communication, vol 11, pp 347-355. 1992

[18] J.M. Gutiérrez-Arriola, Y.S. Hsiao, J.M. Montero, J.M. Pardo, D.G. Childers. “Voice Conversion Based on Parameter Transformation”. Proc. ICSLP'98, vol. 3, pp 987-990. Sidney, 1998

[19] H. M. Hanson. “Glottal characteristics of female speakers: Acoustic correlates”. J. Acoust. Soc. Amer., vol. 101, pp 466-481. 1997

[20] H. M. Hanson, E. S. Chuang. “Glottal characteristics of male speakers: Acoustic correlates and comparison with female data”. J. Acoust. Soc. Amer., vol. 106 (2), pp 1064-1077. 1999

[21] D.E. Hartman, J.L. Danhauer. “Perceptual features of speech for males in four perceived age decades”. J. Acoust. Soc. Amer. Vol.59, pp 713-715. 1976

[22] N. Higuchi, T. Hirai, Y. Sagisaka. “Effect of Speaking Style on Parameters of Fundamental Frequency Contour”. Progress in Speech Synthesis, Cap. 33. Ed. Springer, 1997

[23] C.H. Ho, S. Vaseghi, A. Chen. “Voice Conversion Between UK and US Accented English”. Proc. of Eurospeech’99, vol. 5, pp 2079-2082. Budapest, 1999

[24] N. Iwahashi, Y. Sagisaka. “Speech Spectrum Conversion Based on Speaker Interpolation and Multi-Functional Representation with Weighting by Radial Basis Function Networks”. Speech Communication, vol.16, pp 139-151. 1995

[25] A. Kain, M. Macon. “Personalizing a Speech Synthesizer by Voice Adaptation”. The Thrid ESCA /COCOSDA Workshop on Speech Synthesis. Jenolan, 1998

[26] I. Karlsson. “Glotal wave forms for normal female speakers”. J. Phonetics, vol.14, pp 415-419. 1986

[27] I.Karlsson. “Glottal waveform parameters for different speaker types”. Proc. Speech ’88, 7th FASE Symposium, Vol. 1, pp 225-231. 1988

[28] I. Karlsson. “Female voices in speech synthesis”. J. Phonetics, vol. 19, pp 111-120. 1991

[29] I. Karlsson. “Modelling voice variations in female speech synthesis”. Speech Communication, vol. 11, pp 491-495. 1992

[30] H. Kuwabara, Y. Sagisaka. “Acoustic Characteristics of Speaker Individuality: Control and Conversion”. Speech Communication, vol. 16, pp 165-173. 1995

[31] A.L. Lalwani, D.G. Childers. “Modeling vocal disorders via formant synthesis”. Proc. IEEE ICASSP’91. Vol. 1, pp 505-508. 1991

[32] N.J. Lass, W.S. Brown. “Correlational study of speaker’s height, weight, body surface areas and speaking fundamental frequencies”. J. Acoust. Soc. Amer. Vol.63, pp 1218-1220. 1978

[33] H. Matsumoto, S.Hiki, T.Sone, T. Nimura. “Multidimensional representation of personal quality of vowels and its acoustical correlates”. IEEE Trans. AU, vol. AU-21, pp 428-436. 1973

[34] P.H. Milenkovic. “Voice source model for continuous control of pitch period”. J. Acoust. Soc. Amer. Vol. 93, pp 1087-1096. 1993

[35] H. Mizuno, M. Abe, S. Nakajima. “Development of speech design tool “Sesingn99” to enhance synthesized speech”. Proc. of Eurospeech’99, vol. 5, pp 2083-2086. Budapest, 1999

[36] J.M. Montero, J.M. Gutiérrez-Arriola, S. Palazuelos, S. Aguilera, J.M. Pardo. “Emotional Speech Synthesis: from Speech Database to TTS”. Proc. ICSLP'98, vol. 3, pp 923-926. Sidney, 1998

[37] J.M. Montero, J.M. Gutiérrez-Arriola, J. Colás, E. Enríquez, J.M. Pardo. “Analysis and Modelling of Emotional Speech in Spanish”. Proc. 14th International Congress of Phonetic Sciences, vol. 2, pp 957-960. San Francisco, 1999

[38] E. Moulines, J. Laroche. “Non-parametric techniques for pitch-scale and time-scale modification of speech”. Speech Communication, vol. 16, pp 175-205. 1995

[39] B. Panuthat, T. Funada, N. Kanedera. “Speech Analysis/ Synthesis/ Conversion by Using Sequential Processing”. Proc. IEEE ICASSP’99, vol. I, pp 209-212. Phoenix, 1999

[40] M.D. Plumpe, T.F. Quatieri, D.A. Reynolds. “Modeling of the Glotal Flow Derivative Waveform with Application to Speaker Identification”. IEEE Trans. on Speech and Audio Processing, vol. 7 (5), pp 569-585. 1999

[41] P.J. Price. “Male and female voice source characteristics: Inverse filtering results”. Speech Communication, vol. 8, pp 261-277. 1989

[42] M.F. Schwartz, H.E. Rine. “Identification of speaker sex from isolated, whispered vowels”. J. Acoust. Soc. Amer., vol. 44, pp 1736-1737. 1968

[43] Y. Stylianou, O. Cappé, E. Moulines. “Continuous Probabilistic Transform for Voice Conversion”. IEEE Tran. on Speech and Audio Processing, vol. 6 (2), pp 131-142. 1998

[44] H. Valbret, E. Moulines, J.P. Tubach. “Voice transformation using PSOLA technique”. Speech Communication, vol. 11, pp 175-187. 1992

[45] K. Wu, D.G. Childers. “Gender recognition from speech, Part I; Coarse analysis”. J. Acoust. Soc. Amer. Vol.90, pp 1828-1840. 1991