SISTEMA DE SÍNTESIS ARMÓNICO/ESTOCÁSTICO EN MODO PITCH-ASÍNCRONO APLICADO A CONVERSIÓN DE VOZ

Daniel Erro y Asunción Moreno

Universidad Politécnica de Cataluña, Barcelona

RESUMEN

En el presente artículo se describe un sistema de conversión de voz desarrollado a partir de un sistema de síntesis basado en el modelo armónico/estocástico. A diferencia de otros trabajos similares, no es necesario el funcionamiento pitch-síncrono de la herramienta para que las concatenaciones, modificaciones prosódicas y manipulaciones espectrales se realicen de manera rápida y simple. Los experimentos y comparaciones realizados muestran que el nivel de conversión es altamente satisfactorio a pesar de no usar corpus de entrenamiento paralelos, aunque se pone de manifiesto el compromiso entre el parecido y la calidad de la voz.

1. INTRODUCCIÓN

El propósito de los sistemas de conversión de voz es modificar la voz de un hablante *fuente* de manera que sea percibida como si se tratara de otro hablante *objetivo*, reemplazando las características físicas de la voz sin alterar el mensaje. Las técnicas de conversión de voz encuentran importantes aplicaciones en el campo de la síntesis de habla. En un sintetizador, el habla se genera habitualmente mediante la concatenación de unidades seleccionadas dentro de una base de datos, la cual ha sido construida con anterioridad grabando fragmentos de voz de un locutor cualificado. La inclusión de un sistema de conversión de voz a la salida del sintetizador permite personalizar el sistema sin necesidad de grabar bases de datos de voz de cada uno de los potenciales usuarios. Para ello simplemente se ha de entrenar una función de conversión entre la voz estándar del sintetizador, que será la *fuente*, de la que se dispone de una base de datos completa, y la voz del hablante *objetivo*, del cual se poseen en general pocos datos.

Hay una relación muy cercana entre el sistema de síntesis y el de conversión de voz, hasta el punto de que algunos autores consideran ambos bloques como uno solo. Al convertir una voz en otra no solo se consideran los rasgos espectrales, sino también aspectos prosódicos, y por lo tanto es importante la elección de un sistema de síntesis que permita la manipulación de todas estas características de manera flexible. La calidad de la señal convertida depende en gran medida de la que el sistema de síntesis pueda proporcionar, y se ve limitada por los ruidos de concatenación o discontinuidades espectrales. Por todo ello, es recomendable evaluar el sintetizador y el conversor de voz como un único bloque.

Los sistemas de síntesis basados en TD-PSOLA no son apropiados para la conversión de voz, ya que no asumen un modelo de señal, sino que trabajan directamente con las muestras de audio. Además, son poco resistentes a los ruidos de concatenación de unidades. Existen sistemas basados en LP-PSOLA que ofrecen un buen rendimiento [1, 2], pero se ha observado que a la hora de reconstruir la señal modificada se introducen ruidos derivados de la incoherencia de fase, especialmente cuando se opera sobre el residuo de la codificación LPC. Los sistemas basados en la descomposición armónica o sinusoidal de la señal se han empleado con éxito en tareas de conversión [3, 4], debido a su gran flexibilidad, pero se trata en general de sistemas pitch-síncronos, cuya calidad está condicionada a una correcta detección del inicio y final de cada período de pitch. Todo esto hace pensar en la utilización de un sistema de tipo sinusoidal que permita un mayor control sobre la fase.

En muchos de los trabajos publicados acerca de conversión de voz, se hace uso de un corpus de entrenamiento *paralelo*, es decir, que contiene vectores de parámetros acústicos de los locutores *fuente* y *objetivo* correspondientes al mismo sonido [3, 5]. Para elaborarlo, se acostumbra a grabar las mismas frases pronunciadas por ambos hablantes, alineando convenientemente los fragmentos de sonido equivalentes. Está claro, sin embargo, que no siempre se puede disponer de grabaciones con idénticas frases de cada par de locutores, de modo que recientemente se han propuesto algunas técnicas que permiten resolver este problema [1, 2, 4].

En este trabajo se ha perfeccionado un sistema de síntesis sobre el modelo armónico/estocástico [6] que trabaja con tramas de longitud fija, y se ha implementado un módulo de conversión de voz. Se ha comprobado que su rendimiento es notable aun cuando no se dispone de corpus paralelo de entrenamiento. Esto se debe, en parte, a que aprovecha la enorme flexibilidad del modelo armónico/estocástico para minimizar los errores derivados más propiamente de la reconstrucción de la señal a partir de sus parámetros ya modificados. En la sección 2 se describe el sistema de

síntesis empleado, insistiendo principalmente en los puntos que lo diferencian de otros sistemas parecidos. Se hace especial hincapié en lo referente al tratamiento de las fases de los armónicos. En la sección 3 se explica en detalle el método de conversión de voz. En el marco del proyecto europeo TC-STAR, el sistema completo ha sido evaluado como se detalla en la sección 4, dando lugar a las conclusiones que se exponen en la sección 5.

2. DESCRIPCIÓN DEL SISTEMA DE SÍNTESIS

El modelo armónico/estocástico asume que la señal de voz puede ser representada como la suma de dos componentes: el primero de ellos es una suma de sinusoides armónicamente relacionadas que modelan la parte sonora, y el segundo es un componente de aspecto ruidoso caracterizado por una determinada densidad espectral de potencia, que modela tanto las partes sordas de la señal como cualquier fenómeno de naturaleza no periódica, como puede ser el rozamiento del aire al hablar.

$$s[n] = \sum_{j=1}^{J(n)} A_j[n]\cos\left(\theta_j[n]\right) + e[n] \qquad (1)$$

Ambos componentes cambian a lo largo del tiempo, pero pueden considerarse estables dentro de pequeños intervalos, de modo que la señal es analizada por tramas.

2.1. Análisis y re-síntesis

Se han empleado tramas equiespaciadas centradas en las muestras $n = kN$, con $k = 1, 2, 3\ldots$ N es el número de muestras correspondiente a unos 8 ó 10ms de señal. En adelante, se llamará punto k a la muestra kN.

En cada trama, partiendo de una primera estimación del pitch, se decide si es sorda o sonora. Aunque se puede emplear cualquier método de detección de pitch, en este caso se ha recurrido a la señal electroglotográfica grabada simultáneamente con la voz. Si la k-ésima trama es sonora, se detectan las amplitudes y fases de las sinusoides en el punto k por debajo de 5KHz [7], corrigiendo el valor del pitch si es necesario. La elección de una frecuencia de corte fija se debe a la necesidad de parametrizar las envolventes a la hora de convertir la voz. La parte armónica es interpolada a lo largo de las diferentes tramas [8] y sustraída de la señal original para aislar la parte estocástica, que es analizada en cada trama usando la técnica LPC.

Para resintetizar la señal a partir de sus parámetros, en cada punto k se construye una trama de longitud $2N$ centrada en dicho punto sumando los armónicos detectados con amplitud, frecuencia y fase constantes, y añadiendo la contribución estocástica obtenida al filtrar ruido blanco y gaussiano a través del filtro LPC correspondiente.

$$s^{(k)}[n] = \sum_{j=1}^{J^{(k)}} A_j^{(k)} \cos\left(2\pi j \frac{f_0^{(k)}}{f_s} n + \varphi_j^{(k)}\right) + \sigma[n] * h_{LPC}^{(k)}[n] \qquad (2)$$

con n=–N,…,N–1. Las tramas se solapan entre sí usando ventanas triangulares.

$$s[kN+m] = \left(\tfrac{N-m}{N}\right)\cdot s^{(k)}[m] + \left(\tfrac{m}{N}\right)\cdot s^{(k+1)}[m-N] \qquad (3)$$

para m entre 0 y N–1. La señal resintetizada es prácticamente indistinguible de la original.

2.2. Modificaciones prosódicas

Dado que se están usando tramas de tamaño fijo, se han tenido que desarrollar procedimientos que permiten modificar los parámetros de cada trama de la señal de forma sencilla, sin perder la coherencia de fase entre diferentes tramas. Para ello se considera que las fases medidas son el resultado de sumar un término lineal en frecuencia y la respuesta en fase del tracto vocal, que varía con el tiempo.

2.2.1. Modificación en duración

Para modificar la duración de la señal, basta con cambiar el valor de N en la ecuación (3) para que las variaciones en amplitud y frecuencia se adapten a la nueva escala temporal. No obstante, las fases no se pueden mantener invariantes en cada punto k porque ha variado la distancia entre puntos de análisis, de modo que se rompe la coherencia. Es por eso que hay que corregir el término lineal de fase de modo que solamente la respuesta en fase del tracto vocal se mantenga en cada punto k sin perder la coherencia. Suponiendo que no cambia el tracto vocal y que el pitch varía linealmente entre los puntos k–1 y k, el incremento de fase del primer armónico entre dichos puntos se puede aproximar por medio de la función Ψ:

$$\Psi\left(f_0^{(k-1)}, f_0^{(k)}, N\right) = \left(f_0^{(k-1)} + f_0^{(k)}\right)\cdot \pi N \tfrac{1}{f_s} \qquad (4)$$

siendo f_s la frecuencia de muestreo. Se propone la siguiente corrección para las fases:

$$\Delta\varphi_1^{(k)} = \Psi\left(f_0^{(k-1)}, f_0^{(k)}, N'\right) - \Psi\left(f_0^{(k-1)}, f_0^{(k)}, N\right) \qquad (5a)$$

$$\varphi_j'^{(k)} = \varphi_j^{(k)} + j\sum_{q=1}^{k} \Delta\varphi_1^{(q)} \quad j = 1\ldots J^{(k)} \quad \forall k \qquad (5b)$$

siendo N' la nueva longitud de trama y Ψ la definida en (4). La corrección no afecta al tracto vocal sino solamente al término lineal de fase, como se pretendía. No es necesario modificar los coeficientes LPC de la parte estocástica.

2.2.2. Modificación en pitch

Para modificar el pitch de la señal, se multiplican las frecuencias por el factor deseado y se recalcula el número de armónicos hasta 5KHz. Las amplitudes de los nuevos armónicos se obtienen mediante una simple interpolación lineal entre las log-amplitudes medidas en el análisis [9], manteniendo la energía original con un factor multiplicativo. Si se elimina el término lineal de fase, la respuesta en fase del tracto vocal en los nuevos

armónicos se puede obtener interpolando linealmente la parte real e imaginaria de las amplitudes complejas [9]. El término lineal se suprime del siguiente modo [10]:

$$\varphi_{VT\,j}^{(k)} = \varphi_j^{(k)} - j\alpha^{(k)} \quad (6a)$$

$$\alpha^{(k)} = \arg\min \sum_{j=0}^{J^{(k)}-1} \left| \sqrt{A_j^{(k)}} e^{i\left(\varphi_j^{(k)} - j\alpha^{(k)}\right)} - \sqrt{A_{j+1}^{(k)}} e^{i\left(\varphi_{j+1}^{(k)} - (j+1)\alpha^{(k)}\right)} \right|^2 \quad (6b)$$

tomando $A_0^{(k)}=A_1^{(k)}$ y $\varphi_0^{(k)}=0$. Una vez calculada la fase del tracto vocal en los nuevos armónicos, se repone el término lineal usando el mismo valor de α. También en este caso habría que actualizar el término lineal de fase, pues la longitud del período cambia mientras que N se mantiene constante. La corrección se realiza por medio de la fórmula (5b) con el incremento siguiente:

$$\Delta\varphi_1^{(k)} = \Psi\left(f_0'^{(k-1)}, f_0'^{(k)}, N\right) - \Psi\left(f_0^{(k-1)}, f_0^{(k)}, N\right) \quad (7)$$

Los coeficientes de la parte estocástica no se modifican. Tanto el factor de modificación de pitch como el de duración pueden ser variables con el tiempo, y se pueden efectuar los dos tipos de modificaciones de forma simultánea.

2.3. Concatenación de unidades

Se dispone de una base de datos analizada como se ha descrito. Se desea concatenar una serie de unidades siguiendo unas especificaciones concretas de duración, energía y contorno de pitch. A cada unidad se le aplican las transformaciones prosódicas pertinentes mediante la modificación de sus amplitudes, frecuencias, fases y ganancia de los filtros LPC. De nuevo el esquema pitch-asíncrono plantea el problema de la coherencia de fase entre las unidades. Para evitarlo, se debe imponer que la componente lineal de la fase sea continua en el punto de concatenación. Siguiendo este planteamiento, para cada dos unidades adyacentes A y B se calculan en los puntos limítrofes k_A y k_B las correspondientes $\alpha^{(k_A)}$ y $\alpha^{(k_B)}$ conforme a la ecuación (6b), y se ajustan las fases de la segunda unidad imponiendo que el incremento del término lineal de fase entre k_A y k_B sea el dado por la función Ψ definida en (4):

$$\varphi_j'^{(k)} = \varphi_j^{(k)} + j\cdot\left(-\alpha^{(k_B)} + \alpha^{(k_A)} + \Psi\left(f_0^{(k_A)}, f_0^{(k_B)}, N\right)\right) \quad (8)$$
$$j = 1\ldots J^{(k)} \quad k = k_B, k_B+1, \ldots$$

Por ultimo, se suavizan las amplitudes de los armónicos en torno al punto de concatenación de modo que no haya discontinuidades espectrales que den lugar a ruido.

3. MÉTODO DE CONVERSIÓN DE VOZ

Para que la transformación entre dos hablantes sea óptima, habría que tener en cuenta aspectos muy diversos, desde la curva de entonación de cada uno hasta los vocablos o expresiones que los caracterizan. Este trabajo se centra solamente en la conversión espectral, aunque se realiza también una transformación básica de pitch. A continuación se desarrollan cada una de las etapas que componen el bloque de conversión.

3.1. Normalización del pitch

La frecuencia fundamental se caracteriza por una distribución de tipo log-normal. A partir de los datos disponibles de cada hablante se extrae una estimación de la media y la varianza de su $logf_0$. Se realiza la siguiente normalización:

$$\log f_0^{(\text{converted})} = \mu^{(\text{target})} + \frac{\sigma^{(\text{target})}}{\sigma^{(\text{source})}}\left(\log f_0^{(\text{source})} - \mu^{(\text{source})}\right) \quad (9)$$

3.2. Conversión del tracto vocal

El tracto vocal puede ser modelado por medio de distintos tipos de parámetros: amplitudes normalizadas a una determinada frecuencia fundamental, coeficientes cepstrales discretos, coeficientes LSF, etc. Los coeficientes LSF son los más habituales en conversión de voz, debido a que modelan bien la estructura de formantes y a que el error de estimación en uno de estos coeficientes afecta a una pequeña parte del espectro. En el presente trabajo se tuvieron en cuenta otras ventajas: estos coeficientes codifican también una envolvente de fase mínima, y además en la parte estocástica se emplea el mismo tipo de codificación, lo que permite relacionar ambas componentes de la señal como se verá más adelante.

Para traducir las amplitudes a coeficientes LSF se calcula el filtro todo-polos óptimo mediante la técnica del modelado todo-polos discreto [11], que trata de minimizar la medida de distorsión de Itakura-Saito entre la magnitud del filtro y las amplitudes de los armónicos. El orden del filtro utilizado es 14 para una frecuencia de muestreo de 16KHz, pues órdenes superiores aportan información poco relevante a la vez que complican la conversión. Para la obtención de la función de transformación se ajusta un modelo de mezclas gaussianas de orden m=8 a un corpus de entrenamiento compuesto por vectores de la forma $v=[x^T\ y^T]^T$, donde los vectores columna x e y contienen p=14 parámetros LSF del hablante fuente y objetivo, respectivamente, correspondientes a los mismos fonemas [5]. Se usa el algoritmo EM para buscar los α_i, μ_i y Σ_i tales que sea máxima la verosimilitud del conjunto de vectores v del corpus de entrenamiento, de acuerdo a

$$p(v) = \sum_{i=1}^{m} \alpha_i N(v, \mu_i, \Sigma_i) \quad (10a)$$

$$N(v, \mu, \Sigma) = \frac{1}{(2\pi)^{p/2}} |\Sigma|^{-1/2} \exp\left\{-\tfrac{1}{2}(v-\mu)^T \Sigma^{-1}(v-\mu)\right\} \quad (10b)$$

La función de transformación se define como

$$F(x) = \sum_{i=1}^{m} p_i(x)\cdot\left[\mu_i^y + \Sigma_i^{yx}\left(\Sigma_i^{xx}\right)^{-1}\left(x-\mu_i^x\right)\right] \quad (11a)$$

$$p_i(x) = \frac{\alpha_i N\left(x, \mu_i^x, \Sigma_i^{xx}\right)}{\sum_{j=1}^{m} \alpha_j N\left(x, \mu_j^x, \Sigma_j^{xx}\right)} \quad (11b)$$

$$\mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix} \quad \Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix} \qquad (11c,d)$$

Para la obtención de la función óptima de transformación es necesario disponer de corpus de entrenamiento paralelos. En este trabajo, sin embargo, se han construido corpus paralelizados, generando con ayuda de un sintetizador las mismas frases que se disponen del locutor objetivo [1].

3.3. Estimación de la envolvente de fase

Para que la voz sintetizada sea de alta calidad, las variaciones temporales en la magnitud del tracto vocal deben ir acompañadas de sus correspondientes variaciones en fase, manteniendo a la vez la coherencia entre tramas. Para satisfacer estos dos objetivos, se calcula en primer lugar un término lineal de fase de forma recursiva:

$$\varphi_j^{(k)} = \varphi_j^{(k-1)} + j\Psi\left(f_0^{(k-1)}, f_0^{(k)}, N\right) \qquad (12)$$

Al comienzo de cada región sonora, las fases son inicializadas a cero. Este primer paso no provoca discontinuidades de fase importantes, pero al no haber variaciones en la fase del tracto vocal entre las diferentes tramas, puede ocasionarse un desagradable ruido metálico en la señal sintetizada. Por este motivo se añade un segundo término:

$$\varphi_j'^{(k)} = \varphi_j^{(k)} + \arg\left\{1/A\left(f_j^{(k)}\right)\right\} \qquad (13)$$

siendo $1/A(f)$ el filtro todo-polos convertido a partir de las amplitudes de los armónicos. Este término no representa realmente la envolvente de fase del tracto vocal, pero proporciona variaciones de fase coherentes con las amplitudes, y la calidad que se alcanza es muy aceptable.

3.4. Predicción de la parte estocástica

Es sabido que la relevancia de la parte estocástica en conversión de voz es escasa en relación con la parte armónica de la señal [4]. Sin embargo, se debe tener en cuenta que la separación entre ambas componentes no es perfecta, siendo muy difícil separarlas por completo en tramas sonoras. Además, por encima de la frecuencia de corte de 5KHz se considera que no hay armónicos, cosa que no es realista a pesar de la alta calidad lograda. Por estos y otros motivos, se observa una gran correlación entre la forma del tracto vocal y la de la envolvente LPC que caracteriza la parte estocástica. Aprovechando que sus parametrizaciones son similares y que se dispone de un modelo de mezclas gaussianas que divide en clases solapadas el espacio acústico del tracto vocal, se pueden entrenar vectores η_i y matrices Γ_i tales que se optimice a lo largo del corpus de entrenamiento la relación siguiente:

$$y_{est} = \sum_{i=1}^{m} p_i(y) \cdot \left[\eta_i + \Gamma_i\left(\Sigma_i^{yy}\right)^{-1}\left(y - \mu_i^y\right)\right] \qquad (14)$$

siendo y_{est} el vector LSF estocástico. A la hora de convertir, para obtener la parte estocástica se empleará la fórmula (14) con el vector $F(x)$ calculado previamente en (11a) en lugar de y.

En las zonas sordas no se ha realizado transformación alguna, pues su importancia perceptual es muy relativa y los intentos de conversión efectuados no mejoran el resultado de manera visible, comprometiendo en algunos casos la calidad.

4. EXPERIMENTOS

Para evaluar el sistema de síntesis-conversión de voz, se han usado grabaciones de alta calidad de más de 150 frases, de duración en torno a 5 segundos, pronunciadas por 4 locutores profesionales: 2 hombres y 2 mujeres. El 80% del material grabado se ha utilizado para la elaboración del corpus de entrenamiento, y el resto para la evaluación propiamente. Se han usado otros dos métodos de conversión como referencia, cuyas características son las siguientes [1]:

- M1: se trata de un sintetizador basado en TD-PSOLA creado a partir de las grabaciones del locutor objetivo, de modo que no convierte voces, pero sirve como referencia.
- M2: se trata de un sistema basado en LP-PSOLA. Utiliza como datos de entrada no sólo las muestras de audio sino también información privilegiada de tipo fonético: vocal/ consonante, sordo/sonoro, etc. Usa árboles de decisión para clasificar los sonidos y aplicar distintas funciones de conversión a cada clase. Además, aumenta la resolución de la conversión mediante un proceso de selección+suavizado de tramas residuales de la codificación LPC. El sistema se ha entrenado con frases paralelas entre sí.

A diferencia del método M2, para el entrenamiento de la función de conversión del sistema propuesto se han creado corpus no paralelos con ayuda de un sintetizador, de modo que se evalúa la capacidad del sistema para actuar en situaciones en las que no se dispone de grabaciones de las mismas frases para todos los locutores.

Se han realizado pruebas de conversión en todas las direcciones posibles entre los locutores disponibles, a partir de las frases de test. Los resultados finales se han obtenido promediando los resultados de las conversiones individuales. Para obtener estos resultados, varios oyentes escucharon parejas de fragmentos de audio correspondientes a la voz convertida y la voz objetivo, puntuando el parecido entre 1 (diferentes) y 5 (iguales). Se puntuó también la calidad de la voz convertida entre 1 (mala) y 5 (excelente). Los resultados son los expuestos en la tabla 1:

	Parecido	Calidad
M1	3.35	3.20
M2	3.47	2.25
Propuesto	3.15	2.38

Tabla 1. *Resultados de la evaluación*

Se pueden realizar diversas observaciones. En primer lugar, cabría esperar una mayor puntuación en el parecido obtenido con M1, puesto que usa directamente la voz objetivo. Se puede concluir que el decremento en la calidad de la voz sintetizada debida a los pocos datos usados en síntesis, afecta de manera directa a la percepción del parecido por parte de los oyentes. Aclarado este punto, se puede comprobar que el sistema propuesto se encuentra prácticamente al nivel de M1 y M2, aun cuando estos utilizan información privilegiada, así como corpus paralelos de entrenamiento en el caso de M2. Los resultados obtenidos en cuanto a parecido de la voz son coherentes con el estado del arte.

La calidad obtenida es más baja de lo esperado. Los oyentes parecen preferir la voz de un sintetizador construido con pocos minutos de audio que la voz manipulada y convertida. No obstante, se obtienen mejores resultados en el sistema propuesto que en M2, lo cual parece ser debido sobre todo a haber evitado problemas derivados de la coherencia de fase. Aunque se esperaba una mayor mejora, hay que señalar que el proceso de alineado del corpus no paralelo puede estar influyendo negativamente en la calidad, mientras M2 está libre de estos fallos. En cualquier caso, se puede concluir que la manipulación de la señal ocasiona una pérdida en la naturalidad de la voz que hace que los oyentes la consideren de peor calidad.

Los resultados invitan a plantearse la búsqueda de métodos de manipulación que degraden en menor medida la calidad de la señal, pues el oyente prefiere una alta calidad que un perfecto parecido. Por lo tanto, en futuros trabajos se trabajará en aspectos mejorables como por ejemplo la utilización de envolventes de fase más realistas que la aproximación de fase mínima, así como de funciones de transformación de tracto vocal que provoquen un menor suavizado espectral y ensanchamiento de los formantes.

5. CONCLUSIONES

En este artículo se ha desarrollado un sistema de conversión de voz sobre un sintetizador basado en el modelado armónico/estocástico. Se ha logrado un grado satisfactorio de parecido entre voces convertidas y objetivos, al nivel de otros sistemas que usan información privilegiada. Sin embargo, aunque se superan problemas derivados de la coherencia de fase como se pretendía, no se logra evitar la degradación de la señal debida a la manipulación espectral. Es necesario seguir perfeccionando el método de conversión para que se comprometa menos la calidad de la señal.

6. AGRADECIMIENTOS

Este trabajo se ha llevado parcialmente a cabo con fondos procedentes del proyecto TC-STAR, *Technology and Corpora for Speech-to-Speech Translation* (FP6-506738). Los autores agradecen también a Javier Pérez y Ferran Diego su ayuda en la elaboración de los corpus de entrenamiento.

7. BIBLIOGRAFÍA

[1] H.Duxans, D.Erro, J.Perez, F.Diego, A.Bonafonte, A.Moreno, “Voice Conversion of Non-Aligned Data using Unit Selection”, TC-STAR Workshop on Speech-to-Speech Translation, 2006.

[2] D.Sündermann, H.Höge, A.Bonafonte, H.Ney, J.Hirschberg, “TC-Star: Cross Language Voice Conversion Revisited”, TC-STAR Workshop on Speech-to-Speech Translation, 2006.

[3] Y.Stylianou, O.Cappé, E.Moulines, “Continuous Probabilistic Transform for Voice Conversion”, IEEE Trans. on Speech and Audio Proc., 1998.

[4] H.Ye, S.Young, “Quality-enhanced Voice Morphing using Maximum Likelihood Transformations”, IEEE Trans. On Audio, Speech and Lang. Proc., 2006.

[5] A.Kain, M..Macon, “Spectral Voice Conversion for Text-to-Speech Synthesis”, ICASSP’98, 1998.

[6] D.Erro, A.Moreno, “A Pitch-Asynchronous Simple Method for Speech Synthesis by Diphone Concatenation using the Deterministic plus Stochastic Model”, Proc. SPECOM, 2005.

[7] Ph.Depalle, T.Hélie, “Extraction of Spectral Peak Parameters using a STFT Modeling and no Sidelobe Windows”, Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 1997.

[8] R.J.McAulay, T.F.Quatieri, “Speech Analysis/ Synthesis based on a Sinusoidal Representation”, IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 34 (4), pp. 744-754, 1986.

[9] E.R.Banga, C.García-Mateo, X.Fernández-Salgado, “Concatenative Text-to-Speech Synthesis based on Sinusoidal Modeling”, Improvements in Speech Synthesis, John Wiley and Sons, 2001.

[10] D.Chazan, R.Hoory, Z.Kons, D.Silberstein, A.Sorin, “Reducing the Footprint of the IBM Trainable Speech Synthesis System”, Proc. ICSLP, 2002.

[11] A.El-Jaroudi, J.Makhoul, “Discrete All-Pole Modeling”, IEEE Trans. on Signal Proc., 1991.