Universidad Zaragoza

1542

Trabajo Fin de Grado

Sistemas de alteración de la voz para falsear la identidad en sistemas de verificación de locutor.

Autor

Santiago Ramírez Aretio

Director/es

Eduardo Lleida Solano

Escuela de Ingeniería y Arquitectura

2014

Sistemas de alteración de la voz para falsear la identidad en sistemas de verificación de locutor.

Resumen:

En este trabajo se exploran diferentes técnicas para modificar la información sobre la identidad del locutor de señales grabadas de voz. Aplicando el modelo excitación-filtro para el proceso de producción del habla, el objetivo es modificar la información de la señal de excitación y del tracto vocal por separado.

Para valorar las trasformaciones realizadas se ha usado un sistema de verificación de locutor del estado del arte. Usando las trasformaciones propuestas se simulará un ataque al verificador donde se tratará de modificar u ocultar la identidad de un locutor.

Indice

1-Introducción

La señal de voz captada por un micrófono contiene mucha información. En primer lugar contiene sonidos que forman las palabras de un mensaje que un locutor quiere transmitir. Pero también incluye información acerca del entorno acústico, el idioma, el estado emocional o la identidad el locutor. Este TFG se centra en este último tipo de información.

Por diversas circunstancias, una persona puede querer modificar su voz. Por ejemplo un cantante puede querer subirla o bajarla un tono o un delincuente puede querer modificarla para no ser reconocido, etc. En este TFG vamos a desarrollar un sistema de transformación de la señal de voz que permita modificar las características de dicha señal relacionadas con la identidad del locutor.

La señal de voz puede descomponerse en una excitación y un filtro que modela la forma del tracto vocal. La mayoría que de las aplicaciones para la modificación de la señal de voz que pueden obtenerse en Internet transforman la información de excitación y del tracto vocal al mismo tiempo. En este TFG vamos a explorar diversas técnicas para modificar la información de la excitación y del tracto vocal por separado. La primera transformación consistirá en un desplazamiento frecuencial sobre la señal de excitación para modificar el pitch. La segunda consistirá en un mapeo entre las características del tracto vocal de dos locutores. Después, para valorar las transformaciones realizadas, usaremos un sistema de verificación de locutor.

Los sistemas de verificación de locutor explotan la información sobre la identidad del locutor de la señal de voz para tratar de identificar a personas de manera automática. Otros sistemas de verificación tradicionales basan la identificación de una persona en un objeto o en un conocimiento que solo esa persona puede tener, como por ejemplo una llave o una contraseña. Por el contrario los sistemas de verificación de locutor basan la identificación en la voz, una característica propia de cada persona y que no se puede perder u olvidar. Esta característica hace interesante el uso de estos sistemas en aplicaciones de seguridad tales como el control de acceso o el análisis forense.

El modelo digital de producción del habla permite parametrizar el proceso de generación de la señal de voz. Modificando los parámetros relacionados con la identidad del locutor podemos hacer

transformaciones que puedan ocultar la identidad o suplantar la de otra persona. En los sistemas de verificación de locutor, esto supone dos problemas bien diferenciados:

-Tampering: Una persona intenta ocultar su identidad para no ser detectada.

-Spoofing: Una persona intenta suplantar la identidad de otra para, por ejemplo, tener acceso a sus privilegios en un sistema de seguridad.

En [1] se puede encontrar una clasificación de las diferentes técnicas de spoofing y tampering existentes. En este TFG se proponen varios sistemas para la transformación de parámetros que contienen información del locutor. El objetivo principal es estudiar el comportamiento de estas transformaciones y sus implicaciones en un sistema de verificación de locutor.

La memoria de este trabajo esta dividida en 6 secciones:

1. En esta sección se realiza una introducción a la temática y al trabajo realizado.

2. En esta sección se expone el modelo digital de producción del habla basado en filtro y excitación y se hace una breve introducción a los sistemas de verificación de locutor.

3. En esta sección se proponen varios sistemas para modificar la información de locutor de grabaciones de voz.

4. En esta sección se exponen un serie de experimentos realizados para medir el rendimiento de los sistemas propuestos. También se exponen los resultados de utilizar los sistemas propuestos para realizar un ataque de spoofing y tampering a un sistema de verificación de locutor.

5. En esta sección se realizan conclusiones generales del trabajo realizado y se proponen posibles líneas futuras.

6. En esta sección aparecen todas las referencias que se han aparecido durante la memoria.

2-Modelo digital de la señal de voz

2.1-Sistema de producción del habla

La producción del habla humana es un proceso complejo donde, desde los pulmones hasta los orificios nasales y labios, intervienen un extenso conjunto de órganos. Una manera sencilla de explicar el proceso de producción consiste en dividir el sistema en excitación y tracto vocal.

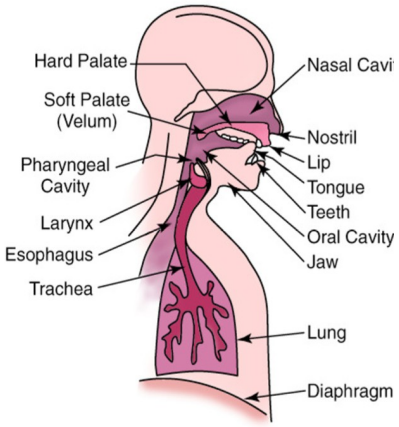

Figura 2.1.1: Órganos que intervienen en el sistema de producción del habla

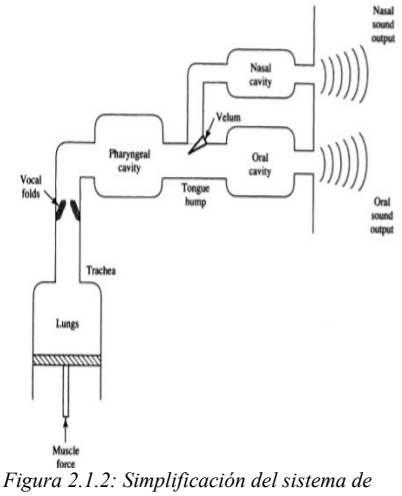

Figura 2.1.2: Simplificación del sistema de producción del habla.

-Excitación:

Los pulmones generan un flujo de aire que atraviesa las cuerdas vocales. Si estas oscilan pueden modular el flujo de aire dándole cierta periodicidad. Los sonidos producidos mediante este fenómeno se llaman sonidos sonoros. Las cuerdas vocales también pueden permanecer relajadas mientras les atraviesa el flujo de aire produciendo así sonidos sordos.

La frecuencia de oscilación se llama frecuencia fundamental o pitch. El valor de pitch depende de la longitud y tensión de las cuerdas vocales y por lo tanto es característico de cada persona. Para hombres el valor de pitch suele estar entre 50 Hz y 250 Hz y para mujeres y niños entre 120 Hz y 500 Hz. El ser humano también puede aumentar y disminuir el valor de pitch con el objetivo de cambiar la entonación del habla.

-Tracto vocal:

El tracto vocal esta formado por todos los órganos entre las cuerdas vocales y la nariz y boca. Cuando la onda acústica atraviesa el tracto vocal, este actúa como un filtro acústico modificando la distribución espectral de la onda original. Se puede demostrar[2] que el filtro acústico se compota como un filtro todo-polos, por lo tanto su respuesta frecuencial está caracterizada por una serie de frecuencias de resonancia. Estas frecuencias de resonancia se llaman formantes y su valor depende de la forma del tracto vocal. Puesto que la forma del tracto vocal de cada persona es diferente, el espectro de la señal de voz contiene información del locutor. La mayoría de los sistemas de verificación de locutor usan únicamente características derivadas de la forma del tracto vocal.

2.2- Modelo del tracto vocal: predicción lineal

El tracto vocal puede modelarse como un tubo donde el área de cada sección $A(x)$ varia con la longitud. Discretizando la longitud en p intervalos equidistantes podemos representar la forma del tracto vocal con el valor del área de la secciones en cada intervalo A_i.

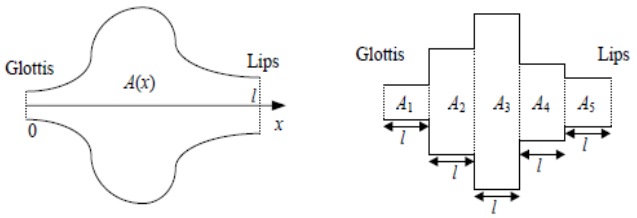

Figura 2.2.1: Modelado de la forma del tracto vocal por 5 tubos de igual longitud y distinto radio.

De acuerdo con este modelo de p tubos, en cada unión un parte de la onda acústica es reflejada y otra transmitida. Se define el coeficiente de reflexión k_i como el cociente entre la onda reflejada e incidente en la i-ésima unión. Considerando tubos sin pérdidas se cumple:

$$k_i = \frac{A_{i+1} - A_i}{A_{i+1} + A_i} \quad A_i > 0 \Rightarrow |k_i| < 1$$

Si consideramos que la velocidad de propagación es constante, el tiempo requerido por la onda

acústica para atravesar cada sección es constante. Este hecho permite una transformación inmediata de las ondas acústicas de entrada y de salida al dominio digital. Se puede demostrar [3] que la transformada-z de una concatenación de p tubos sin pérdidas se corresponde con un sistema auto-regresivo de p polos. De este modo podemos descomponer la señal de voz $s[n]$ en un excitación $e[n]$ filtrada por un filtro todo polos $H(z)$ que modela la forma del tracto vocal:

$$H(z)=\frac{S(z)}{E(z)}=\frac{1}{A(z)}=\frac{1}{1-\sum_{k=1}^{p} a_k \cdot z^{-k}}$$

De la ecuación anterior deducimos que:

$$e[n]=s[n]-\sum_{k=1}^{p} a_k\, s[n-k]$$

En la práctica dispondremos de la señal de voz $s[n]$ y nos interesará estimar los coeficientes a_i para obtener información de la forma del tracto vocal del locutor. Una forma muy común de hacerlo es calcular los coeficientes a_i que minimizan el error cuadrático medio de la señal de error. De este modo $A(z)$ actuará como un filtro de predicción lineal. Sus coeficientes a_i serán los coeficientes LP(Linear predictive) y *e[n]* el residuo de la predicción. La solución [4] a las ecuaciones del predictor lineal son las ecuaciones de Yule-Walker:

$$\begin{pmatrix} R_s[0] & R_s[1] & \cdots & R_s[p-1] \\ R_s[1] & R_s[0] & \cdots & R_s[p-2] \\ \vdots & \vdots & \ddots & \vdots \\ R_s[p-1] & R_s[p-2] & \cdots & R_s[0] \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} R_s[1] \\ R_s[2] \\ \vdots \\ R_s[p] \end{pmatrix}$$

Donde $R_s[k]$ es la función de autocorrelación de la señal $s[n]$. La matriz anterior es de tipo Toeplitz ya que es simétrica y los elementos de todas sus diagonales son iguales. El algoritmo Levinson-Durbin[5] permite hallar el inverso este tipo de matrices de forma eficiente.

-*Consideraciones prácticas del análisis LP localizado*

La señal de voz puede considerarse localmente estacionaria para ventanas temporales de aproximadamente 40 ms. Definimos $s_m[n]$ como un segmento de N muestras de la señal de voz $s[n]$ en torno a la muestra m-ésima.

$$s_m[n] = s[n]w[n-m]$$

Donde $w[n]$ es la función de enventanado y cumple:

$$w[n] = 0 \quad si \quad |n| > N/2$$

En la práctica elegiremos un valor N de tal forma que la duración del segmento sea de aproximadamente 20 ms y supondremos que las variables estadísticas de la señal se mantienen constantes en ese segmento. Estimaremos la autocorrelación de la señal en aquellos instantes m que sean múltiplo de un determinado número de muestras *hop* como:

$$R_m[k] = \sum_{n=0}^{N-k-1} s_m[n]s_m[n-k]$$
$$m = i \cdot hop \quad i = 1,2,3...$$

Después, utilizando el algoritmo Levinson-Durbin, podremos calcular los coeficientes a_i del filtro de predicción lineal $A(z)$ que modelarán la forma del tracto vocal de un locutor en cada instante m. Siguiendo [6] se pueden calcular los coeficientes de reflexión a partir de los coeficientes LP. La señal de excitación $e_m[n]$ se podrá calcular directamente filtrando $s_m[n]$ con $A(z)$.

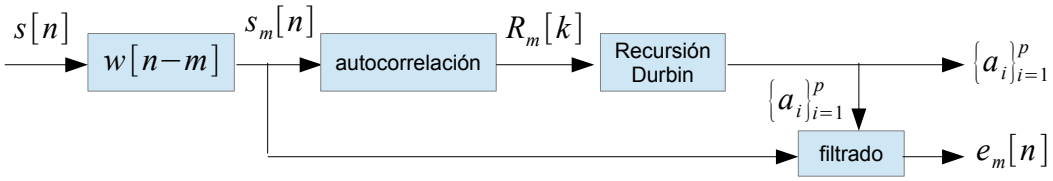

Figura 2.2.2: Esquema del análisis LP localizado

2.3- Cepstrum real

Mediante predicción lineal, la forma del tracto vocal queda representada por los coeficientes de un filtro IIR. Sería interesante encontrar otra representación del tracto vocal cuya manipulación resultara más sencilla y versátil. El cepstrum real es una transformación homomórfica que permite convertir las convoluciones en sumas en el dominio transformado. El cepstrum real $c[n]$ de una señal $x[n]$ puede calcularse como:

$$c[n]=TDF^{-1}\left\{\log\left|TDF\left\{x[n]\right\}\right|\right\}$$

Si $x[n]=x_1[n]\otimes x_2[n]$ es fácil comprobar que se cumple:

$$c[n]=TDF^{-1}\left\{\log\left|TDF\left\{x_1[n]\otimes x_2[n]\right\}\right|\right\}=TDF^{-1}\left\{\log\left|X_1[k]X_2[k]\right|\right\}=c_1[n]+c_2[n]$$

Siendo $c_1[n]$ y $c_2[n]$ las transformadas cepstrum de $x_1[n]$ y $x_2[n]$ respectivamente. Si consideramos el modelo excitación y filtro de la señal de voz:

$$s[n]=e[n]\otimes h[n]$$

Donde $h[n]$ es la respuesta impulsional del filtro que modela el tracto vocal. La excitación $e[n]$ puede aproximarse como ruido blanco para sonidos sordos o como un tren de deltas separadas N_0 muestras para sonidos sonoros, donde N_0 es el periodo de pitch en muestras. Si se calcula la transformada cepstrum $c_s[n]$ de la señal de voz:

$$c_s[n]=c_e[n]+c_h[n]$$

Donde $c_e[n]$ y $c_h[n]$ son las transformadas cepstrum de $e[n]$ y $h[n]$ respectivamente. En [7] se puede encontrar una demostración de que para sonidos sordos se cumple:

$$c_e[n]\approx 0$$

mientras que para sonidos sonoros:

$$c_e[n] \approx \sum_{r=1}^{\infty} \frac{\delta[n - rN_0]}{r}$$

Por lo que tomando los primeros N_C coeficientes de $c_s[n]$, con $N_C < N_0$, se pueden separar excitación y tracto vocal.

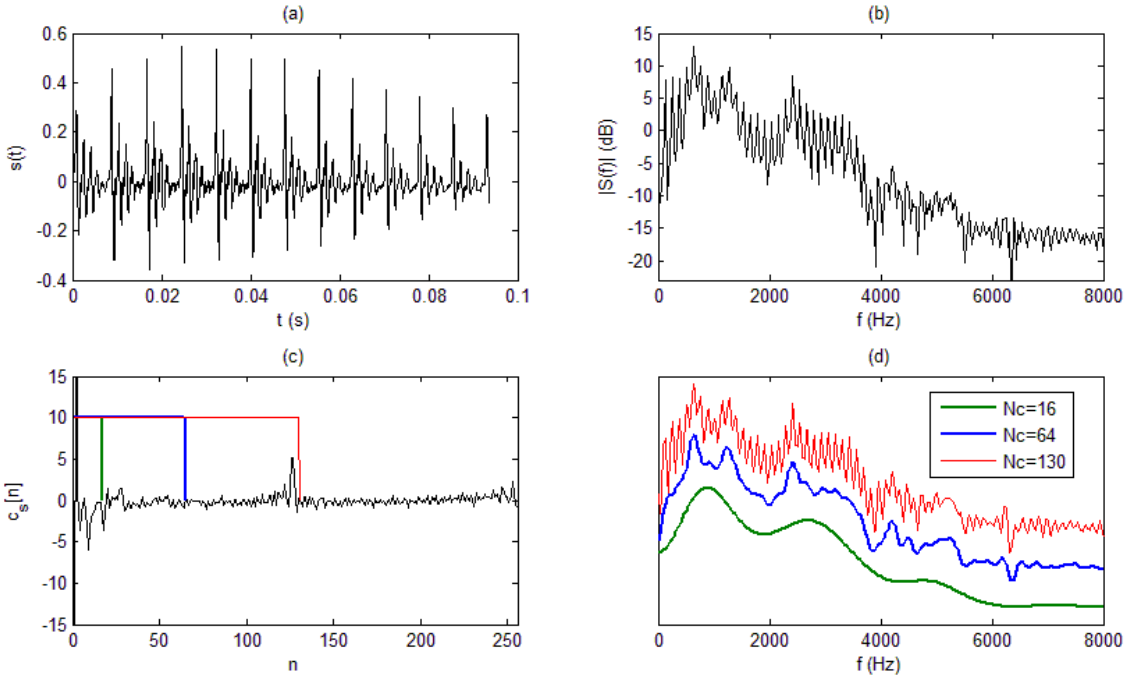

Figura 2.3.1: (a) Grabación de la vocal a. (b) Espectro de la señal. (c) Tranformada cepstrum de la señal. (d) Espectro recuperado con distinto número de coeficeintes cepstrum

En la figura 2.3.1.c se representa la transformada cepstrum con 256 coeficientes de una grabación a 16 kHz de la vocal a. En torno a n = 126 se puede observar el pico debido a la periodicidad de la señal de excitación, para este caso N_0 = 126.

La figura 2.3.1.d muestra el espectro reconstruido a partir de los primeros N_C coeficientes cepstrum. Se observa que al aumentar el valor de N_C de 16 a 64 se obtiene una representación con más detalles de la forma del espectro de la señal. En el caso N_C = 130 se cumple $N_C > N_0$ y por lo tanto se están extrayendo detalles del espectro propios de la señal de excitación.

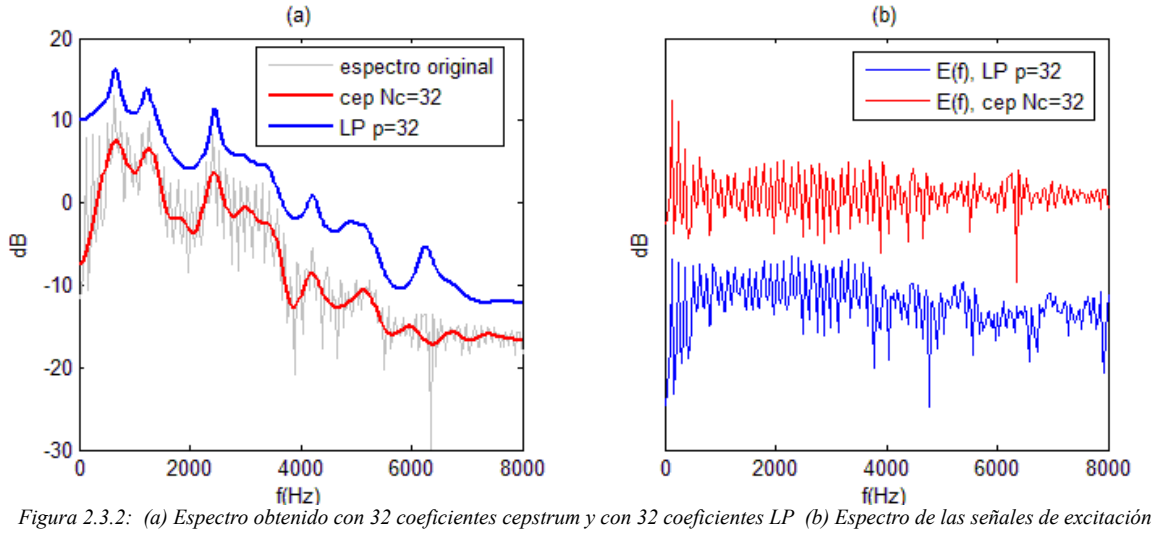

Figura 2.3.2: (a) Espectro obtenido con 32 coeficientes cepstrum y con 32 coeficientes LP (b) Espectro de las señales de excitación

La transformada cepstrum consigue una separación mejor entre el tracto vocal y excitación [8] que las técnicas de predicción lineal. Además los coeficientes cepstrum son una representación más compacta de la información del tracto vocal. En la figura 2.3.2.a puede apreciarse como para un mismo número de coeficientes la transformada cepstrum consigue extraer más detalles.

-Consideraciones prácticas del análisis cepstrum localizado

Al igual que en caso de análisis LP localizado se calculará la transformada cepstrum $c_m[n]$ de los sucesivos fragmentos $s_m[n]$ de la señal $s[n]$ en los instantes de tiempo m, donde m será un múltiplo de *hop* muestras.

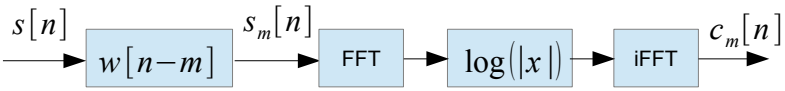

Figura 2.3.3: Esquema del análisis cepstrum localizado

2.4- Sistemas de verificación de locutor

Los sistemas de verificación de locutor pueden dividirse en tres procesos: entrenamiento, registro y test. Durante el entrenamiento y el registro el sistema genera modelos de locutores que después serán usados en la fase de test para tratar de identificar a locutores desconocidos.

La fase de entrenamiento consiste en extraer parámetros de varias señales de voz de diferentes locutores para obtener un modelo estadístico general de locutor UBM (universal background model). Durante la fase de registro el sistema genera modelos para cada locutor en particular. La forma más común de hacerlo es partir del UBM y adaptarlo a cada locutor usando la información de los parámetros extraídos de una grabación de voz. El UBM trata de representar a todos los locutores en general y sirve como un modelo estadístico para los posibles impostores.

Empíricamente se ha demostrado el buen funcionamiento de los modelos de locutor basados en mezclas de gaussianas (GMM). Sea λ el modelo estadístico de un locutor representado por la mezcla de M distribuciones gaussianas. Se define la función de verosimilitud entre λ y un vector de parámetros $\vec{v}$ de dimensión $D\times 1$ como:

$$p(\vec{v}\,|\lambda)=\sum_{i=1}^{M} w_i\, p_i(\vec{v})$$

La expresión anterior es una combinación lineal de M distribuciones de densidad gaussianas unimodales. Cada distribución unimodal esta definida por una vector de medias $\vec{\mu}_i$ de dimensión $D\times 1$ y por una matriz de covarianza Σ_i de dimensión $D\times D$

$$p_i(\vec{v})=\frac{1}{(2\pi)^{D/2}\,|\Sigma_i|^{1/2}}\exp\left(-\frac{1}{2}(\vec{v}-\vec{\mu}_i)^T\Sigma_i^{-1}(\vec{v}-\vec{\mu}_i)\right)$$

Donde para que se cumpla la condición de función de densidad:

$$\sum_{i=1}^{M} w_i=1$$

Por lo tanto el modelo λ del locutor estará definido por:

$$\lambda = \left\{ w_i, \vec{\mu}_i, \Sigma_i \right\}_{i=1}^{M}$$

Dado un conjunto de vectores de entrenamiento los parámetros del modelo pueden estimarse usando el algoritmo iterativo EM (expectation-maximization) [9]. Para adaptar el modelo general UBM a cada locutor se utiliza la adaptación MAP (máximo a posteriori) [10]. En el siguiente esquema se muestra un resumen del proceso de entrenamiento:

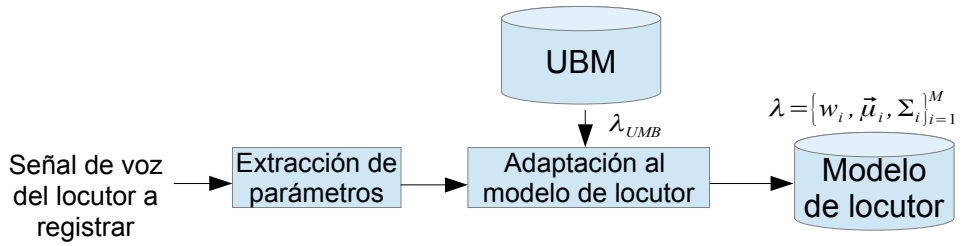

Figura 2.4.1: Esquema del proceso de entrenamiento de un sistema de verificación de locutor.

Durante la fase de test se extraen los parámetros de la señal de voz de test del locutor desconocido y se comparan con el modelo del locutor objetivo y con el modelo UBM. Los resultados con cada modelo se comparan y se genera un estadístico Λ, de modo que cuanto más alto sea el valor de este estadístico, más se parece el locutor desconocido al locutor objetivo. La idea principal de este procedimiento es que si el locutor desconocido es quien dice ser, sus vectores de parámetros se ajustarán bien al modelo del locutor. Si es un impostor, sus parámetros se ajustarán mejor al modelo UBM.

De la grabación de voz del locutor desconocido se extrae un conjunto de vectores $V = \left\{ \vec{v}_t \right\}_{t=1}^{T}$. Puede calcularse la log-verosimilitud entre λ y el conjunto de vectores de test V como:

$$\log p(V|\lambda) = \frac{1}{T} \sum_{t=1}^{T} \log p(\vec{v}_i|\lambda)$$

Sea λ_{UBM} el conjunto de parámetros estadísticos del modelo general UBM. El valor del estadístico Λ puede calcularse comparando la log-verosimilitud de V con los modelos del locutor objetivo y con el modelo UBM:

$$\Lambda(V) = \log p(V|\lambda) - \log p(V|\lambda_{UBM})$$

Para reducir la varianza del estadístico resultante se suele realizar una normalización[11]

$$\widetilde{\Lambda}(V) = \frac{\Lambda(V) - \mu_\lambda}{\sigma_\lambda}$$

Donde μ_λ y σ_λ son los parámetros de normalización del modelo de locutor λ. Finalmente se fija un umbral y se compara con el estadístico normalizado. Si este es mayor que el umbral el sistema decide que el locutor desconocido es quien dice ser. En el siguiente esquema se muestra un resumen del proceso de test:

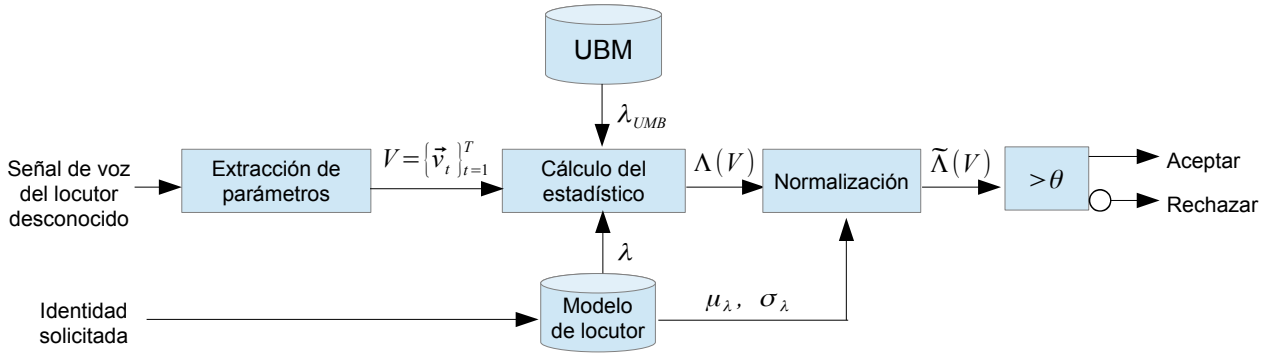

Figura 2.4.2: Esquema del proceso de test de un sistema de verificación de locutor.

Los sistemas de verificación de locutor pueden encontrarse con dos situaciones. La primera es que el locutor desconocido sea quien diga ser y la segunda que sea un impostor. La probabilidad de pérdida P_{loss} se define como la probabilidad de que el sistema rechace a un locutor que sea quien dice ser. Por otro lado la probabilidad de falsa alarma P_{FA} se define como la probabilidad de que el sistema acepte a un impostor.

La figura 2.4.3 muestra las distribuciones de los estadísticos obtenidas por un sistema de verificación de locutor real. Se observa que las distribuciones de los impostores y de los no impostores no están completamente separadas. La elección del umbral por lo tanto generará un compromiso entre los valores de P_{FA} y P_{loss} En la figura 2.4.4 muestra el compromiso entre P_{FA} y P_{loss} a la hora de elegir el umbral.

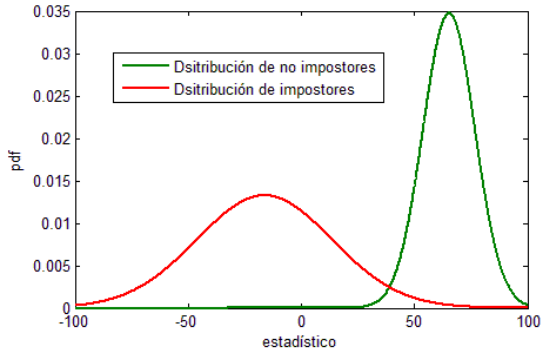

Figura 2.4.3: Distribución de los estadísticos obtenidos por un sistema de verificación de locutor.

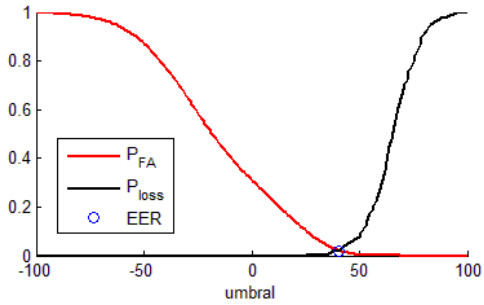

Figura 2.4.4: P_{FA}, P_{loss} vs umbral en un sistema de verificación de locutor.

La curva DET(detection error tradeoff) es una visualización de las prestaciones del sistema. Consiste en representa P_{loss} en función de $P_{FA.}$

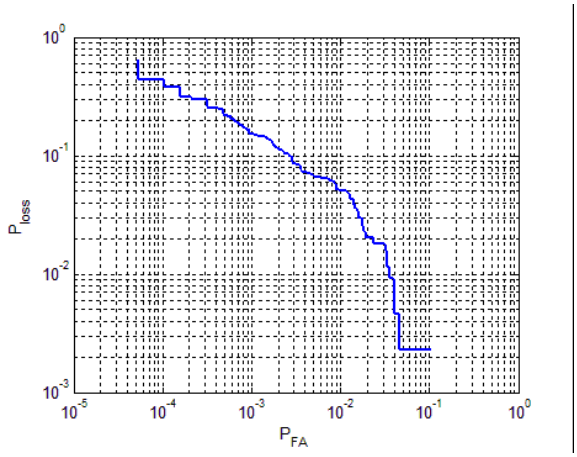

Figura 2.4.4: Curva DET de un sistema de verificación de locutor

Se denomina EER (equal-error-rate) al punto de operación en el que se cumple $P_{FA} = P_{loss}$**.** El valor del EER es una representación simplificada de las prestaciones de un sistema de verificación de locutor.

-Sistemas de verificación de locutor basados en i-vectors

Uno de los principales problemas de los sistemas de verificación es la influencia del canal por el que se transmite la señal de voz. Para resolver este problema se ha propuesto una solución basada en el análisis de factores[12], que conforma lo que hoy en día es el estado del arte en los sistemas de verificación de locutor.

Recientemente los modelos de locutor de total variabilidad basados en i-vectors[13] han ganado importancia debido a su buenos resultados y a su bajo coste computacional. Sea *S* el supervector de medias del modelo GMM de un locutor. *S* será un vector formado por la concatenación de $\{\vec{\mu}_i\}_{i=1}^{M}$ y de dimensión *DM*×1, según el modelo de total variablidad:

$$S = S_{UBM} + T\,\omega_i$$

Donde S_{UBM} es el supervector de medias del modelo UBM. T es la matriz de total variabilidad de dimensión *DM*×*L* con *L* << *DM*. ω_i es un vector aleatorio de distribución normal estándar y dimensión *L*×1 denominado vector intermedio o i-vector. La reducción dimensional hace que los i-vectors sean una representación más eficiente de la información cada locutor. En [12, 13] se pueden encontrar diversos métodos para entrenar la matriz *T*.

La mayoría de los sistemas de verificación de locutor basados en i-vectors calculan su estadístico mediante la distancia coseno:

$$\Lambda(\omega_{1,}\omega_2) = \frac{\langle \omega_{1,}\omega_2 \rangle}{||\omega_1||||\omega_2||}$$

Donde ω_1 es el i-vector obtenido del locutor desconocido y ω_2 es el i-vector obtenido del locutor objetivo.

3- Modificación de la señal de voz

En la sección anterior se han visto algunas técnicas para separar la información del tracto vocal y la información sobre la excitación de la señal de voz. En esta sección se hacen uso de esas técnicas y se proponen sistemas para transformar tracto vocal y la señal de excitación por separado.

3.1- Modificación de la señal de excitación

El phase vocoder es una técnica de procesado digital de la señal que permite realizar escalados temporales y desplazamientos del pitch de gran calidad. Su funcionamiento fue descrito por primera vez[14] en 1966 y debido a su bajo coste computacional y a su buen funcionamiento hoy en día su uso es muy frecuente en aplicaciones musicales.

El phase vocoder usa la información de fase obtenida en el análisis STFT (short-time Fourier tranform) para modificar la amplitud o fase de las componentes frecuenciales de una señal. Después, sobre la señal modificada en el dominio frecuencial, realiza la STFT inversa para obtener su representación en el dominio temporal.

La mayoría de aplicaciones existentes para la modificación del pitch usan técnicas basadas en phase vocoder para realizar un desplazamiento frecuencial sobre la señal de voz. Al no separar las componentes de excitación y de tracto vocal estas aplicaciones también desplazan la posición de los formantes modificando así la información de la forma del tracto vocal. En esta subsección se propone un sistema para la modificación del pitch que actúa solo sobre la señal de excitación con el objetivo de preservar la información del tracto vocal.

3.1.1- Desplazamiento frecuencial del pitch

A continuación se propone un sistema para modificar el valor de pitch de una señal de voz grabada. El objetivo es realizar un desplazamiento frecuencial sobre la señal de excitación tratando de no alterar los parámetros del tracto vocal. El sistema se resume en el siguiente esquema:

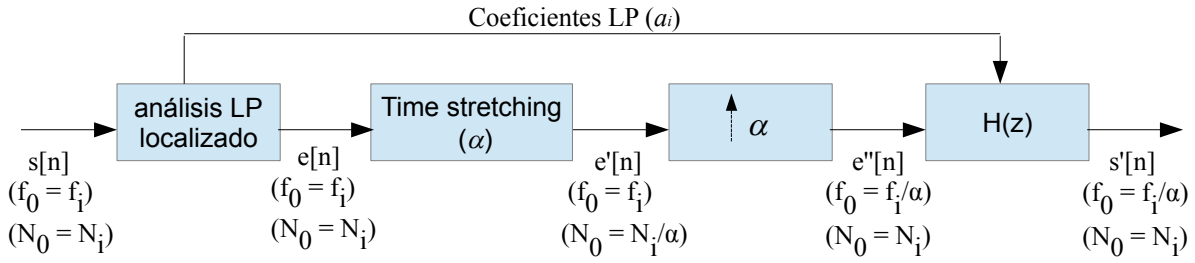

Figura 3.1.1: Esuquema del proceso de modificación del pitch

La señal grabada se divide en segmentos de *N* muestras con un avance temporal entre segmentos de *hop* muestras. Para cada segmento se realiza análisis de predicción lineal obteniendo un segmento de señal de excitación y unos coeficientes LP.

El efecto de desplazamiento frecuencial lo conseguiremos interpolando o diezmando en tiempo la señal de excitación en un factor α. De este modo al interpolar (α >1), se producirá un desplazamiento frecuencial negativo y el valor de pitch se reducirá α veces. Al diezmar (α <1), se producirá un desplazamiento frecuencial positivo y el valor de pitch aumentará α veces. El proceso de diezmado involucra un problema de alisasing, por lo que en esos casos cada fragmento se pasará antes por un filtro antialiasing. Por otro lado, durante el proceso de interpolación, el espectro original se ve reducido y replicado, efecto que debido a la periodicidad de la señal de excitación resulta interesante conservar.

El proceso de interpolación y diezmado a su vez modifica la duración de la señal en un factor α. Puesto que nos interesa conservar la duración original previamente modificaremos la duración de esta en un factor $1/\alpha$ sin alterar su distribución espectral. Esta técnica se conoce como *time stretching* y es ampliamente utilizada en los procesos de modificación de la escala temporal en los sistemas phase vocoder.

Una vez hecho el desplazamiento frecuencial sin haber modificado la duración temporal, cada

segmento de la señal de excitación recuperado es filtrado por sus correspondientes coeficientes LP.

En el sistema anterior quedan ciertos parámetros por definir como:

-Parámetros de los segmentos de señal (duración, tipo de ventana, factor de overlap)

-Orden del análisis LP.

-Técnica de interpolación.

La elección de estos parámetros y su influencia en la señal sintetizada se discuten en la sección de resultados.

3.1.2- Time stretching

Conseguiremos el efecto time stretching sintetizando con un número de muestras de avance temporal entre segmentos diferente del de la fase de análisis. El factor de dilatación temporal α será por lo tanto:

$$\alpha = h_o / h_i$$

Donde h_i y h_o son el número de muestras de avance temporal u overlap entre segmentos en la fase de análisis y síntesis respectivamente. El objetivo es calcular el nuevo incremento de fase asociado a h_o. Dividimos la señal $x[n]$ en segmentos de N muestras con h_i muestras de overlap. Denotamos $x_m[n]$ como el m-ésimo fragmento. Definimos:

$$X_m[k] = DFT_N\{x_m[n]\}$$
$$\varphi_m[k] = \angle X_m[k]$$

El incremento de fase entre los instantes m y m-1 puede expresarse en términos de la frecuencia instantánea ω_m de s_m como:

$$\varphi_m[k] - \varphi_{m-1}[k] = h_i \omega_m[k]$$

El incremento total de fase puede dividirse en un incremento de fase nominal debido al salto temporal entre los instantes *m* y *m*-1 de h_i muestras y un incremento de fase adicional:

$$\varphi_m[k]-\varphi_{m-1}[k]=h_i\,\omega_m[k]=h_i\frac{2\pi k}{N}+\Delta\varphi_m[k]$$

En términos de frecuencia instantánea:

$$\omega_m[k]=\frac{2\pi k}{N}+\frac{\Delta\varphi_m[k]}{h_i}=\frac{2\pi k}{N}+\Delta\omega_m[k]$$

Puesto que en la práctica calcularemos la DFT mediante el algoritmo FFT, los valores de fase instantánea estarán comprendidos entre -π y π. Por lo tanto no podemos calcular el incremento instantáneo de fase directamente. Una solución es calcular el incremento de fase adicional como:

$$\Delta\varphi_m[k]=arg_\pi\left\{\varphi_m[k]-\varphi_{m-1}-h_i\frac{2\pi k}{N}\right\}\quad\text{donde}\quad arg_\pi\{x\}=x-floor(x/\pi)\pi$$

Podemos calcular el incremento de frecuencia instantánea adicional como:

$$\Delta\omega_m[k]=\frac{\Delta\varphi_m[k]}{h_i}$$

Finalmente el incremento de fase entre los instantes *m* y *m*-1de la señal a sintetizar puede expresarse como:

$$\varphi'_m[k]-\varphi'_{m-1}[k]=h_o\left(\frac{2\pi k}{N}+\Delta\omega_m[k]\right)\quad\text{normalmente C.I:}\quad\varphi'_0[k]=\varphi_0[k]$$

El espectro del segmento de señal a sintetizar podrá calcularse como:

$$X'_m[k]=\left|X_m[k]\right|e^{j\varphi'_m[k]}$$

3.2- Modificación del tracto vocal

En esta sección se propone un sistema para modificar los parámetros del tracto vocal de un locutor A y transformarlos en los de un segundo locutor B. Partimos de un conjunto de grabaciones de la misma frase de A y B. En la fase de entrenamiento se procesarán estos pares de grabaciones y se establecerá un código de mapeo entre los parámetros del tracto vocal del locutor A y los del locutor B. La fase de síntesis modificará una nueva grabación del locutor A no vista en la fase de entrenamiento usando este código de mapeo.

3.2.1- Entrenamiento

A continuación se muestra un esquema de la fase de entrenamiento del sistema propuesto:

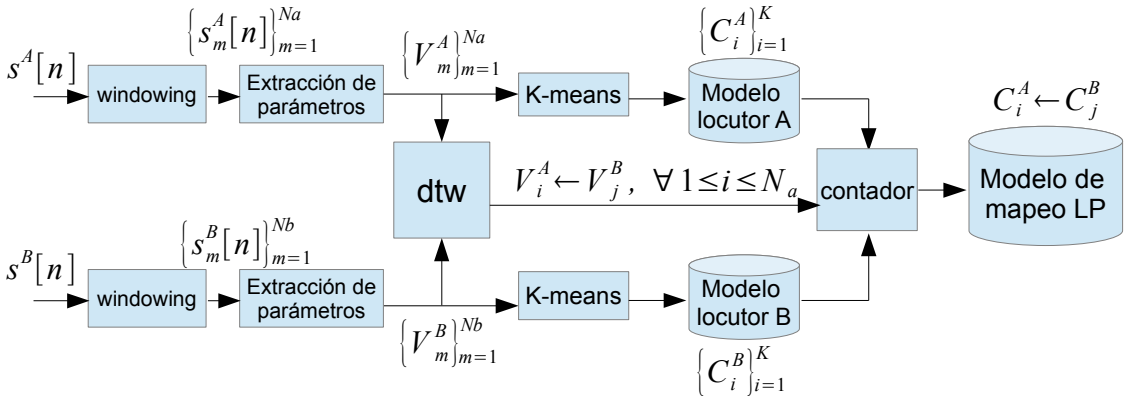

Figura 3.2.1: Esuquema del proceso de entrenamiento para la modificación de los parámetros del tracto vocal

1) Las grabaciones de A y de B se enventanan y se extraen los parámetros del tracto vocal de cada fragmento.

2) Cada vector de parámetros de A se empareja con un vector de B mediante DTW[15] (dynamic time warping) usando una determinada métrica de distancias. El algoritmo DTW permite alinear secuencias de vectores con distinta velocidad de producción.

3) Después se aplica un cuantificador vectorial basado en k-means[16] sobre todos los vectores de A y de B por separado. El resultado de esta operación es una división de todo el

conjunto de vectores de parámetros en diferentes grupos denominados clusters. El vector más representativo de cada cluster se denomina centroide y el conjunto de centroides representará un modelo del tracto vocal de cada locutor.

4) Usando los modelos obtenidos se transforma cada emparejamiento obtenido en el paso 2 en un emparejamiento entre centroides.

5) Se contabilizan todos los emparejamientos. El código de mapeo se establece asignando a cada centroide de A el centroide de B con el que más veces se ha emparejado.

Aplicando el algoritmo k-means conseguimos una representación eficiente de todos los vectores de parámetros obtenidos en la fase de entrenamiento. El objetivo de usar este algoritmo es resumir todas las formas del tracto vocal estimadas durante el entrenamiento.

El algoritmo DTW trata de emparejar vectores de parámetros de A y de B que se corresponden con el mismo sonido de cada frase de entrenamiento. Puesto que tras el algoritmo k-means todos los vectores ya han sido clasificados en un determinado cluster, este emparejamiento puede traducirse inmediatamente en un emparejamiento entre clusters. Si los centroides son una representación eficiente de las posibles formas del tracto vocal, un emparejamiento entre centroides de A y de B será una representación eficiente de todos los emparejamientos obtenidos mediante DTW. La hipótesis es que si el número de muestras de entrenamiento es lo suficientemente grande estos emparejamientos de centroides serán un buen modelo para mapear el tracto vocal del locutor B en el locutor A.

3.2.2- Síntesis

A continuación se muestra un esquema de la fase de síntesis del sistema propuesto:

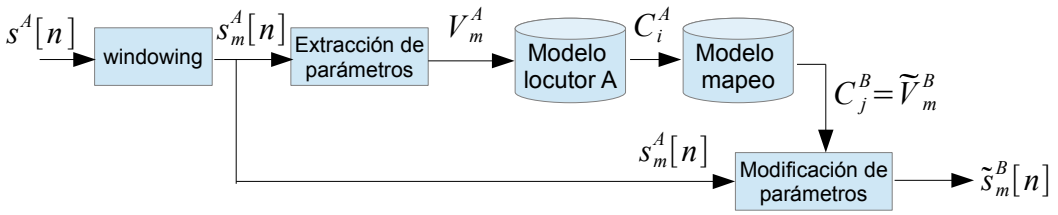

Figura 3.2.2: Esuquema del proceso de síntesis para la modificación de los parámetros del tracto vocal

1) La grabación a modificar del locutor A se enventana y se extraen los parámetros del tracto vocal para cada fragmento.

2) Cada vector de se clasifica usando el modelo obtenido para el locutor A en la fase de entrenamiento y es substituido por el centroide del cluster al que pertenece. Haciendo uso del modelo de mapeo se vuelve a substituir por un centroide del modelo del locutor B.

3) Finalmente cada segmento de señal es modificado por su correspondiente estimación del vector de parámetros del tracto vocal del locutor B.

Se han implementado dos versiones del sistema propuesto usando diferentes parámetros del tracto vocal. Uno utiliza coeficientes LP y otro un determinado número de coeficientes Cespstrum.. A continuación se explican algunos detalles sobre la extracción, clasificación y modificación de estos parámetros para cada versión.

3.2.3- Modificación mediante predicción lineal

Como ya se ha explicado en la primera sección los coeficientes LP serán estimados mediante predicción lineal. El orden de predicción es un parámetro que queda por determinar.

Para poder clasificar cada vector como perteneciente a un cluster se debe decidir una métrica de distancias, en este caso se utiliza la distancia Itakura[17]. En cada iteración del alogritmo k-menas se calculan nuevos centroides haciendo la media aritmética de un conjunto de vectores. La media aritmética de un conjunto de vectores de coeficientes LP no tiene sentido desde un punto de vista de distancia espectral y además la combinación lineal de coeficientes LP puede dar como resultado unos coeficientes inestables. Esta es la razón por lo que los centroides se calculan mediante una modificación del algoritmo k-means[18] que está explicada en el anexo. La idea básica es recalcular los centroides en el dominio de coeficientes de reflexión para poder aplicar la media artimética.

En al fase de síntesis al realizar análisis de predicción lineal también se estimará la señal de excitación de cada fragmento. La modificación consistirá en filtrar cada fragmento de señal de excitación por su

correspondiente estimación de coeficientes LP del locutor B.

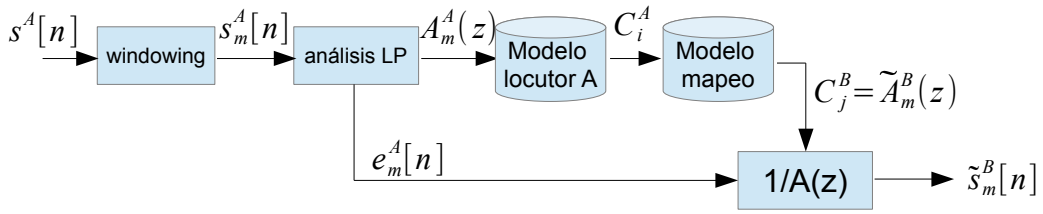

Figura 3.2.3: Esuquema del proceso de síntesis para la modificación del tracto vocal mediante coeficeintes LP

3.2.4- Modificación mediante cepstrum.

Tras el enventanado se obtendrá para cada fragmento de señal $s_m^A[n]$ su STFT (short-time Fourier tranform) $S_m^A[k]$ aplicando el algoritmo FFT con el mismo número de muestras que el fragmento. Para el enventanado se elegirá un número de muestras que sea potencia de dos para que la implementación del algoritmo FFT sea más eficiente. Se calculará el cepstrum real de cada fragmento aplicando la transformada coseno al logaritmo del valor absoluto de las muestras no redundantes de cada STFT :

$$c_m^A[n] = DCT\left\{\log\left|S_m^A[k]\right|\right\}$$

El vector de parámetros del tracto vocal estará formado por N_C coeficientes cepstrum:

$$V_m^A[k] = c_m^A[k],\ \ 1 \le k \le N_C$$

Donde N_C es el número de coeficientes cepstrum que serán utilizados durante la clasificación y la modificación.

Para la clasificación se utilizará la versión original del algoritmo k-means, distancia euclídea como métrica y la media aritmética para calcular los centroides en cada iteración.

La modificación se realizará mediante el cálculo de un nuevo cepstrum $\tilde{c}_m^B[n]$ sustituyendo en el

cepstrum original de cada fragmento del locutor A $c_m^A[n]$ el vector de parámetros estimado del locutor B $\tilde{V}_m^B[k]$:

$$\tilde{c}_m^B[n]=\begin{Bmatrix} \tilde{V}_m^B[n], & 1\le n\le N_C \\ c_m^A[n], & resto \end{Bmatrix}$$

El módulo de la STFT del fragmento modificado $\tilde{S}_m^B[k]$ puede calcularse invirtiendo el cálculo del cepstrum como:

$$\left|\tilde{S}_m^B[k]\right|=e^{DCT^{-1}\left\{\tilde{c}_m^B[n]\right\}}$$

Como fase de $\tilde{S}_m^B[k]$ tomaremos la fase de la STFT del fragmento original:

$$\tilde{S}_m^B[k]=\left|\tilde{S}_m^B[k]\right|\cdot e^{j\angle S_m^A[k]}$$

Finalmente haciendo la FFT inversa podemos obtener el fragmento de señal modificado:

$$\tilde{s}_m^B[n]=FFT^{-1}\left\{\tilde{S}_m^B[k]\right\}$$

El siguiente esquema resume el proceso de síntesis para la modificación del tracto vocal mediante coeficientes cepstrum:

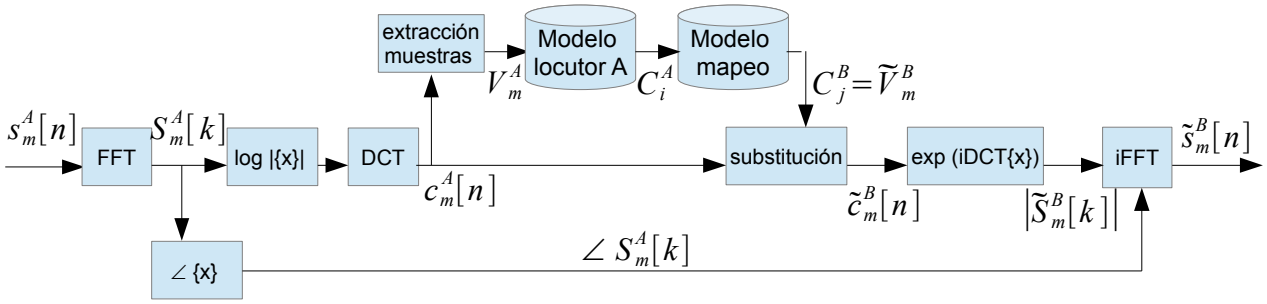

Figura 3.2.4: Esuquema del proceso de síntesis para la modificación del tracto vocal mediante coeficeintes cepstrum

4-Resultados

Los sistemas propuestos en la sección 3 han sido implementados con la herramienta software Matlab. En esta sección se exponen y se analizan las prestaciones de cada sistema por separado.

4.1-Desplazamiento frecuencial del pitch

Se ha implementado el sistema propuesto para modificar el valor de pitch en Matlab. Las primeras versiones utilizaban un interpolador lineal con el que se conseguían resultados de muy baja calidad. Finalmente se decidió usar la función *resample* de Matlab donde se usan filtros polifase para realizar la interpolación.

A continuación se muestran los espectrogramas de una señal de voz y de su excitación antes y después de modificar su valor de pitch. Se ha realizado un aumento en un factor 1.5 sobre una grabación a 16 kHz donde se pronuncia la palabra “Francia” :

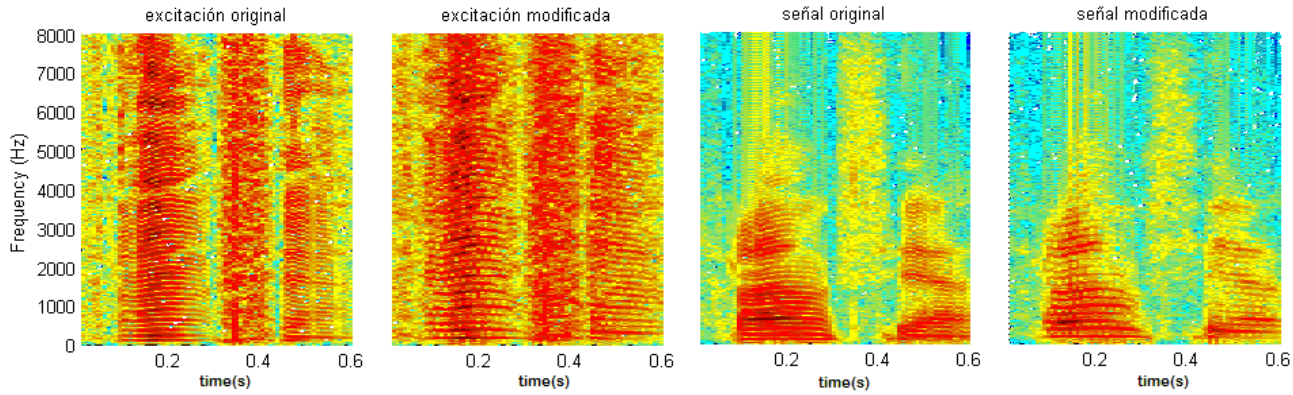

Figura 4.1.1: Espectrogramas de la palabra "Francia" antes y despues de modificaciar el pitch en un factor 1.5

La forma de los espectrogramas reflejan la transformación deseada. En los intervalos de tiempo correspondientes a sonidos sonoros (0.1-0.3s y 0.4-0.6s) se puede ver un aumento en la distancia entre armónicos mientras que la envolvente del espectro de la señal no se ve modificada.

4.1.1- Análisis cualitativo

Se ha realizado un análisis sobre como influyen los parámetros de la función de transformación sobre la calidad de la señal recuperada.

-*N* (número de muestras del fragmento de señal)

Para ventanas pequeñas (<< 20 ms) no hay suficientes muestras para estimar adecuadamente los parámetros del tracto vocal y se nota una gran distorsión en los fonemas sintetizados. A medida que nos acercamos a 20 ms la calidad aumenta. Por encima de los 40 ms no se aprecian errores de fase pero los sonidos se escuchan distorsionados.

-*hop* (número de muestras de avance temporal entre fragmentos consecutivos)

Con valores por encima de *N*/2 la señal recuperada contiene muchos artefactos debido a la pérdida de coherencia de la fase. Conforme disminuimos el valor hasta *N*/4 el salto de fase a estimar disminuye y se aprecia una clara mejora en al calidad de la señal recuperada. A partir de *N*/8 no se aprecia mejora.

-*p* (orden del análisis de predicción)

Valores pequeños de orden de predicción (< 8) generan una estimación de la señal de excitación con mucha información del tracto vocal. En estos casos se está desplazando frecuencialmente parte de la energía espectral de los formantes, generando sonidos más agudos o más graves de lo deseado. Para valores muy grandes de orden de predicción (> 64) la estimación de los parámetros del tracto vocal contiene información de la señal de excitación original. En estos casos al escuchar la señal sintetizada se puede apreciar componentes de la señal original y no se produce el cambio de pitch deseado.

-*H* (Precisión de la interpolación)

A medida que aumentamos el valor de H aumenta la calidad de la señal sintetizada puesto que se realiza una interpolación más precisa. A partir de 100 no se nota ninguna mejora por lo que por defecto se fija ese valor.

-*r* (Factor de escalado del valor de pitch)

Para valores inferiores a 1 se aprecia como la entonación de la señal recuperada es más grave y más aguda para valores superiores a 1. Se ha modificado una grabación con distintos valores de r y se ha estimado el pitch de la señal sintetizada usando la función *fxrapt* del toolbox para Matlab *voicebox*. En el siguiente gráfico se puede observar como se consigue el efecto deseado:

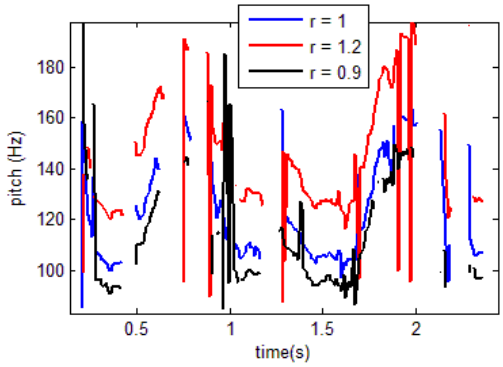

Figura 4.1.2: Estimación del valor de pitch de señales tranformadas

4.1.2- Medida de la distorsión espectral

Se ha realizado un experimento para cuantificar la distorsión que la transformación de pitch introduce sobre los parámetros del tracto vocal. Sobre varias grabaciones se han realizado modificaciones de pitch y se ha calculado una distancia entre la envolvente espectral de la señal original y de la modificada.

Se ha usado la distancia Itakura[18] que permite medir distancias espectrales independientes de la ganancia a partir de los coeficientes LP. A continuación se muestra un esquema del sistema que se ha usado para realizar la medida:

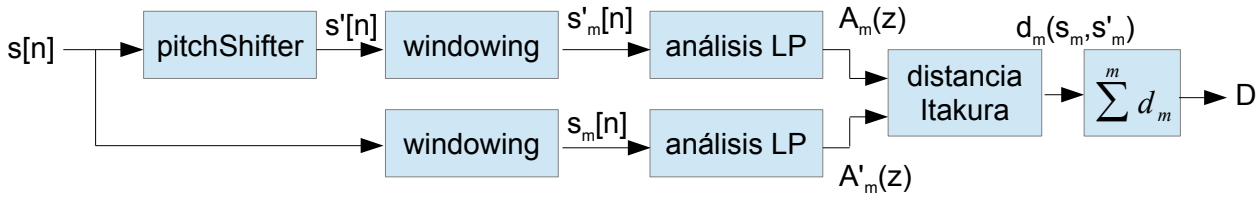

Figura 4.1.3: Esquema del sistema de medida de la distorsión espectral

La señal original y modificada se dividen en fragmentos de 20 ms. Sobre cada segmento se realiza análisis LP de orden 16 para obtener los coeficientes de predicción de los filtros $A_m(z)$ y $A'_m(z)$. Después para cada par de coeficientes de la señal original y modificada se calcula la distancia Itakura y se guarda en memoria. Finalmente se calcula la media de todas las distancias obtenidas para cada fragmento.

Para tener una distancia de referencia se ha realizado la misma medida comparando pares de grabaciones de la misma frase de un mismo locutor. El problema de alineado debido a la diferencia de velocidad del habla entre grabaciones de la misma frase se ha solucionado utilizado DTW[15]. En total se ha medido la diferencia espectral de trece pares de grabaciones de 13 locutores diferentes obteniendo una distancia media de 0.865.

En el experimento se han utilizado 100 grabaciones del corpus Albayzin[19] de las cuales 50 son de hombres y 50 de mujeres. Las grabaciones tienen una frecuencia de muestreo de 16 kHz y las muestras están linealmente cuantificadas con 16 bits. Se han realizado dos conjuntos de medidas para comprobar la influencia del factor de escalado y del orden de predicción en la distorsión espectral. A continuación se muestran los parámetros con los que se han realizado las medidas y los resultados:

	r	*N*	*hop*	*p*	*H*
1	variable	512 (≈20ms)	*N*/4	12	100
2	1.3	512 (≈20ms)	*N*/4	variable	100

Tabla 4.1.1: Parámetros para la medida de la distorsión espectral.r(factor de escalado), N(tamaño de la ventana), hop(avance temporal entre fragmentos), p(orden de predicción), H(precisión de la interpolación).

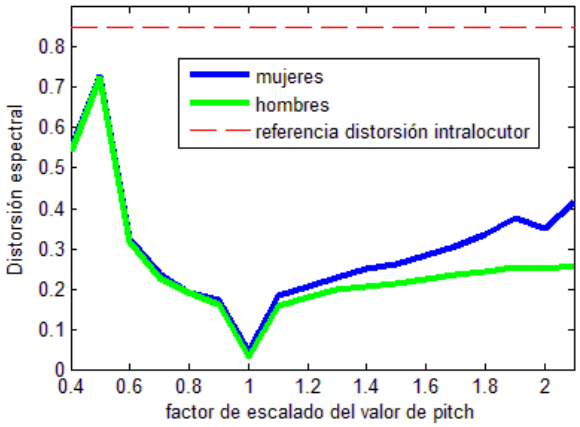

Figura 4.1.4: Distorsión espectral en dunción del factor de escalado del pitch.

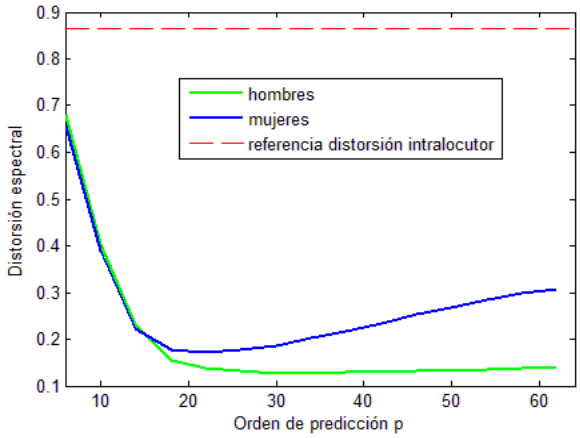

Figura 4.1.5: Distorsión espectral en dunción del factor del orden de predicción

Observamos que la distorsión espectral obtenida en todo el rango de valores de test está por debajo de la referencia. Esto significa que la modificación que introduce el sistema sobre los parámetros del tracto vocal es inferior a la variabilidad temporal intralocutor de los parámetros del tracto vocal.

En la figura 4.1.4 llama la atención que para $r = 1$ la distorsión no es nula. Puesto que en este caso la fase teórica con la que sintetiza es la misma que la fase original, la distorsión puede deberse a errores durante la interpolación.

En la figura 4.1.5 puede observarse que para valores de p pequeños la distorsión es mayor. Esto se debe a que si el orden de predicción no es lo suficientemente grande, la forma del espectro de la señal de excitación estimada contiene componentes del tracto vocal. En esos casos al desplazar en frecuencia la excitación estamos modificando el tracto vocal también.

También se puede observar como la distorsión es superior para mujeres que para hombres cuando el factor de conversión es mayor que uno. Puede deberse a que el valor de pitch de las voces femeninas es superior y por lo tanto su espectro contiene armónicos lo suficientemente separados para que la predicción lineal contenga detalles de la seña de excitación para órdenes bajos de predicción.

4.2-Modificación del tracto vocal

Los sistemas propuestos para la modificación del tracto vocal han sido implementados en Matlab orientados a la transformación de señales con una frecuencia de muestreo de 16 kHz. Se ha usado una ventana Hamming de 512 muestras y un factor de solapamiento del 50%. El resto de parámetros como el número de coeficientes o el número de clusters no han sido fijados y su influencia sobre el resultado se discute en los siguientes apartados.

-Sistema de referencia

Para tener una cota del máximo rendimiento del sistema se ha tomado una medida de referencia. Consiste en comparar la frase a transformar del locutor de A con la misma frase pronunciada por B, alinearlas con DTW[15] y hacer la transformación de parámetros del tracto vocal. A continuación se muestra un esquema del sistema de medida de referencia.

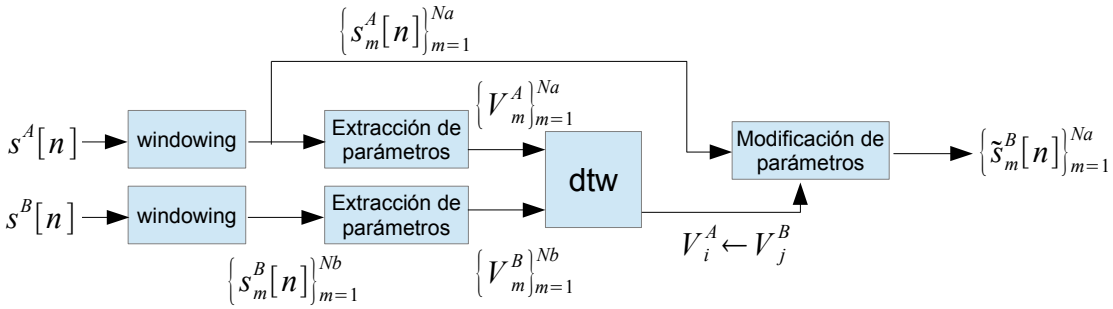

Figura 4.2.1: Esquema del sistema de medida de referencia.

De esta manera no necesitamos generar un modelo de mapeo para estimar los parámetros del tracto vocal del locutor B y el resultado solo depende del método de modificación. Esta medida será por lo tanto una referencia del rendimiento máximo que se puede esperar de los sistemas propuestos para modificar el tracto vocal.

4.2.1- Análisis cualitativo

Se ha realizado un análisis cualitativo sobre cómo influyen los parámetros de la función de transformación sobre la señal recuperada. Se ha prestado atención a la calidad y al parecido entre la voz de la señal sintetizada y del locutor objetivo.

En general, para todas las señales sintetizadas, los sonidos de la grabación original se conservan y se puede entender la frase pronunciada, por lo que parece que el modelo de mapeo funciona correctamente. No obstante se aprecian algunos artefactos en las señales recuperadas debidos a la pérdida de coherencia en la fase. Tanto para el caso de transformación mediante coeficientes LP como en el caso de coeficientes cepstrum, no se ha tenido en cuenta la fase a la hora de sintetizar.

-Tranformación LP

En las señales modificadas mediante LP además se escuchan cambios bruscos durante la pronunciación de un sonido de larga duración. Al disminuir el número de clusters este efecto se ve reducido. Este hecho puede deberse a una mala clasificación, formas del tracto vocal propias de un mismo sonido han sido asignadas a diferentes clusters.

También se observa que a medida que incrementamos el orden de predicción hasta 64 aumenta la sensación de que la señal sintetizada ha sido pronunciada por el locutor objetivo. Para órdenes altos superiores a 128 comienza a distorsionarse mucho la señal sintetizada. Esto se debe a que con un orden de predicción tan alto se están extrayendo detalles del espectro propios de la información de pitch.

-Transformación cepstrum

En el caso de señales modificadas mediante coeficientes cepstrum no se aprecian diferencias significativas al cambiar el número de clusters. Parece que el proceso de clasificación mediante el algoritmo k-means funciona mejor con vectores de coeficientes cepstrum que con vectores de coeficientes LP.

A medida que aumentamos el número de coeficientes cepstrum aumenta la sensación de que la señal sintetizada ha sido pronunciada por el locutor B y disminuye la calidad de la señal. Este hecho se debe

a que al emplear más coeficientes estamos modelando mejor el tracto vocal pero a la hora de sintetizar estamos aumentando la porción del espectro cuya fase estamos estimando.

A partir de un cierto número de coeficientes, aproximadamente 130 para hombres y 80 para mujeres, la señal recuperada se ve muy distorsionada. A partir de ese punto se pierde por completo la sensación de transformación hacia el locutor objetivo y se escucha una voz "metálica". Este fenómeno se debe a que a partir de cierto número de coeficientes estamos mapeando información de pitch. Como se ha visto en la primera sección, la información sobre el pitch en el dominio transformado cepstrum aparece en torno a la muestra que se corresponde con el periodo de pitch. El hecho de que las mujeres tengan una frecuencia de pitch más alta explica que este efecto aparezca antes en mujeres que en hombres. Podemos calcular un teórico valor de pitch a partir de los resultados teniendo en cuenta que en nuestro caso la frecuencia de muestreo F_s es de 16 kHz:

$$\widetilde{f}_0^H = F_s/130 = 123\text{Hz}$$
$$\widetilde{f}_0^M = F_s/80 = 200\text{Hz}$$

Los valores obtenidos encajan entre los valores típicos de pitch para hombres y para mujeres.

4.2.2-Análisis cuantitativo

Una vez hecho un análisis cualitativo del funcionamiento de los sistemas, queremos cuantificar el parecido entre el tracto vocal de la señal sintetizada y el tracto vocal del locutor B. Como medida del parecido del tracto vocal usaremos el estadístico resultante de un sistema de verificación de locutor basado en i-vectors.

Para este experimento se ha utilizado la partición de pruebas del subcorpus fonético de Albayzin[19]. En esta partición hay un total de 2000 grabaciones de 40 locutores diferentes (50 grabaciones por cada locutor). Cada grabación tiene una duración aproximada de 3 s, una frecuencia de muestreo de 16 kHz y muestras linealmente cuantificadas con precisión de 16 bits. En total hay 500 frases diferentes divididas en 10 particiones de tal forma que hay 10 grupos de 4 locutores (2 hombres y dos mujeres) que pronuncian las mismas frases.

Primero se ha entrenado el sistema de verificación de locutor con las 20 primeras frases de cada uno de los 40 locutores. Las 20 frases siguientes se han utilizado para entrenar al sistema de modificación del tracto vocal. Sean $\{h_1\ h_2\ m_1\ m_2\}$ los locutores de un grupo donde h_1 y h_2 son hombres y m_1 y m_2 mujeres. Para cada grupo se han generado los siguientes seis modelos de mapeo:

$$h_1 \leftarrow h_2, \quad h_2 \leftarrow h_1, \quad m_1 \leftarrow m_2, \quad m_2 \leftarrow m_1,$$
$$h_1 \leftarrow m_2, \quad m_2 \leftarrow h_1$$

Las últimas 10 frases de cada locutor han sido tranformadas con los modelos de mapeo anteriores. Finalmente se han recogido todos los estadísticos que el sistema de verificación de locutor ha obtenido con cada una de las frases modificadas y con las 10 ultimas frases de cada locutor sin modificar.

-Análisis del sistema de verificación de locutor

Denotamos S_1 como el estadístico resultante de un locutor tratando de verificar su verdadera identidad y S_0 como el estadístico resultante de un locutor tratando de verificar una identidad falsa. A continuación se muestran las distribuciones de S_1 y S_0 obtenidas a partir de señales sin modificar:

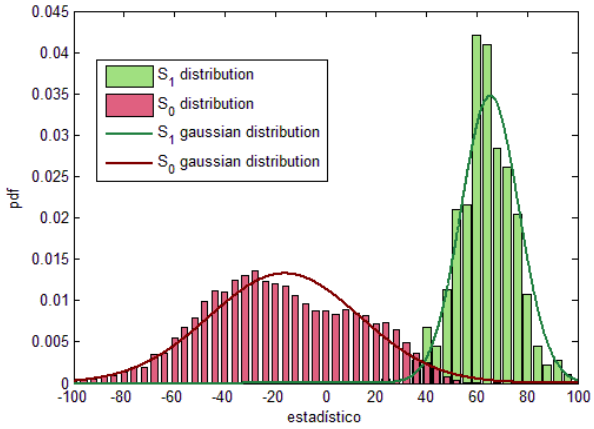

Figura 4.2.2: Sistema de verificación de locutor en ausencia de spoofing

La distribución de los estadísticos se ajusta muy bien a una distribución gaussiana. Para hacer más fácil la interpretación de las gráficas de aquí en adelante se mostrará la curva gaussiana como resumen de la distribución de los estadísticos obtenidos.

Una vez realizada una transformación del locutor origen A al locutor objetivo B, nos interesa obtener el resultado de comparar la señal modificada con el modelo del locutor origen y con el modelo del locutor objetivo. Sea S_a el estadístico resultante de una señal modificada tratando de verificar la identidad del locutor origen y S_b el estadístico resultante de una señal modificada tratando de verificar la identidad del locutor destino. Una transformación perfecta daría como resultados $S_a \approx S_0$ y $S_b \approx S_1$.

-Transformación LP

Se ha usado el sistema de referencia con diferentes órdenes de predicción M con el objetivo de obtener la influencia del número de coeficientes en el resultado. También se han realizado transformaciones con diferente número de clusters K con $p = 64$. A continuación se muestran los resultados:

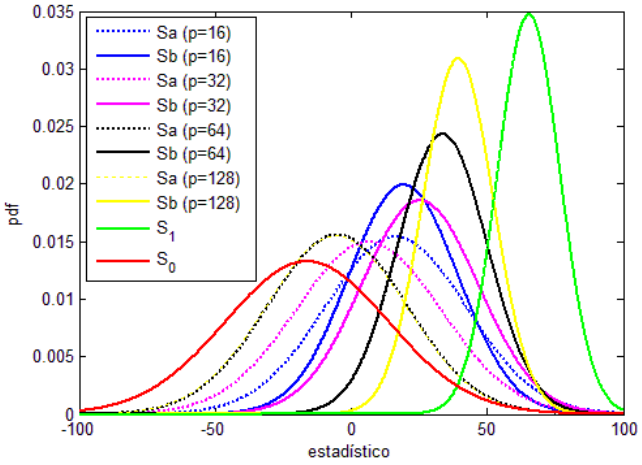

Figura 4.2.5: Referencia spoofing LP para varios órdenes de predicción

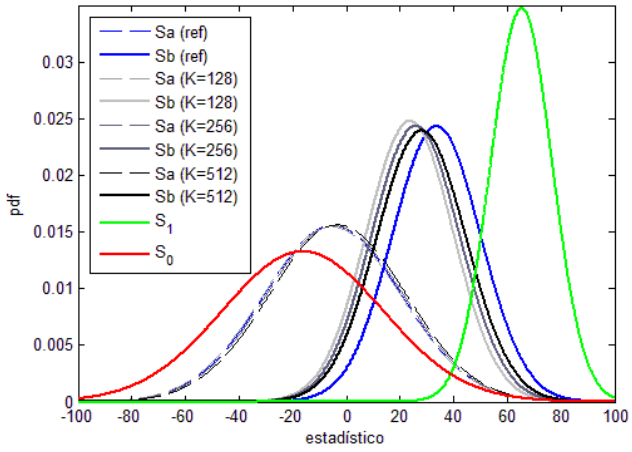

Figura 4.2.6: Spoofing LP con diferente número de clusters

En la figura 4.2.5 se observa que a medida que aumenta p la distribución de S_b se aproxima a la de S_1 y la distribución de S_a se aproxima a la de S_0. Resultado que concuerda con la teoria puesto que un orden superior permite a la predicción lineal extraer más detalles del tracto vocal del locutor origen y destino. La notable diferencia en media entre las distribuciones de S_1 y S_b indica que no se están mapendo todos los parámetros del tracto vocal que usa el verificador de locutor

En la figura 4.2.6 puede apreciarse el correcto funcionamiento de los modelos de mapeo puesto que los resultados obtenidos para diferentes valores de K se parecen mucho a los resultados del sistema de referencia. También se observa que al doblar el valor de K la distribución de S_b se aproxima un poco a la de referencia mientras que la de S_a no cambia.

A la vista de las dos gráficas se saca como conclusión que el orden de predicción tiene más influencia sobre el resultado que el número de clusters. El coste computacional asociado a doblar el tamaño del cluster no compensa la mejora en los resultados.

-Transformación cepstrum

Al igual que en el caso anterior se ha usado el sistema de referencia con diferente número de coeficientes cepstrum N_C. Después se ha fijado N_C = 80 y se han realizado transformaciones con distintos valores de K. A continuación se muestran los resultados:

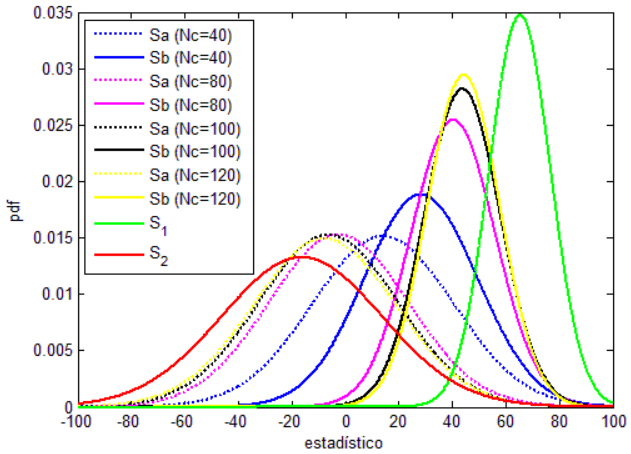

Figura 4.2.7: Referencia spoofing cepstrum para distintos N_C

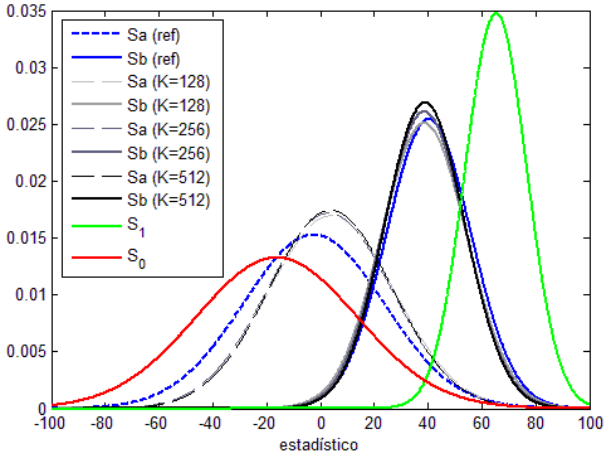

Figura 4.2.8: Spoofing cepstrum para distinto número clusters

En la figura 4.2.7 se observa, al igual que en el caso anterior, que aumentando N_C se consigue que S_a y S_b se separen y se acerquen a S_0 y S_1 respectivamente. Al incrementar N_C se están utilizando coeficientes cepstrum con cada vez menos información del tracto vocal. En la figura 4.2.7 se puede apreciar como la mejora al pasar de 100 a 120 coeficientes es menor que al pasar de 80 a 100.

La figura 4.2.8 muestra que el modelo de mapeo funciona correctamente y que K tiene poca influencia en los resultados. Un tamaño K=128 parece ofrecer las mismas prestaciones que el sistema de referencia. Comparando lo resultados del gráfico 4.2.8 con los de 4.2.6 podemos sacar como conclusión que la clasificación y la generación del modelo de mapeo funciona mejor con coeficientes cepstrum que con coeficientes LP.

A continuación se muestran los resultados de transformaciones LP y cepstrum separando transformaciones entre locutores del mismo sexo y de distinto sexo.

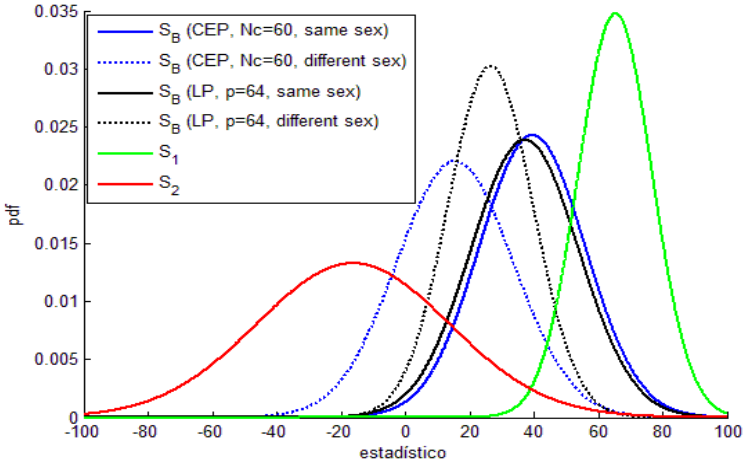

Figura 4.2.9: spoofing entre locutores de igual y distinto sexo

Se puede observar que, tanto para la transformación mediante LP como mediante cepstrum, se obtienen mejores resultados en transformaciones entre locutores del mismo sexo. Una explicación sencilla es que estos casos el tracto vocal del locutor origen es más similar al del locutor objetivo que en los casos contrarios.

Si nos fijamos en las transformaciones entre locutores del mismo sexo se puede observar que la transfomación mediante cepstrum genera mejores resultados que la transformación LP con un menor número de coeficientes. Este resultado pone de manifiesto que los coeficientes cepstrum son una representación más compacta del tracto vocal que los coeficientes LP.

No obstane también se puede apreciar que la transformación LP genera mejores resultados que la transformación cepstrum en transformaciones entre locutores de diferente sexo.

4.2.3- Spoofing

En esta subsección se expone un estudio sobre las implicaciones que tendrían los sistemas de transformación del tracto vocal propuestos a la hora de intentar falsear una identidad. En este escenario el locutor origen de la transformación es un intruso que trata de identificarse como el locutor destino. La probabilidad de engañar al sistema por lo tanto será la probabilidad de falsa alarma obtenida en presencia de spoofing.

-Sistema de verificación de locutor en ausencia de spoofing

Con los resultados del gráfico 4.2.2 para distintos umbrales de decisión se obtienen las siguientes probabilidades de falsa alarma P_{FA} y probabilidades de pérdida P_{loss}:

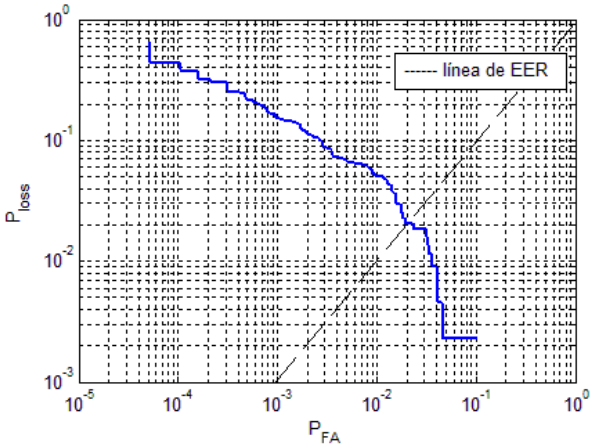

Figura 4.2.10: P_{FA}, P_{loss} vs umbral en ausencia de spoofing

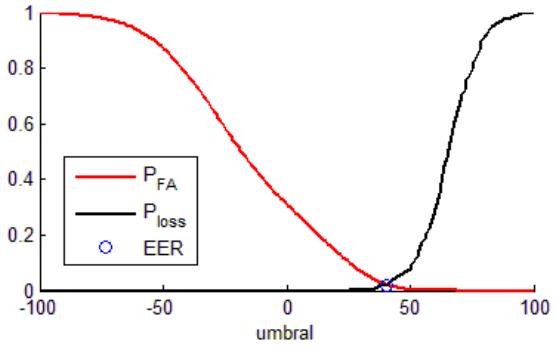

Figura 4.2.11: P_{FA}, P_{loss} vs umbral en ausencia de spoofing

Tomamos el punto EER como punto de operación del sistema es ausencia de spoofing:

umbral	P_{FA}	P_{loss}
40.87	0.02	0.02

Tabla 4.2.1: Punto de operación del verificador de locutor sin spoofing

-Sistema de verificación de locutor en presencia de spoofing

Sustituyendo ahora S_0 por S_b puede obtenerse el comportamiento del sistema ante un ataque de spoofing usando los sistemas de transformación propuestos. En este caso el locutor origen de la transformación

es un intruso que trata de suplantar la identidad del locutor objetivo. Puesto que la finalidad es engañar al verificador solo se han considerado los valores de S_b que se corresponden a transformaciones entre locutores del mismo sexo. Fijando K=256 a continuación se muestra la degradación del sistema para transformaciones con distintos parámetros:

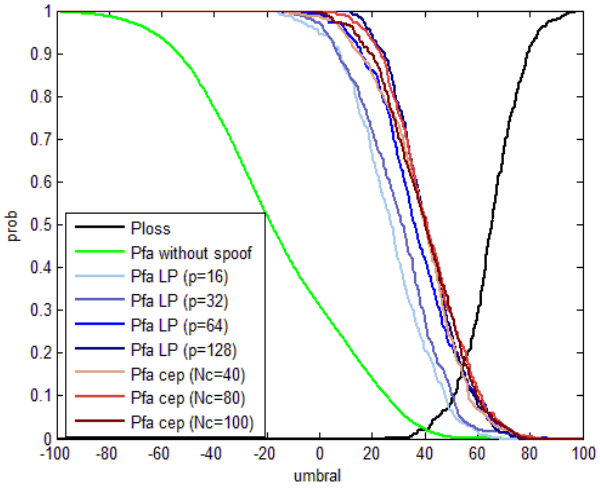

Figura 4.2.12: $P_{FA,}$ P_{loss} vs umbral con spoofing (K=256)

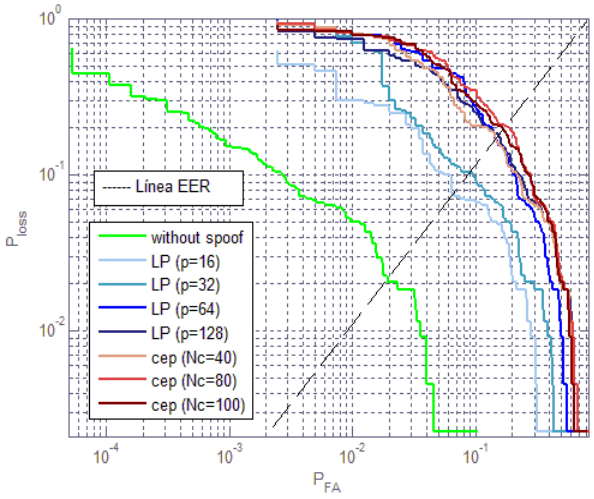

Figura 4.2.13: P_{FA} vs P_{loss} con spoofing (K=256)

En ambos gráficos se observa que el aumento del número de coeficientes tiene más influencia en transformaciones mediante coeficientes LP que mediante cepstrum. Ademas se puede apreciar que la degradación de las prestaciones del verificador con 40 coeficientes cepstrum es mayor que con 64 coeficientes LP. Este hecho pone de manifiesto que los coeficientes cepstrum son una representación más compacta de la forma de tracto vocal.

En la figura 4.2.13 se puede apreciar que asintóticamente se consigue aumentar la probabilidad de falsa alarma más de 10 veces en los casos más favorables.

Para medir la probabilidad de engañar al verificador de locutor antes es necesario elegir un punto de operación. Se han considerado dos:

(I) Umbral = 40.87 y P_{loss} = 0.02 (EER sin spoofing).

(II) EER en presencia de spoofing.

El primer caso respondería a una situación en el que el sistema de verificación desconoce la presencia de señales modificadas y es más susceptible al spoofing. El segundo caso respondería a una situación en el que el sistema de verificación conoce la presencia de spoofing y ha reajustado su umbral de manera óptima.

-*Punto de operación* (I)

En las siguientes tablas se muestran las probabilidades de falsa alarma P_{FA}^{I} obtenidas en el punto de operación (I) :

p	16	16	16	32	32	32	64	64	64	64	128	128
K	ref	128	256	ref	128	256	ref	128	256	512	ref	256
P_{FA}^{I}	*0.20*	0.19	0.20	*0.39*	0.21	0.26	*0.41*	0.21	0.23	0.29	*0.49*	0.39

Tabla 4.2.2: Probabilidades de falsa alarma con spoofing LP en el punto de operación EER sin spoofing

N_C	32	32	64	64	80	80	80	80	100	100	120	120	120	256	256	256
K	ref	256	ref	256	ref	128	256	512	ref	128	ref	128	256	ref	128	256
P_{FA}^{I}	*0.38*	0.38	*0.49*	0.48	*0.52*	0.53	0.51	0.50	*0.60*	0.52	*0.64*	0.54	0.56	*0.80*	0.64	0.64

Tabla 4.2.3: Probabilidades de falsa alarma con spoofing cepstrum en el punto de operación EER sin spoofing

Comparando las dos tablas se observa que para un mismo número de coeficientes se consiguen probabilidades de falsa alarma más altas mediante cepstrum que mediante predicción lineal.

En la tabla 4.2.2 se puede ver que la transformación mediante LP no llega a generar una probabilidad de falsa alarma por encima del 50%. Para todos los órdenes de predicción las probabilidades obtenidas mediante modelo de mapeo están significativamente por debajo de las obtenidas por el sistema de referencia. Este hecho parece indicar que el modelo de mapeo mediante LP no funciona de manera correcta.

En la tabla 4.2.3 se puede ver que con la transformación mediante más de 80 coeficientes cepstrum si que se supera el 50% de probabilidad de falsa alarma. Para valores pequeños de N_c los resultados obtenidos se parecen mucho a los resultados del sistema de referencia. Sin embargo a medida que N_c aumenta, los resultados cada vez se alejan más de la referencia. Parece que los modelos de mapeo funcionan mejor con un número limitado de coeficientes cepstrum (N_C<80).

-*Punto de operación* (II)

En las siguientes tablas se muestran las probabilidades de falsa alarma P_{FA}^{II} obtenidas en el punto de operación (II) :

N_C	16	16	16	32	32	32	64	64	64	64	128	128
K	ref	128	256	ref	128	256	ref	128	256	512	ref	256
$P_{FA}^{II} \equiv EER$	*0.08*	0.07	0.07	*0.14*	0.08	0.09	*0.15*	0.09	0.09	0.11	*0.16*	0.11

Tabla 4.2.4: EER obtenido mediante spoofing LP.

N_C	32	32	64	64	80	80	80	80	100	100	120	120	120	256	256	256
K	ref	256	ref	256	ref	128	256	512	ref	128	ref	128	256	ref	128	256
$P_{FA}^{II} \equiv EER$	*0.12*	0.11	*0.19*	0.18	*0.20*	0.21	0.20	0.17	*0.19*	0.20	*0.19*	0.20	0.19	*0.25*	0.21	0.20

Tabla 4.2.5: EER obtenido mediante spoofing cepstrum.

Comparando los EER obtenidos con spoofing con los de la tabla 4.2.1 en ausencia de spoofing, se puede apreciar que las transformación LP consigue aumentar el EER 5 veces para $p > 32$ mientras que la transformación cepstrum consigue aumentar el EER 10 veces para $N_C > 64$.

En la tabla 4.2.5 se puede ver un comportamiento similar al de la tabla 4.2.3, a medida que N_C aumenta los resultados cada vez se alejan más de la referencia.

-*Sistema cepstrum modificado*

A la vista de los resultados se ha realizado una modificación en la función de transformación mediante cepstrum. En esta nueva versión se clasifica y se genera el modelo de mapeo con 80 coeficientes y se realiza la transformación con 256 coeficientes. En la siguiente tabla se pueden ver los resultados:

K	P_{FA}^{I}	$P_{FA}^{II} \equiv EER$
128	0.67	0.21
255	**0.72**	**0.21**

Tabla 4.2.6: Resultados del sistema modificado.

Esta modificación permite obtener la probabilidad de falsa alarma y el EER más altos.

4.2.4- Tampering

En esta subsección se expone un estudio sobre las implicaciones que tendrían los sistemas de transformación del tracto vocal propuestos a la hora de intentar ocultar una identidad. En este escenario el locutor origen de la transformación está registrado en el sistema de verificación de locutor y tratará de no ser identificado. La probabilidad de engañar al sistema por lo tanto será la probabilidad de pérdida obtenida en presencia de tampering.

-Sistema de verificación de locutor en ausencia de tampering

Las prestaciones del sistema en ausencia de señales transformadas son las mismas que en la subsección anterior:

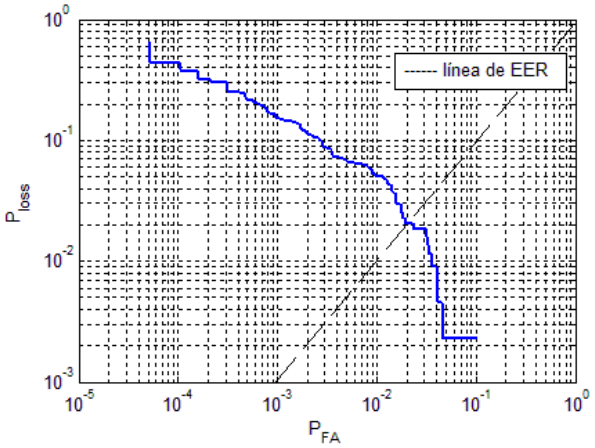

Figura 4.2.14: P_{FA}, P_{loss} vs umbral en ausencia de tampering

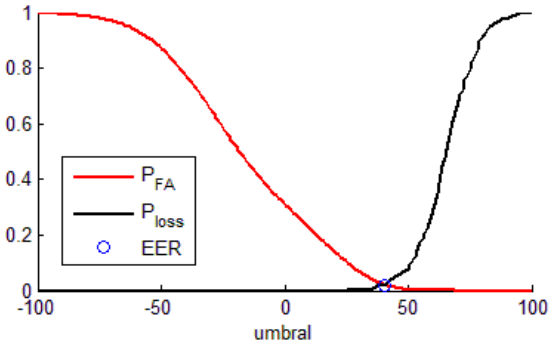

Figura 4.2.15: P_{FA}, P_{loss} vs umbral en ausencia de tampering

Tomamos el punto EER como punto de operación del sistema es ausencia de tampering:

umbral	P_{FA}	P_{loss}
40.87	0.02	0.02

Tabla 4.2.7: Punto de operación del verificador de locutor sin tampering

-Sistema de verificación de locutor en presencia de tampering

Sustituyendo ahora S_1 por S_a puede obtenerse el comportamiento del sistema ante un ataque de tampering usando los sistemas de transformación propuestos. En este caso el locutor origen de la

transformación está registrado en el sistema y tratará de no ser identificado. Para este escenario se han considerado los valores de S_a que se corresponden tanto a transformaciones entre locutores del mismo sexo como del sexo opuesto . Fijando K=256 a continuación se muestra la degradación del sistema para transformaciones con distintos parámetros:

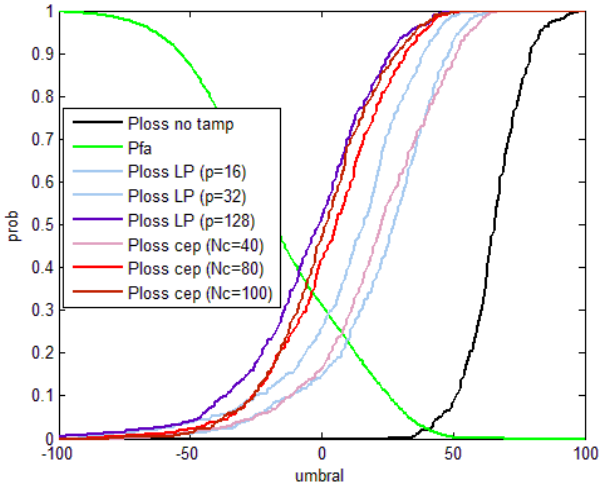

Figura 4.2.16: $P_{FA,}$ P_{loss} vs umbral con tampering (K=256)

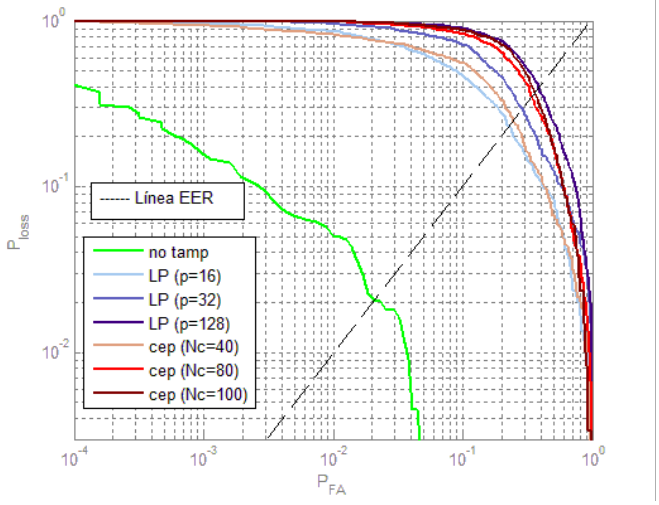

Figura 4.2.17: P_{FA} vs P_{loss} con tampering (K=256)

En los dos gráficos se observa que el aumento del número de coeficientes supone un aumento en la probabilidad de pérdida para ambas transformaciones. En la figura 4.2.17 se puede apreciar que para probabilidades de falsa alarma inferiores a 0.02 se consiguen probabilidades de pérdida muy cercanas a la unidad.

Para medir la probabilidad de ocultar la identidad al verificador de locutor se han elegido dos puntos de operación:

(I) Umbral = 40.87 y P_{FA} = 0.02 (EER sin tampering).

(II) EER en presencia de tampering.

El primer caso respondería a una situación en el que el sistema de verificación desconoce la presencia de tampering. El segundo caso respondería a una situación en el que el sistema de verificación conoce la presencia de tampering y ha reajustado su umbral de manera óptima.

-Punto de operación (I)

En las siguientes tablas se muestran las probabilidades de pérdida P_{loss}^{I} obtenidas en el punto de operación (I) :

p	16	16	16	32	32	32	64	64	64	64	128	128
K	ref	128	256	ref	128	256	ref	128	256	512	ref	256
P_{loss}^{I}	*0.85*	0.77	0.78	*0.95*	0.93	0.92	*0.99*	0.99	0.99	0.99	*0.99*	0.97

Tabla 4.2.8: Probabilidades de falsa alarma con tampering LP en el punto de operación EER sin tampering.

N_C	32	32	64	64	80	80	80	80	100	100	120	120	120	256	256	256
K	ref	256	ref	256	ref	128	256	512	ref	128	ref	128	256	ref	128	256
P_{loss}^{I}	*0.83*	0.74	*0.95*	0.91	*0.98*	0.97	0.96	0.98	*0.98*	0.99	*0.99*	0.98	0.99	*0.98*	0.99	0.99

Tabla 4.2.9: Probabilidades de falsa alarma con tampering cepstrum en el punto de operación EER sin tampering.

En las dos tablas se observa que se consiguen probabilidades de pérdida muy cercanas a la unidad con transformaciones LP con $p > 32$ y con transformaciones cepstrum con $N_C > 80$. Comparando las dos tablas se observa que para un mismo número de coeficientes se consiguen probabilidades de pérdida más altas mediante predicción lineal que mediante cepstrum.

-Punto de operación (II)

En las siguientes tablas se muestran las probabilidades de falsa alarma P_{loss}^{II} obtenidas en el punto de operación (II) :

N_C	16	16	16	32	32	32	64	64	64	64	128	128
K	ref	128	256	ref	128	256	ref	128	256	512	ref	256
$P_{loss}^{II} \equiv EER$	*0.26*	0.23	0.23	*0.33*	0.30	0.29	*0.40*	0.39	0.38	0.39	*0.40*	0.39

Tabla 4.2.10: EER obtenido mediante tampering LP.

N_C	32	32	64	64	80	80	80	80	100	100	120	120	120	256	256	256
K	ref	256	ref	256	ref	128	256	512	ref	128	ref	128	256	ref	128	256
$P_{loss}^{II} \equiv EER$	*0.27*	0.22	*0.35*	0.30	*0.39*	0.34	0.34	0.34	*0.41*	0.38	*0.41*	0.39	0.38	*0.41*	0.39	0.38

Tabla 4.2.11: EER obtenido mediante tampering cepstrum.

Comparando los resultados obtenidos con y sin tampering se puede apreciar que ambas transformaciones consiguen aumentar el EER 20 veces a partir de un cierto numero de coeficientes.

5-Conclusiones y líneas futuras

En este trabajo, a partir de un modelo digital del proceso de producción el habla, se han propuesto e implementado sistemas para modificar la señal de voz.

El sistema para la modificación del pitch consigue señales de buena calidad con determinados parámetros y con factores de escalado cercanos a uno. También se ha visto que el sistema no modifica significativamente los parámetros del tracto vocal. Una posible mejora consistiría en poder cambiar el factor de escalado del pitch a lo largo del tiempo. De este modo se podría elegir un valor de pitch en cada instante de tiempo y poder cambiar la entonación de la frase sintetizada.

A la vista de los resultados obtenidos en la transformación del tracto vocal se puede concluir que hay un compromiso entre la calidad de la señal recuperada y la cantidad de información del tracto vocal que se mapea de un locutor a otro. Las transformaciones con 80 coeficientes cepstrum permiten sintetizar señales con bastante calidad y que generan distribuciones cercanas a las del locutor objetivo. Las distribuciones más cercanas a las del locutor objetivo se han obtenido modificando con 256 coeficientes cepstrum. En modificaciones con tantos coeficientes la señal recuperada no parece natural y cualquier ser humano que la escuchara detectaría la transformación.

Con el sistema para la modificación del tracto vocal se ha simulado un ataque de spoofing y de tampering a un sistema de verificación de locutor del estado del arte y se ha conseguido reducir sus prestaciones en un orden de magnitud. Suponiendo que el verificador opera en el punto de EER en ausencia de señales modificadas, el sistema de transformación propuesto conseguiría suplantar una identidad en el **72%** de los casos y ocultar una identidad en el **99%** de los casos, resultados que ponen de manifiesto la problemática del uso de los sistemas de verificación de locutor en aplicaciones de seguridad.

Siguiendo el esquema básico del sistema, generación del modelo de mapeo y modificación de los parámetros del tracto vocal, una posible continuación del trabajo pasaría por mejorar cada uno de estos dos procesos por separado:

-Mejorar el modelo de mapeo entre locutores:

-Usar modelos ocultos de Markov para tener en cuenta la evolución temporal.

-Incluir derivadas de los parámetros en el proceso de clasificación.

-Usar modelos perceptibles como la escala Mel en el proceso de clasificación.

-Mejorar la modificación de parámetros

-Tener en cuenta la fase a la hora de sintetizar para recuperar señales con más calidad.

-Investigar otras técnicas más sofisticadas del estado del arte.

En este trabajo se ha puesto de manifiesto la problemática del spoofing y del tampering en los sistemas de verificación de locutor. Los artefactos generados con los sistemas de transformación podrían ser estudiados para después ser utilizados en la detección de spoofing y tampering.

6- Referencias

[1] Jesús Villaba, Eduardo Lleida*,"Speaker Verification Performance Degradation against Spoofing and Tampering Attacks".* VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop. 2010

[2] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, "*Spoken Language Processing*" , Chap 6.2 Acustical Model of Speech Production. ISBN 0-13-022616-5.

[3] Atal, B.S. and M.R. Schroeder, *"Predictive Coding of Speech Signals".* Report of the 6th Int. Congress on Acoustics, 1968, Tokyo, Japan.

[4] Anexo: 1-Principio de ortogonalidad y solución del análisis LP. p 48.

[5] Anexo: 2-Recursión de Durbin.p 49.

[6] Anexo: 5-Recursión para calcular los coeficientes de reflexion a partir de los coeficientes LP. p 51.

[7] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, "*Spoken Language Processing*" , Chap 6.4 Cepstral Processing. ISBN 0-13-022616-5.

[8] A.V. Oppenheim, R.W. Schafer, From Frequency to Quefrency: *"A History of the Cepstrum"*. IEEE Signal Processing Magazine, September (2004), pp. 95-99, 106.

[9] A. Dempster, N. Laird, and D. Rubin, *"Maximum likelihood from incomplete data via the EM algorithm,"* Journal of the Royal Statistical Society, vol. 39, no. 1, pp. 1–38, 1977.

[10] Bimbot, Frédéric et al. *"A Tutorial on Text-Independent Speaker Verification."* EURASIP Journal on Advances in Signal Processing 2004.4 (2004): 430-451. Web.

[11] K. P. Li and J. E. Porter, *"Normalizations and selection of speech segments for speaker recognition scoring,"* in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '88), vol. 1, pp. 595–598, New York, NY, USA, April 1988.

[12] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, *"Front end factor analysis for speaker verification".* In print IEEE Trans. Audio, Speech and Language Processing, 2010.

[13] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, , and P Dumouchel, *"A study of inter-speaker variability in speaker verification,"* IEEE Trans. Audio, Speech, and Language Processing, vol. 16, pp. 980–988, 2008.

[14] Flanagan, J.L, Golden, R.M, "Phase vocoder", in Bell System Technical Journal, vol 45, pp 1493-1509. November 1966.

[15] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, "*Spoken Language Processing*" , Chap 8.2.1 Dynamic Programming and DTW. ISBN 0-13-022616-5.

[16] Anexo: 6-Algorimto k-means.pp 51-52.

[17 Anexo: 3-Distancia Itakura. p 50.

[18] Anexo: 7-Algorimto k-means modificado. p 53.

[19] A. Moreno, D.Poch, A.Bonafonte, E.Lleida, J.Llisterri, J.B.Marino, C.Nadeu *"Albayzin speech data base: design of the phonetic corpus"* EUROSPEECH'93, Berlin, 21-23 Sept. 1993, pp.175-8.

ANEXO

1/1

1- Principio de ortogonalidad y solución del análisis LP

Queremos predecir el valor de la señal $s[n]$ a partir de sus p valores anteriores:

$$\tilde{s}[n]=\sum_{k=1}^{p} a_k s[n-k] \quad (1)$$

el error de predicción por lo tanto será:

$$e[n]=s[n]-\sum_{k=1}^{p} a_k s[n-k] \quad (2)$$

Si J es el error cuadrático medio de la señal error:

$$J=E\{e^2[n]\}=E\{(s[n]-\sum_{k=1}^{p} a_k s[n-k])^2\} \quad (3)$$

Derivando respecto a un coeficiente arbitrario a_i e igualando a cero:

$$\frac{\partial J}{\partial a_i}=2E\{e[n]\frac{\partial e[n]}{\partial a_i}\}=2E\{e[n]s[n-i]\}=0 \qquad ,1\leqslant i\leqslant p \quad (4)$$

De la ecuación anterior se deduce que los coeficientes de predicción que minimizan el error cuadrático medio son aquellos que generan un residuo cuyo producto escalar con las p muestras anteriores de $s[n]$ es cero. Esto se conoce como el principio de ortogonalidad.

Substituyendo $e[n]$ por la ecuación (2) en el principio de ortogonalidad obtenemos:

$$E\{s[n-i]s[n]\}=\sum_{k=1}^{p} a_k E\{s[n-i]s[n-k]\} \qquad ,1\leqslant i\leqslant p \quad (5)$$

Si $R_s[k]$ es la función de autocorrelación de $s[n]$ obtenemos las siguientes ecuaciones lineales:

$$R_s[i]=\sum_{k=1}^{p} a_k R_s[i-k] \qquad ,1\leqslant i\leqslant p \quad (6)$$

En formato matricial:

$$\begin{pmatrix} R_s[0] & R_s[1] & \cdots & R_s[p-1] \\ R_s[1] & R_s[0] & \cdots & R_s[p-2] \\ \vdots & \vdots & \ddots & \vdots \\ R_s[p-1] & R_s[p-2] & \cdots & R_s[0] \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}=\begin{pmatrix} R_s[1] \\ R_s[2] \\ \vdots \\ R_s[p] \end{pmatrix} \quad (7)$$

2- Recursión de Durbin

Dado un sistema de ecuaciones cuyos coeficientes dependientes forman una matriz tipo Toepliz como por ejemplo las ecuaciones del análisis LP:

$$\begin{pmatrix} R_s[0] & R_s[1] & \cdots & R_s[p-1] \\ R_s[1] & R_s[0] & \cdots & R_s[p-2] \\ \vdots & \vdots & \ddots & \vdots \\ R_s[p-1] & R_s[p-2] & \cdots & R_s[0] \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} R_s[1] \\ R_s[2] \\ \vdots \\ R_s[p] \end{pmatrix}$$

Los coeficientes a_i pueden calcularse siguiendo la siguiente recursion:

Inicialización:

$$E^0 = R_s[0]$$

Iterar para $i = 1, \ldots, p$:

$$k_i = \left(R_s[i] - \sum_{k=1}^{i-1} a_k^{i-1} R_s[i-k] \right) / E^{i-1}$$

$$a_i^i = k_i$$

$$a_k^i = a_k^{i-1} - k_i a_{i-k}^{i-1} \qquad , 1 \leqslant k < i$$

$$E^i = (1 - k_i^2) E^{i-1}$$

Final:

$$a_k = a_k^p \qquad , 1 \leqslant k \leqslant p$$

Los coeficientes intermedios k_i son los coeficientes de reflexión.

3- Distancia Itakura

Sean A_0 y A_1 los coeficientes de predicción lineal obtendos de los segmentos de señal de voz x_0 y x_1 respectivamente. La distancia Itakura permite medir la separación entre la represetnación espectral de A_0 y A_1.

Sean E_{00} y E_{01} las energías de los errores de predicción obtenidos al filtrar x_0 con los coeficientes A_0 y A_1 respectivamente:

$$E_{00} = E\left(\left|A_0^T x_0\right|^2\right) = A_0^T R_0 A_0$$

$$E_{01} = E\left(\left|A_1^T x_0\right|^2\right) = A_1^T R_0 A_1$$

Donde R_0 es la matriz de autocorrelación de x_0. La distancia Itakura entre A_0 y A_1 se calcula como el cociente entre E_{00} y E_{01}:

$$D(A_0, A_1) = \frac{A_0^T R_0 A_0}{A_1^T R_0 A_1}$$

Puesto que $D(A_0, A_1) \neq D(A_1, A_0)$ normalemnte se calcula la versión simétrica de la distancia Itakura como:

$$D_s(A_0, A_1) = \log \frac{A_0^T R_0 A_0}{A_1^T R_0 A_1} + \log \frac{A_1^T R_1 A_1}{A_0^T R_1 A_0}$$

4- Recursión para calcular los coeficientes LP a partir de los coeficientes de reflexión

Iterar para $i=1, \ldots, p$:

$$a_i^i = k_i$$
$$a_k^i = a_k^{i-1} - k_i a_{i-k}^{i-1} \qquad , 1 \leqslant k < i$$

Finalmente:

$$a_i = a_i^p \qquad , 1 \leqslant i \leqslant p$$

5- Recursión para calcular los coeficientes de reflexion a partir de los coeficientes LP

Inicialización:

$$a_i^p = a_i$$

Iterar para $i = p, \ldots, 1$:

$$k_i = a_i^i$$
$$a_k^{i-1} = \frac{a_k^i + a_i^i a_{i-k}^i}{1 - k_i^2} \qquad , 1 \leqslant k < i$$

6- Algoritmo k-means

Dado un conjunto de vectores de observación $\{X_i\}_{i=1}^N$ el algoritmo iterativo k-menas contruye una partición de las observaviones en *K* conjuntos $\{S_n\}_{n=1}^K$ tratando de minimizar:

$$\sum_{n=1}^{k} \sum_{X_j \in S_n} \| X_j - C_n \|^2$$

donde los vectores $\{C_n\}_{n=1}^K$ se conocen como centroides y son la media de vectores asignados a cada conjunto.

Sea S_n^t el conjunto de observaciones de vecotores pertenecientes al *n*-esimo cluster en la iteración

número t y $\left\{C_n^t\right\}_{n=1}^K$ los K centroides en la iteración numero t.

El comportamiento del algoritmo se describe a continuación:

-Inicialización:

Se toman K vectores al hazar del conjunto $\left\{X_i\right\}_{i=1}^N$ como centroides.

-Iteraración:

Asignación:

$$S_n^t = \left\{ X_i : \left\| X_i - C_n^t \right\| \leq \left\| X_i - C_j^t \right\| \forall\, 1 \leq j \leq K \right\}$$

Actualización:

$$C_n^{t+1} = \frac{1}{card\left(S_n^t\right)} \sum_{X_i \in S_n^t} X_i$$

El algoritmo ha convergido cuando las asiganciones no cambian.

7- Algoritmo k-means modificado

Se define $\{A_i\}_{i=1}^{N}$ como el conjunto de observaciones de vectores de coeficientes LP a clasificar y $\{r_i\}_{i=1}^{N}$ el conjuntio de coeficientes de reflexión asociados.

Sean $\{C_n^t\}_{n=1}^{K}$ los *K* centroides en la iteración numero *t* y $\{R_n^t\}_{n=1}^{K}$ su conjuntio de coeficientes de reflexión asociados.

Sea S_n^t el conjunto de vecotores de coeficientes LP pertenecientes al *n*-esimo cluster en la iteración número *t*.

Se define $d(A, B)$ como la distancia Itakura entre los vectores *A* y *B*.

El comportamiento del algoritmo consiste en:

-Inicialización:

Se calcula $\{r_i\}_{i=1}^{N}$ a partir de $\{A_i\}_{i=1}^{N}$ usando el algoritmo [5] y se toman *K* vectores al hazar del conjunto $\{A_i\}_{i=1}^{N}$ como centroides.

-Iteraración:

Se calculan las distancias entre cada observacion y cada centroide:

$$d(A_i, C_n^t) \forall \begin{matrix} 1 \le i \le N \\ 1 \le n \le K \end{matrix}$$

Asignación

$$S_n^t = \{r_i : d(A_i, C_n^t) \le d(A_i, C_j^t) \forall\, 1 \le j \le K\}$$

Actualización de los centroides mediante la media artimética de los

coeficeintes de reflexión:

$$R_n^{t+1} = \frac{1}{card(S_n^t)} \sum_{r_i \in S_n^t} r_i$$

$$C_n^{t+1} \underset{\leftarrow}{lp} R_n^{t+1}$$

El algoritmo ha convergido cuando las asiganciones no cambian.