TECNIACUSTICA
BARCELONA • 1996

Reconocimiento de locutor independiente de texto (en ambientes ruidosos)

Francisco García López , Marcos Faúndez Zanuy

Escola Universitaria Politécnica De Mataró (EUPMAT).Departamento de Telecomunicaciones y arquitectura de computadores.Avda. Puig i Cadafalch 101-111 08303 MATARÓ
e-mail: frangar@redestb.es // marcos@gps.tsc.upc.es

INTRODUCCIÓN

En los últimos años ha surgido un gran interés por los sistemas automáticos que identifican a las personas por sus voces. El reconocimiento de locutor es un ejemplo de la identificación biométrica personal [1]. Este término se usa para diferenciar técnicas que basan la identificación en ciertas características intrínsecas de la persona (como voz, huellas digitales, estructura genética,...) frente a aquellas que usan objetos artificiales para la identificación (como llaves, tarjetas magnéticas, contraseñas,...). Las aplicaciones básicas están relacionadas con la seguridad, tales como control de accesos físicos, control de accesos a bases de datos, control automático de transacciones por telefóno, validación de tarjetas de crédito,... y aplicaciones técnicas como evaluación de calidad de codificadores de voz,... Existen dos tareas básicas dentro del ámbito de reconocimiento de locutor: identificación y verificación. La identificación de locutor parte de la premisa de que el locutor pertecene a un conjunto de N locutores (N posibles identidades) siendo la misión de la identificación determinar a quién pertenece la voz procedente del locutor desconocido, mientras que en la verificación de locutor, el locutor desconocido debe suministar su identidad (mediante una clave,...) además de su voz, siendo por lo tanto el objetivo decidir si la voz desconocida pertenece o no a un locutor concreto del conjunto. Otra característica importante en el reconocimiento de locutor es la dependencia o no del texto pronunciado por el locutor a reconocer. Los sistemas de texto fijo implican la cooperancia de locutor que desea ser reconocido. Los sistemas independienes de texto son más libres pero menos fiables. Los sistemas que necesitan de una gran fiabilidad deberán usar verificación de locutor con texto fijo, ya que el entorno de trabajo está muy delimitado y se consiguen grandes tasas de reconocimiento. Las últimas tendencias se dirigen hacia el desarrollo de sistemas independientes de texto de alta fiabilidad.

INFORMACIÓN DE LA IDENTIDAD DEL LOCUTOR EN LA SEÑAL DE VOZ.

Existen muchas y diferentes fuentes de información sobre la identidad del locutor, incluyendo las de alto nivel como dialectos, información contextual, estilos del habla (niveles fonéticos, prosódicos y lingüísticos),... estas características sirven de gran ayuda en el reconocimiento de locutores por las personas pero son de poca ayuda para los sistemas automáticos. Las características de bajo nivel, tales como las características de la señal de voz, amplitudes espectrales, frecuencia fundamental, ... son más útiles para los sistemas automáticos de reconocimiento. La voz es una señal acústica que no transporta de forma explícita información anatómica del locutor, lo que la distingue de la identificación personal por huellas digitales, por ejemplo.

Considerando la señal de voz como una consecuencia de articulaciones determinadas por el aparato fonador y por el control neuronal, la fuente de la información del locutor son básicamente las características físicas y estructurales del tracto vocal. Esta información se imparte a la señal de voz durante la articulación además de todas las otras fuentes de información (mensajes lingüísticos,estado emocional, edad, salud,...).

Codificación robusta de la señal de voz en ambientes ruidosos

La señal de voz es el resultado de la convolución de la respuesta impulsional de filtro resonante que representa al tracto vocal h(n) y de la excitación e(n) (pulsos periódicos/ruido) (1). Considerando que $S_e(\omega)$ -espectro de la excitación- es constante (2), H(z) se puede considerar un filtro todo polo, dando lugar a la codificación LPC clásica que realiza la deconvolución entre la señal de voz y la señal de excitación obteniendo las características del filtro del tracto vocal, que es inherente a cada persona. En ambientes ruidosos el modelo todo polos no es váli-

do y por lo tanto, tampoco la codificación LPC. Sabemos que la autocorrelación es más robusta al ruido que la propia señal de voz, ya que si consideramos un ruido blanco y gaussiano, la secuencia de autocorrelación sólo se ve afectada en su primer término R(0) (3), donde R_c es en ausencia de ruido.Es por ello, que se prevee que una codificación (por ejemplo, LPC) que parta del dominio de la autocorrelación será más robusta al ruido que si partimos de la señal de voz. Al ser la señal de voz real su autocorrelación tendrá simetría par, por lo que estudiando su parte causal obtendremos la información de toda ella. Si definimos R'(n) como la parte causal de la autocorrelación (4)[3], su transformada de fourier se corresponde con el espectro complejo (5), donde S(ω) es el espectro (FFT) de R(n) y $S_H(\omega)$ es la transformada de Hilbert de S(ω).Debido a la analogía entre S'(ω) y la señal analítica usada en modulación de amplitud, se puede definir una "envolvente" espectral como E(ω) (6). Esta característica de envolvente, junto con el gran margen dinámico del espectro de la señal de voz, origina que E(ω) enfatize las frecuencias de mayor potencia, que son precisamente las más robustas a un ruido de banda ancha. Como conclusión podemos decir que el cuadrado de la envolvente espectral $E^2(\omega)$, que es además el espectro de R'(n), será más robusto al ruido que el propio espectro. Estas propiedades sugieren que es más factible obtener una representación espectral robusta al ruido aplicando un modelo de predicción lineal a R'(n), que directamente de la señal de voz. Al igual que la técnica LPC estándar asume un modelo todo polo para S(ω), la técnica consistente en aplicar LPC a la parte causal de la autocorrelación ,llamada OSALPC [3], equivale a suponer un modelo, todo polo para $E^2(\omega)$ (8). Se demuestra [3] que B(ω) depende de $S_e(\omega)$ y A(ω), y por lo tanto, no puede ser considerada constante. Esto implica que $E^2(\omega)$ no puede considerarse como un filtro únicamente con polos, sino que también tendrá ceros. Por lo tanto la parametrización OSALPC no realiza una deconvolución entre el filtro y la señal de voz, como la LPC clásica, en cambio, realiza una deconvolución parcial de la señal de voz, que si bien no se corresponde con un modelo todo polos, si que proporciona mejores resultados en presencia de ruido que la LPC clásica.

$$s(n) = h(n)*e(n) \quad (1)$$

$$S(\omega) = Se(\omega) \,/\, |A(\omega)|^2 \quad (2)$$

$$R(m) = \begin{cases} R_c(m) + \eta^2 \ \ para\ m = 0 \\ R_c(m)\ \forall\ m \neq 0 \end{cases} \quad (3)$$

$$R'(m) \begin{cases} R(m) & m > 0 \\ R(0)/2 & m = 0 \\ 0 & m < 0 \end{cases} \quad (4)$$

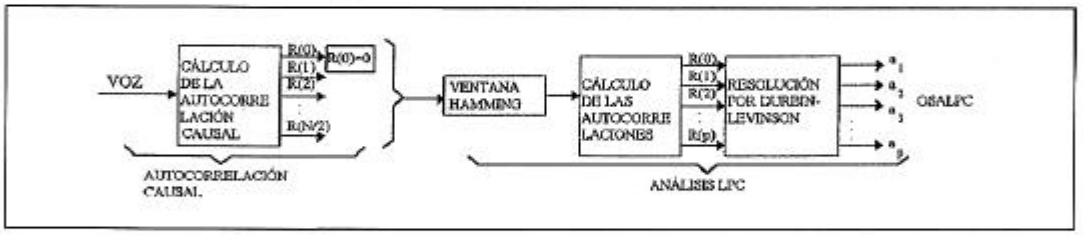
$$S'(\omega) = 1/2[S(\omega) + jS_H(\omega)] \quad (5)$$

$$E(\omega) = |S'(\omega)| \quad (6)$$

$$R'(n) = h(n)*b(n) \quad (7)$$

$$S'(\omega) = B(\omega) \,/\, A(\omega) \Rightarrow$$

$$\Rightarrow E^2(\omega) = |B(\omega)|^2 \,/\, |A(\omega)|^2 \quad (8)$$



El proceso de cálculo de los coeficientes OSALPC [4] es el siguiente:El término R(0) se coloca a 0 ya que es el más susceptible al ruido. Para comprobar los excelentes resultados de la parametrización OSALPC compararemos los espectros de un segmento de la vocal 'a' (8bits / 11KHz) codificado mediante LPC y OSALPC en ausencia y en presencia de ruido con una relación de 10 dB. En las gráficas podemos observar dos hechos:

1) Sin ruido la parametrización LPC es mejor que OSALPC, ya que esta última sólo representa las frecuencias de mayor potencia (debido a la deconvolución parcial)
2) LPC se ve gravemente degradada en presencia de ruido mientras que OSALPC es mucho más robusta y mantiene sus características. Es de uso común el realizar un procesado homomórfico a los coeficientes LPC o OSALPC obteniendo como resultado los coeficientes cepstrales o CEPSTRUM, que simulan la respuesta aproximadamente logarítmica del oído.

MÉTODO DE RECONOCIMIENTO

El resultado de la parametrización es una sucesión de M vectores cepstrales de orden p:$\{X_t\}$ (un vector por cada ventana estudiada). Para obtener un único elemento que represente la identidad del locutor se hallará la matriz de covarianza (9)[5]. La matriz de covarianza , cuyos elementos de la diagonal son la covarianza de las componentes de los vectores cepstrales, nos informa de las capacidades articulatorias del locutor (como por ejemplo, la velocidad media de variación de los parámetros espectrales,...). Una vez definida la matriz como unidad que identificará

a cada locutor, debemos plantearnos como medir la distancia entre la matriz de un locutor de test y las de referencias (que se habrán creado en el proceso de entramiento).

Distancia aritmético-armónica de esfericidad

Sea Y la matriz de test y X la de referencia. Existe una gran familia de medidas que se pueden expresar como una función de los valores propios de $Y \cdot X^{-1}$ (λ_i), cuanto más cercanos estén a la unidad más parecidas son las matrices X e Y. Si llamamos A a la media aritmética de los valores propios y H a la media armónica, podemos definir la distancia aritmético-armónica de esfericidad (10). Si las matrices son iguales la distancia será cero. Para evitar calcular de forma explícita los valores propios, se puede expresar a partir de las trazas (11) siendo más rápido de calcular.

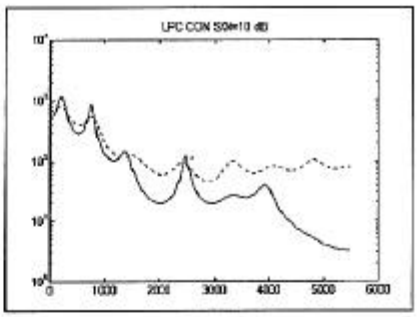


Figura 1: Espectro LPC, en ausencia de ruido (linea continua) y con 10dB (linea discontinua).

FASE DE ENTRENAMIENTO Y DE TEST

Fase de entrenamiento

Cualquier sistema de reconocimiento para poder ser operativo debe tener un grupo de locutores que deseen ser identificados o verificados, ellos configurarán la base de locutores de referencia. Para crear dicha base deberán pronunciar cada locutor un texto de 15'' de duración (aprox), y de él se extraerá la matriz de covarianza que representará a cada locutor. Como para calcular las distancias hará falta su matriz inversa también se realizará en esta fase, guardando para cada locutor la matriz y su inversa.

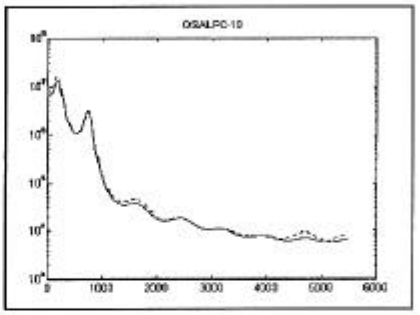


Figura 2: Espectro OSALPC :$E^2(\omega)$, en ausencia de ruido (linea continua) y con 10dB (linea discontinua)

Fase de identificación

Cuando un locutor quiera ser identificado (por lo tanto debe pertener a la base de referencia) deberá pronunciar una frase (de 3'' aprox), a partir de ella se extraerá su matriz de covarianza y su inversa. Luego se calculará la distancia entre el locutor test y todas y cada una de las referencias. Aquella que de distancia mínima será la identidad del locutor.

Fase de verificación

Al ser independiente de texto, las distancias suelen ser muy variables en función de su duración, por lo que es muy difícil elegir un umbral. Para evitarlo se deben normalizar las distancias calculando su verosimilitud [6] según (12), donde d(X,Y) es la distancia entre el test y la referencia correspondiente a la identidad pretendida y el sumatorio se corresponde a las k (13) referencias más cercanas, siendo N el nº total de locutores. Si definimos la tasa de falsa aceptación (FA) como la posibilidad de que un locutor que no perteneciendo al conjunto de locutores-referencia, pretenda acceder y sea aceptado; y la tasa de falso rechazo (FR) a que un locutor perteneciente sea rechazado, el umbral deberá elegirse tal que iguale ambas tasas. En el proceso de test se calculará la verosimilitud y si es menor que el umbral se aceptará, si no se rechazará.

$$X = 1/M \sum_{i=1}^{i=M} X_i \cdot X_i^T \quad (9)$$

$$\mu(X;Y) = \log (A / H) =$$

$$= \log[(\sum_{i=1}^{p} \lambda_i) \cdot (\sum_{i=1}^{p} 1 / \lambda_i)] - 2 \cdot \log (p) \quad (10)$$

$$\mu(X;Y) = \log [(tr(YX^{-1}) \cdot tr(XY^{-1})] - 2 \cdot \log (p) \quad (11)$$

Figura 3. Matriz de covarianza y distancia aritmético armónico de esfericidad

$$L_Y(X) = d(X,Y) / \sum_{j=1}^{\lambda} d(X,Y_j) \quad (12)$$

$$k = \sqrt{N / 2} \quad (13)$$

RESULTADOS

Identificación

Se ha denominado proyecto LOCUS [7] a la implementación de un sistema automático de identificación y verificación de locutor basado en un PC y adquirido mediante una tarjeta de sonido estándar. Todo el software de reconocimiento así como el control de la tarjeta ha sido realizado en C++. Para probar el sistema se ha creado una base de 50 locutores grabados a 16KHz y 16 bits (base LOCUS-16), la grabación se realizó en una única sesión por lo que no se pudo observar el problema de la variabilidad temporal de la voz. Cada locutor grabó 15'' para la fase de entrenamiento y 5 frases de 3'' para realizar los test. Se usó un filtro de preénfasis $1-0.95 \cdot z^{-1}$. Se implementaron las parametrizaciones LPC-Cepstrum y OSALPC-Cepstrum. El proceso de test se dividió en:

a cada locutor, debemos plantearnos como medir la distancia entre la matriz de un locutor de test y las de referencias (que se habrán creado en el proceso de entramiento).

Distancia aritmético-armónica de esfericidad

Sea Y la matriz de test y X la de referencia. Existe una gran familia de medidas que se pueden expresar como una función de los valores propios de $Y \cdot X^{-1}$ (λ_i), cuanto más cercanos estén a la unidad más parecidas son las matrices X e Y. Si llamamos A a la media aritmética de los valores propios y H a la media armónica, podemos definir la distancia aritmético-armónica de esfericidad (10). Si las matrices son iguales la distancia será cero. Para evitar calcular de forma explícita los valores propios, se puede expresar a partir de las trazas (11) siendo más rápido de calcular.

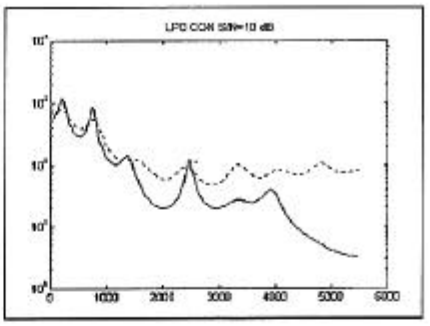


Figura 1: Espectro LPC, en ausencia de ruido (linea continua) y con 10dB (linea discontinua).

FASE DE ENTRENAMIENTO Y DE TEST

Fase de entrenamiento

Cualquier sistema de reconocimiento para poder ser operativo debe tener un grupo de locutores que deseen ser identificados o verificados, ellos configurarán la base de locutores de referencia. Para crear dicha base deberán pronunciar cada locutor un texto de 15'' de duración (aprox), y de él se extraerá la matriz de covarianza que representará a cada locutor. Como para calcular las distancias hará falta su matriz inversa también se realizará en esta fase, guardando para cada locutor la matriz y su inversa.

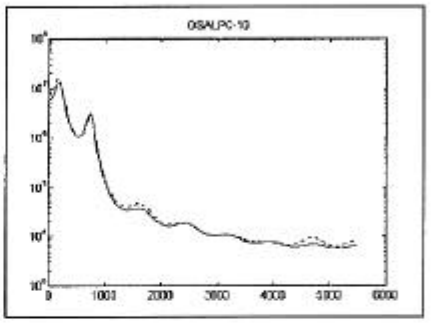


Figura 2: Espectro OSALPC :$E^2(\omega)$, en ausencia de ruido (linea continua) y con 10dB (linea discontinua)

Fase de identificación

Cuando un locutor quiera ser identificado (por lo tanto debe pertener a la base de referencia) deberá pronunciar una frase (de 3'' aprox), a partir de ella se extraerá su matriz de covarianza y su inversa. Luego se calculará la distancia entre el locutor test y todas y cada una de las referencias. Aquella que de distancia mínima será la identidad del locutor.

Fase de verificación

Al ser independiente de texto, las distancias suelen ser muy variables en función de su duración, por lo que es muy difícil elegir un umbral. Para evitarlo se deben normalizar las distancias calculando su verosimilitud **[6]** según (12), donde d(X,Y) es la distancia entre el test y la referencia correspondiente a la identidad pretendida y el sumatorio se corresponde a las k (13) referencias más cercanas, siendo N el nº total de locutores. Si definimos la tasa de falsa aceptación (FA) como la posibilidad de que un locutor que no pertenenciendo al conjunto de locutores-referencia, pretenda acceder y sea aceptado; y la tasa de falso rechazo (FR) a que un locutor perteneciente sea rechazado, el umbral deberá elegirse tal que iguale ambas tasas. En el proceso de test se calculará la verosimilitud y si es menor que el umbral se aceptará, si no se rechazará.

$$X = I/M \sum_{i=1}^{i=M} X_i \cdot X_i^T \quad (9)$$

$$\mu(X;Y) = \log (A / H) =$$

$$= \log[(\sum_{i=1}^{p} \lambda_i) \cdot (\sum_{i=1}^{p} 1 / \lambda_i)] - 2 \cdot \log (p) \quad (10)$$

$$\mu(X;Y) = \log [(tr(YX^{-1}) \cdot tr(XY^{-1})] - 2 \cdot \log (p) \quad (11)$$

Figura 3. Matriz de covarianza y distancia aritmético armónico de esfericidad

$$L_Y(X) = d(X,Y) / \sum_{j=1}^{k} d(X,Y_j) \quad (12)$$

$$k = \sqrt{N / 2} \quad (13)$$

RESULTADOS

Identificación

Se ha denominado proyecto LOCUS [7] a la implementación de un sistema automático de identificación y verificación de locutor basado en un PC y adquirido mediante una tarjeta de sonido estándar. Todo el software de reconocimiento así como el control de la tarjeta ha sido realizado en C++. Para probar el sistema se ha creado una base de 50 locutores grabados a 16KHz y 16 bits (base LOCUS-16), la grabación se realizó en una única sesión por lo que no se pudo observar el problema de la variabilidad temporal de la voz. Cada locutor grabó 15'' para la fase de entrenamiento y 5 frases de 3'' para realizar los test. Se usó un filtro de preénfasis $1\text{-}0.95 \cdot z^{-1}$. Se implementaron las parametrizaciones LPC-Cepstrum y OSALPC-Cepstrum. El proceso de test se dividió en: