Verificación de hablante basado en Dynamic Time Warping

Lácides Antonio Ripoll Solano*

Resumen

Este proyecto desarrolla un sistema de verificación de hablante, texto dependiente basado en alineamiento dinámico del tiempo (Dtw, del inglés Dynamic Time Warping), utilizando el espectro de la señal de voz como elemento de juicio. Al final del proyecto se realizan unas pruebas para determinar los parámetros óptimos para el funciónamiento del Dtw en la verificación de habla continua en el idioma castellano.
Palabras claves: Dynamic Time Warping (DTW), lenguaje.

Abstract

This project develops a text-based speaker verification system based upon Dynamic Time Warping (DTW) by using the voice signal spectrum as the judgement element. At the final stage of the project, some tests are carried out for determining the optimum parameters for DTW functioning when verifying the continuos Spanish speech.
Key words: Dynamic Time Warping (DTW), speech.

1. Introducción

La identificación de las personas ha sido una necesidad desde tiempos remotos, pues les permite la individualizacion y la afirmación de su personalidad. En los ultimos años la identificación personal ha tomado mayor importancia por la necesidad de seguridad en diferentes sítuaciones: Acceso a recintos, transacciones bancarias y especialmente por la concentración de información en redes de computadores.

La verificación automática de hablantes es el proceso por el cual una máquina acepta o rechaza la identidad que aduce una persona, cuando lo único que se posee para confirmarla es su señal de voz. La Fig. 1 muestra el dia-

* Magíster en Ingeniería Eléctrica de la Universidad de los Andes. Ingeniero Eléctrico de la Universidad del Norte, docente y Director (e) del Programa de Ingeniería Electrónica de esa misma institución *(Dirección: lripoll@guayacan.uninorte.edu.co)*

grama de bloques general de un sistema de verificación de hablante.[1]

La operación de la mayoría de los sistemas de verificación de hablantes consta de dos fases. En la fase de entrenamiento, el sistema adquiere la señal de voz del usuario, la analiza y extrae de ella los parámetros que le permiten caracterizarlo. Dichos parámetros determinan un patrón asociado con la identidad de la persona y es almacenado para ser utilizado luego en la verificación. En la fase de reconocimiento, el sistema adquiere una señal de voz que analiza y de la cual extrae los parámetros que compara con el patrón de referencia. De la comparación se obtiene un valor que si es inferior a un cierto umbral ya predeterminado, el hablante es aceptado; en caso contrario es rechazado.

El sistema de verificación que se implementó es un texto dependiente,

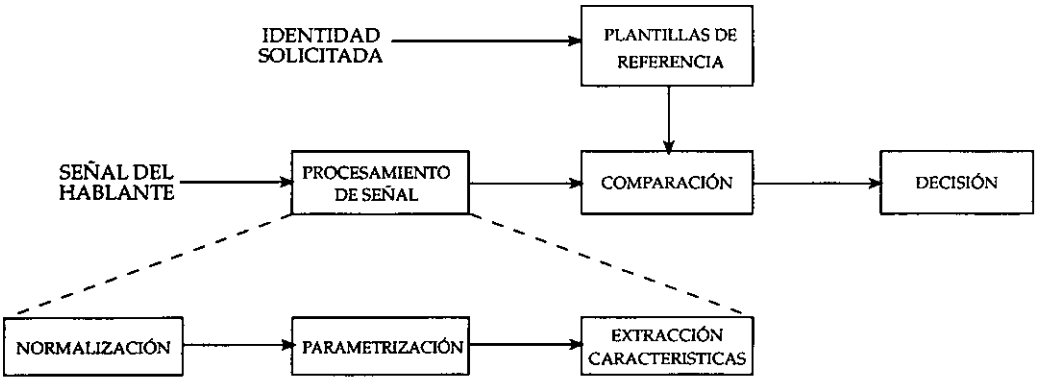


Fig. 1. Diagrama en bloques de un sistema general de verificación de hablante

en el cual se necesita que las palabras utilizadas en el reconocimiento sean las mismas que se utilizaron en el entrenamiento del sistema. El conjunto de características de la voz que fueron utilizadas como elementos de juicio son las contenidas en el espectro de la señal de voz. Teniendo en cuenta que el tracto vocal determina el contenido espectral de los sonidos, se han desarrollado técnicas (Codificación predictiva lineal, LPC) para estimar los parámetros característicos de éste. Para la generación y comparación de patrones se utilizó Alineamiento dinámico de los ejes de tiempo de los dos patrones (DTW, del inglés *Dynamic Time Warping)*, con

[1] SAVIC, Michael y GUPTA, Sunil. *Variable parameter speaker verification system based on hidden markov modeling.*

el propósito de que se compensaran de manera óptima las diferencias en las velocidades de pronunciación de la señal de prueba y las señales de entrenamiento.

Teniendo en cuenta el tipo de decisión que el sistema de verificación debe tomar (identidad verdadera o falsa), existen dos tipos posibles de error: Falso rechazo y falsa aceptación.

mejor desempeño posible del sistema, es decir, mínima probabilidad de error y ejecución suficientemente rápida. [2]

2.1. *Selección de las características*

Las características que se utilizaron son las contenidas en el espectro de la señal de voz. Las características espectrales de un sonido dependen del sistema que lo produce. En el caso de la voz humana,

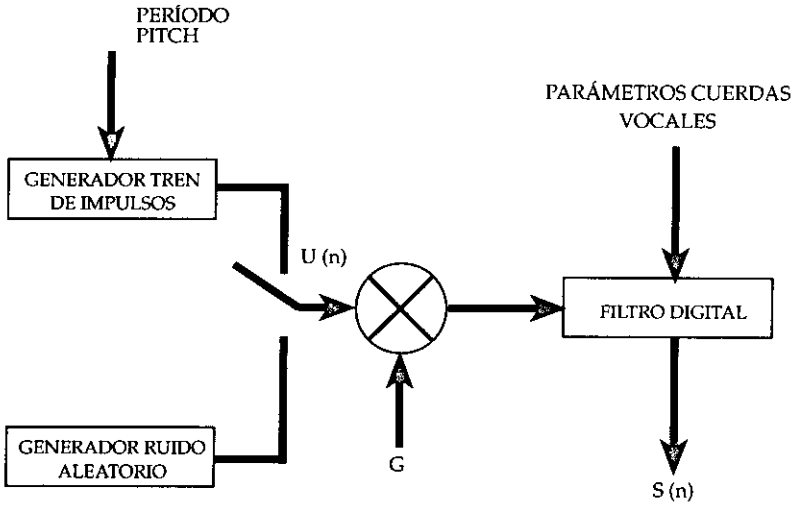


Figura 2. Modelo simplificado de la producción de voz

2. Descripción del sistema

Para solucionar el problema de verificación de identidad por voz se consideraron tres factores. El primero de ellos fue el conjunto de características de la voz que iban a ser utilizadas como elementos de juicio; el segundo, la forma en que dichas características serían extraídas, y el tercero, la técnica que se utilizaría sobre ellas para garantizar el

dicho sistema puede modelarse así [3]:

En este modelo el bloque V(z) representa la función de transferencia (en tiempo discreto) del tracto vocal hu-

[2] BELLO, Marco y GARCÍA, Pablo. "Caracterización de la voz, para verificación automática de hablantes." Tesis Universidad Javeriana.

[3] RABINER, Lawrence y SCHAFER, Ronald. *Digital processing of speech signals.* Prentice-Hall, 1978.

mano; en cuya entrada se ubica el sistema excitador, constituido por los pulmones y las cuerdas vocales, y es modelado por dos bloques distintos, dependiendo del fonema que se intenta producir (sonoro o no sonoro). El sistema excitador determina características como la sonoridad, tono fundamental o intensidad de la voz. El tracto vocal, a su vez, determina el contenido espectral de los sonidos, es decir, los fonemas que se pronuncian y el timbre particular de éstos. Se entiende, por lo tanto, que el tracto vocal es un sistema variable en el tiempo, cuyos parámetros dependen exclusivamente de dos factores:

- El fonema que se produce, pues éste determina la disposición de los órganos que constituyen el tracto vocal (lengua, paladar, velo del paladar, dientes, labios, etc.).

- Las dimensiones específicas del tracto vocal (largo, área transversal, etc.) dependen de la constitución física de la persona.

Tanto en la disposición como en las dimensiones particulares del tracto vocal existe información asociada exclusivamente al hablante. Por esta razón, y gracias a que se han desarrollado técnicas que permiten estimar dichos parámetros a partir de la señal acústica, se prefiere utilizar el bloque V(z) como el elemento básico de la voz que caracteriza la identidad.[4]

[4] BELLO, Marco y GARCÍA, Pablo, *op. cit.*

2.2. ***Procesamiento de la señal de voz***

La técnica que se utilizó para estimar los parámetros característicos del trato vocal se denomina Codificacion predictiva lineal (LPC), y es equivalente a un modelo Auto-regresivo (AR) que utiliza la función de auto correlación de la señal para generar una función de transferencia racional que sólo tiene polos en el plano Z.

$$V(z) = \frac{\theta}{1 + \sum_{k=1}^{N} a_k z^{-k}}$$

Los coeficientes a_k se denominan coeficientes de predicción lineal o coeficientes LPC y se estiman de la función de autocorrelación r(n), de la señal de voz s(n), resolviendo el siguiente sistema de ecuaciones:

$$\sum_{k=1}^{p} a_k \, Rn(|i-k|) = Rn(k)$$

donde la función de autocorrelación es

$$\sum_{m=0}^{N-1-k} Sn(m)Sn(m+k) = Rn(k)$$

y $Sn(m) = S(m+n)W(m)$

con Sn(m) un segmento de la señal y W(m) una ventana de longitud finita $0 \leq m \leq N-1$

Debido a las características de variación de la señal de voz, los coeficientes

del predictor deben ser estimados sobre segmentos cortos de la señal, donde se puede considerar a ésta como estacionaria. Típicamente este tiempo se encuentra entre 20 y 30 min.

Aunque los parámetros LPC permiten representar la señal de voz, contienen información referente tanto al texto como al hablante. Por ello, con el fin de extraer la información del hablante, puede ser necesario hacer un proceso de selección de estos parámetros. O mejor aun, realizar algun tipo de transformación sobre ellos, que permita obtener nuevas variables que caractericen la voz y no el texto. Una de dichas transformaciones se denomina cepstrum, y es de gran popularidad en sistemas de verificación de hablantes[5].

2.2.1. *Coeficientes cepstrales*

Los coeficientes cepstrales representan una transformación de la señal de salida de un filtro AR con dos propiedades[6]:

1. Lo representativo de la señal de entrada está separado de lo representativo del filtro.
2. Los cepstrales son una combinación lineal de lo representativo de la señal de entrada y lo representativo del filtro.

Uno de los métodos para calcular los cepstrales parte de los coeficientes LPC. La recursión está dada por[7]:

$$\gamma(n,m) = \log \Theta \qquad \text{para } n=0$$
$$\gamma(n,m) = -a(n,m) + \Sigma(k/n)\, \gamma o(k,m) a(n-k,m) \qquad \text{para } n>0$$

en donde a(n,m) es cero para n>M, donde M corresponde al orden del modelo LPC.

La forma de medir la similitud entre dos tramas es calculando la distancia euclidiana entre los vectores cepstrales asociados a cada una, es decir,

$$d[c1(m),c2(m)] = \{[c_1(m)-c_2(m)]^T[c_1(m)-c_2(m)]\}^{1/2}$$

En la realidad, los procesos implicados para extraer los parámetros cepstrales son[8]:

- *Preénfasis:* Se realiza para acentuar las altas frecuencias en la señal de voz.
- Segmentación en tramas del mismo tamaño y de duración pequeña (≈50 milisegundos) para que representen un solo fonema.
- *Ponderación:* Cada trama se multiplica por una ventana; típicamente una ventana Hamming.

---

[5] *Ibid.*

[6] DELLER, JHON; PROAKIS, John y HANSEN, Jhon. *Discrete-time processing of speech signals.* Prentice-Hall, 1987.

[7] RABINER, Lawrence y SCHAFER, Ronald, *op. cit.*

[8] BEDOYA, Mauricio y ROSELLO Alberto. "Evaluación del método dtw para reconocimiento de habla continua." Tesis Universidad Javeriana.

- *Análisis de autocorrelación:* Se calculan los primeros p+1 coeficientes de autocorrelación para todas las tramas.

- *Análisis LPC-Cepstral:* A partir de los valores de autocorrelación se calcula para cada trama un vector de coeficientes LPC. Se sigue con la extracción de los coeficientes cepstrales, a partir de los coeficientes LPC y las autocorrelaciones.

2.3. *Generación, comparación y reconocimiento de patrones*

Para la generación y el reconocimiento de patrones se utilizó la técnica *Dynamic Time Warping* (DTW).

2.3.1. *DTW (Alineamiento Dinámico del Tiempo)*

Teniendo en cuenta el problema de que dos repeticiones de una misma palabra nunca se pronuncian igual, y que en general no son de la misma duración cada uno de los fonemas que la constituyen, hay que utilizar una técnica capaz de alinear de manera adecuada los sonidos registrados en dos patrones distintos de una misma palabra, para garantizar así una comparación razonable de los datos. DTW tiene su punto de partida en el apareamiento de plantillas *(template matching)*. Para llevar a cabo la prueba de apareamiento de éstas es indispensable que cada palabra se encuentre alineada en el tiempo con la plantilla que se encuentra en observación. Con el fin de resolver este problema se hace uso de la técnica de Programación Dinámica (DP)[9].

En la práctica se comparan cadenas de vectores cepstrales extraídos de la señal de voz de cada palabra. Las cadenas de vectores de la frase de prueba y de la frase de referencia *(template)* se colocan sobre los ejes I y J, respectivamente; los primeros vectores de cada cadena quedan en los puntos 1 de cada eje. Estas cadenas son de la forma:

Cadena de prueba $\mathbf{t}(1),\ldots\ldots ,\mathbf{t}(I)$
Cadena de referencia $\mathbf{r}(1),\ldots\ldots ,\mathbf{r}(J)$

Cada nodo en el plano ha sido indexado con un número entero positivo. El problema básico consiste en encontrar el camino con la distancia más corta o con el menor costo asociado a éste a través de la rejilla que comienza en el nodo origen (0,0) y termina en el nodo terminal (I,J).[10] Las distancias o costos son asignadas a los caminos de tres formas: Asignación de costo tipo T, tipo N y tipo B.[11] En el tipo T hay un costo asociado con la transición a un nodo cualquiera desde su predecesor. En el tipo N, los costos son asignados con los nodos en sí, en vez de la transición entre ellos (éste fue el tipo de costo que se utilizó en el algoritmo). El tipo B es aquel costo donde tanto la transición como los nodos poseen costos asociados.

---

[9] BEDOYA, Mauricio y ROSELLO, Alberto, *op. cit.*

[10] *Ibidem.*

[11] DELLER, Jhon; PROAKIS, John y HANSEN, Jhon, *op. cit.*

La distancia asociada con el camino completo es usualmente tomada como la suma de los costos de transición y/o nodos a través de la trayectoria completa. El principio de la DP es que para calcular el costo mínimo a cada punto de la rejilla se debe calcular el camino más corto hasta los puntos anteriores posibles y sumarles el costo de pasar de esos puntos al actual y tomar como camino el más económico entre los hallados con este método. Este recurso se repite hasta llegar al punto de inicio.[12]

Ahora podemos asociar a cada punto del plano cartesiano (i,j) un costo igual a la distancia entre t(i) y r(j):

$$d_N(i_k,j_k) = d_2[t(i_k),r(j_k)] = \|t(i_k) - r(j_k)\|$$

Esta distancia se aplica en el caso de los coeficientes cepstrales.

El camino óptimo se halla calculando Dmin con DP. La ecuación para este camino es de la forma[13]:

$$Dmin = \sum_{k=1}^{\#\,p} d(i_k,j_k \mid i_{k-1}, j_{k-1})$$

donde (io , jo) = (0 , 0) , el # de p depende de la trayectoria del camino y

$d(i_k,j_k \mid i_{k-1}, j_{k-1})$ es el costo de pasar de (i_{k-1}, j_{k-1}) a (i_k,j_k).

- Para buscar el camino de costo mínimo a un punto de la grilla es necesario buscar el punto anterior entre puntos tales, $i_{k-1} \leq i_k$ y que $j_{k-1} \leq j_k$, para ahorrar tiempo de cómputo; esta condicion se extrae de la física del problema. La única forma de que no se cumpliera sería que la persona hablara "al revés."[14]

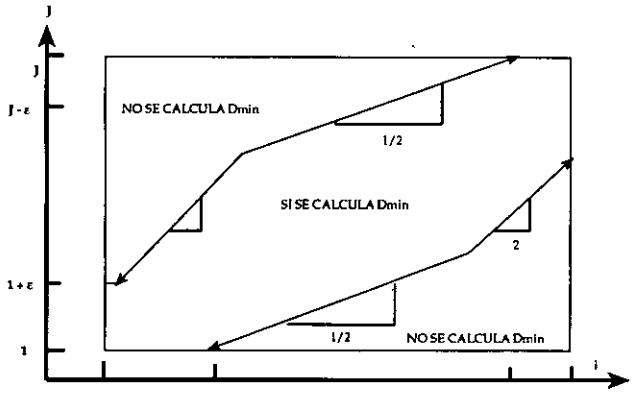
Con el fin de evitar una compresión o una expansión excesiva de tiempos entre las palabras de referencia y las de prueba, aparecen las restricciones locales.

- Para independizar la selección de la frase de referencia con relación a su tamaño se divide el costo asociado a cada punto de la grilla por el número de pasos requeridos para llegar a ellos.[15]

Para economía de cómputo, y teniendo en cuenta que las variaciones entre diferentes iteraciones de la misma frase no son muy grandes, sólo se calcula el Dmin(i,j) en los puntos que se indican en la figura.[16]

A continuación se muestra en la Fig. 4 el diagrama de bloque del algoritmo para el entrenamiento.

[12] *Ibid.*
[13] *Ibid.*
[14] BEDOYA, Mauricio y ROSELLO, Alberto, *op. cit.*
[15] *Ibid.*
[16] *Ibid.*


J
J
J - ε
NO SE CALCULA Dmin
1/2
SI SE CALCULA Dmin
2
1 + ε
1/2
NO SE CALCULA Dmin
1
i

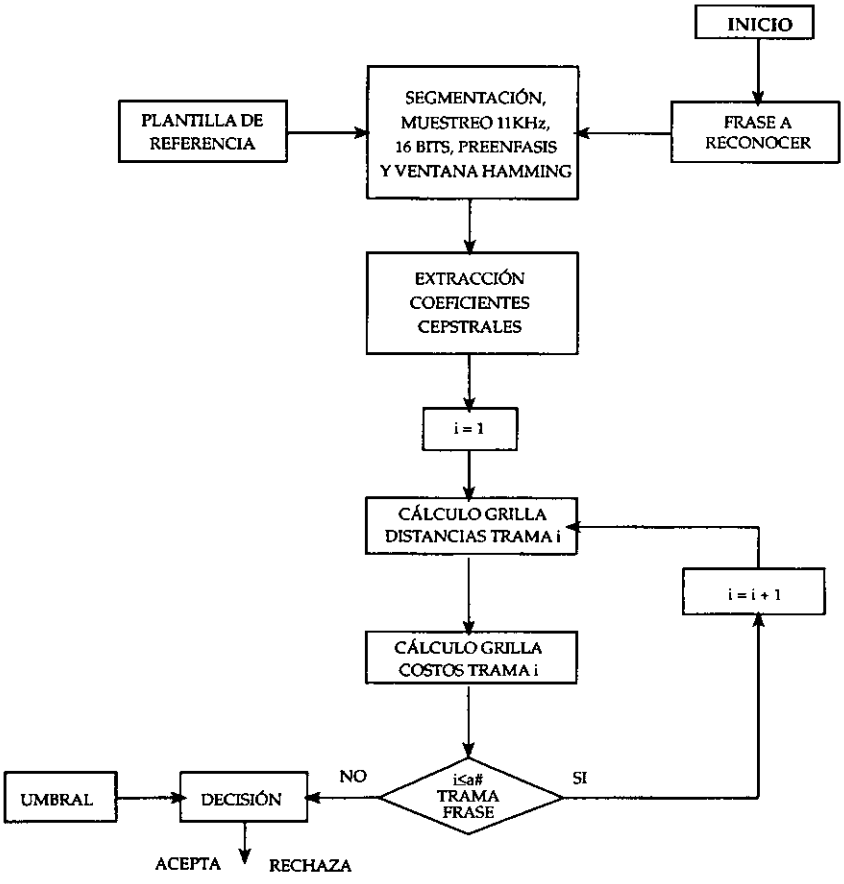

INICIO
PLANTILLA DE REFERENCIA
SEGMENTACIÓN, MUESTREO 11KHz, 16 BITS, PREENFASIS Y VENTANA HAMMING
FRASE A RECONOCER
EXTRACCIÓN COEFICIENTES CEPSTRALES
i = 1
CÁLCULO GRILLA DISTANCIAS TRAMA i
i = i + 1
CÁLCULO GRILLA COSTOS TRAMA i
UMBRAL
DECISIÓN
NO
i≤a# TRAMA FRASE
SI
ACEPTA
RECHAZA

Figura 4

3. Experimento

Para llevar a cabo el diseño de un sistema de verificación de hablante basado en DTW, se realizaron los siguientes objetivos particulares:

3.1. *Creación de un banco de voces*

Se contó con la colaboración de 10 personas (5 hombres y 5 mujeres), con edades entre 24 y 32 años.

Cada una de estas personas digitalizó la secuencia "Cero, uno, dos, tres, cuatro, seis, ocho", en 10 sesiones diferentes, separadas entre sí una semana como mínimo, lo cual generó una base de datos de 100 señales de voz (17 Mbytes Aprox.).

3.2. *Procesamiento de la señal de voz*

El banco de voces fue procesado realizando los pasos descritos en el ítem 2.2.1.

3.3. *Diseño del sistema verificador*

Este diseño se observa en la estructura del diagrama de bloques de la figura 5. Las especificaciones básicas son:

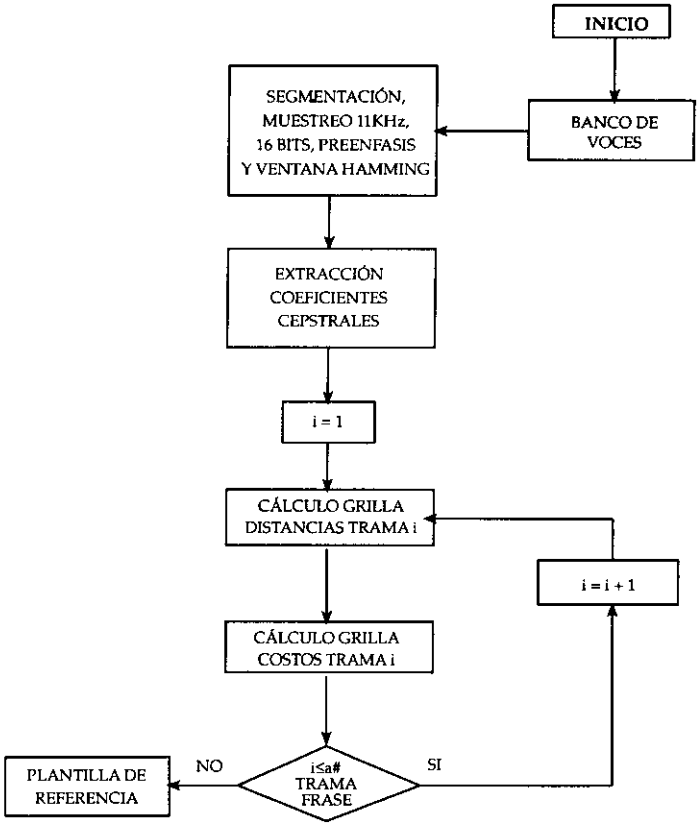


Figura 5.

- Texto fijo
- Procesamiento de las señales basado en predicción lineal
- Patrones individuales por persona
- DTW para la generación, comparación y reconocimiento de patrones
- 10 personas en total: 5 hombres y 5 mujeres.
- Condiciones de bajo ruido (cuarto cerrado y en silencio)

3.3.1. *La generación y comparación de patrones*

El proceso que se utiliza para generar el patrón de una persona es el siguiente:

- Se procesan las 10 señales de cada persona utilizando el diagrama de bloques de la figura 4.

- Se evalúa y se escoge la señal de cada persona que mejor representa a las nueve señales restantes, siendo éste el patrón que debe utilizar cada persona.

3.3.2. *La regla de decisión*

El bloque de decisión del sistema de verificación determina una distancia o costo resultante entre la señal de prueba con los respectivos patrones.

El criterio que se utiliza para aceptar como verdadera la identidad del usuario es comparar las distancias obtenidas en el bloque de decisión con un umbral previamente establecido. Si cualquiera de las distancias es menor a este umbral, el usuario es aceptado; en caso contrario es rechazado.

3.3.3. *Evaluación del sistema*

Las condiciones utilizadas para realizar la evaluación del sistema fueron:

- Se utilizó un computador personal con procesador pentium de 166 MHZ y 32 MB de RAM.

- Dicho computador posee una tarjeta de sonido *sound blaster* de 16 bits y un micrófono convencional.

- Frecuencia de muestreo de 11025 Hz.

Se utilizaron 10 señales de cada una de las personas del banco de voz y se generaron con ellas patrones de voz característicos para cada usuario. Luego, utilizamos 9 señales de cada persona y las empleamos como señales de prueba en el sistema de verificación. De esta forma el sistema intentó verificar 9 señales verdaderas y 90 falsas para cada uno de los 10 usuarios.

En total se hicieron 90 intentos de verificaciones válidas y 900 intentos de suplantación (todos los intentos de verificaciones válidas y de suplantación se realizaron con la frase de verificación).

Con el fin de determinar qué parámetros son los óptimos para el funcionamiento del DTW en la verificación de habla continua en el idioma castellano, se realizaron pruebas en las que se variaron los parámetros que afectan el desempeño del software desarrollado.

Cada prueba contenía una sección de los parámetros que se mantuvieron constantes, seguida de otra en la que se encuentra el parámetro que se varió en dicha prueba. Para finalizar, se muestran los resultados obtenidos.

Para obtener el umbral óptimo de operación se graficaron los porcentajes de error (Falsa aceptación y falso rechazo) con base en diferentes umbrales para cada prueba.

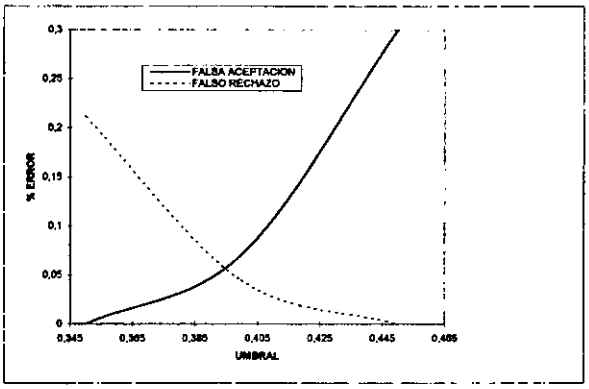
4. Pruebas

4.1. *Variación del factor de multiplicación*

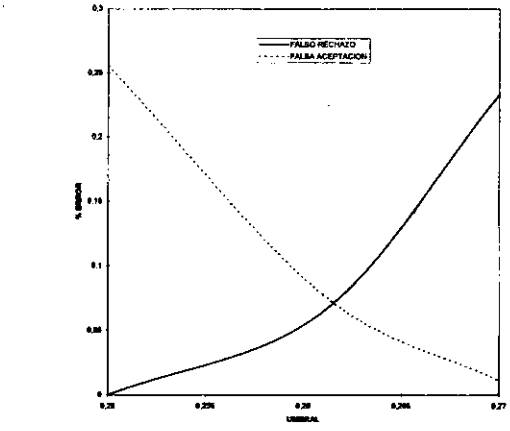
Condiciones de la prueba:

- Con preénfasis
- Ventana Hamming
- Orden LPC = 8
- Orden cepstrales = 10
- Tamaño de la ventana = 16 m.
- Restricción local

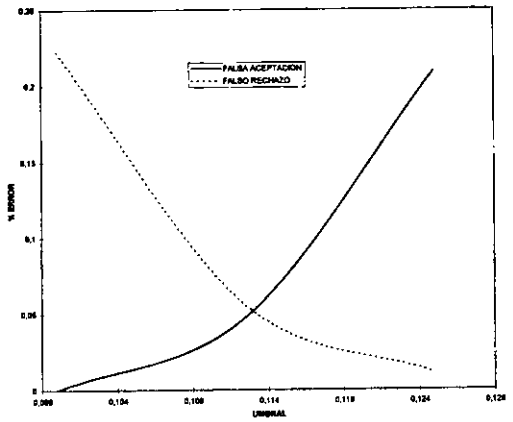
Factor de multiplicación = 0.7



Factor de multiplicación = 0.5



Factor de multiplicación = 0.2



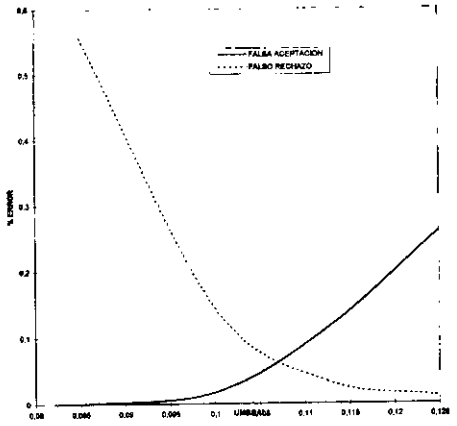
El factor de multiplicación escogido fue el 0.2

4.2. *Tamaño de la ventana*

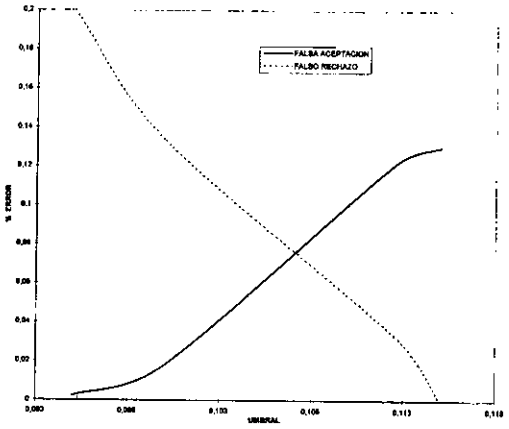
Condiciones de la prueba:

- Con preénfasis
- Ventana Hamming
- Orden LPC = 8
- Orden cepstrales = 10
- Factor = 0.2
- Restricción local

Tamaño = 24 ms



Tamaño = 32 ms



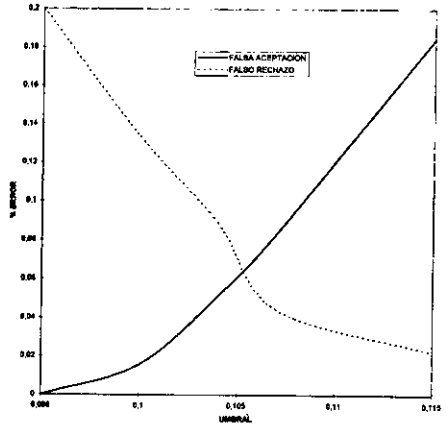
Permaneció la de 16 ms

4.3. *Variación del orden LPC*

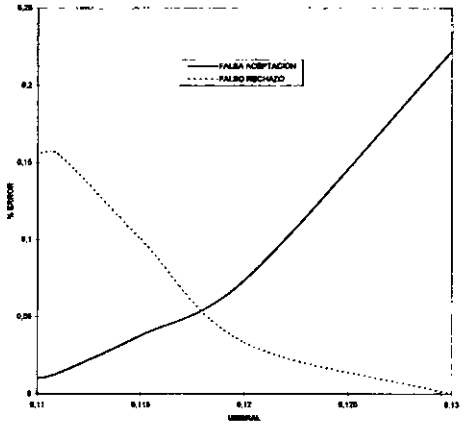
Condiciones de la prueba:

- Con preénfasis
- Ventana Hamming
- Factor = 0.2
- Orden cepstrales = 10
- Tamaño de la ventana = 16 ms
- Restricción local

Orden LPC = 6



Orden LPC = 10



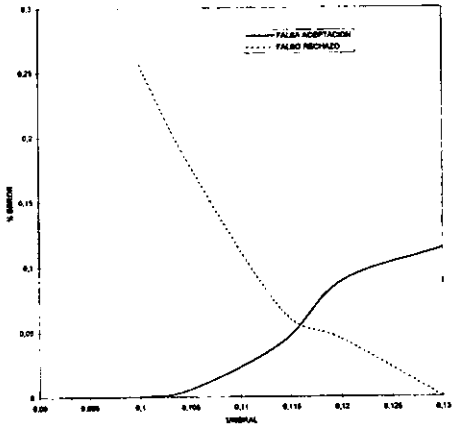
Permaneció el orden de 8

4.4. *Variación orden cepstrales*

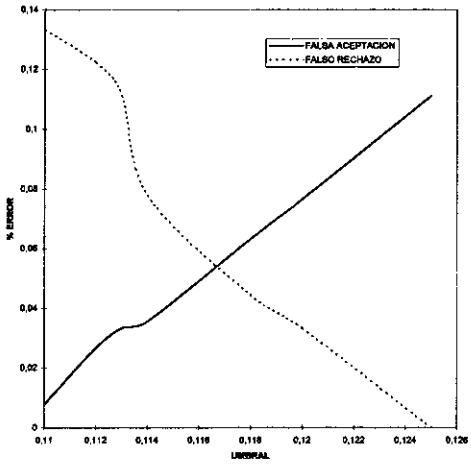
Condiciones de la prueba:

- Con preénfasis
- Ventana Hamming
- Factor = 0.2
- Orden LPC = 8
- Tamaño de la ventana = 16 ms
- Restricción local

Orden coeficientes cepstrales = 15



Orden coeficientes cepstrales = 20



5. Conclusiones y posibles proyectos

Se ha presentado un sistema de verificación basado en *Dynamic Time Warping*, donde se obtuvieron buenos resultados (porcentajes de error de falsa aceptación y falso rechazo de aproximadamente 5%), si se consideran los siguientes factores:

- Homogeneidad de la población.
- Señales de prueba y entrenamiento tomadas con semanas de diferencia.
- Tamaño de la frase de verificación. Usualmente los sistemas comerciales utilizan frases más largas, o por lo menos admiten repeticiones de éstos.
- La no utilización de un algoritmo de inicio y fin de frase.

Se pudo observar que el mejor tamaño de ventana es la de 16 ms. Esto significa que este tamaño es lo suficientemente corto para representar un segmento estacionario del tracto vocal, y lo suficientemente largo para captar el espectro de la señal de voz en estado estacionario.

En la teoría, entre mayor sea el orden del modelo LPC, la señal estará mejor representada, y por lo tanto el porcentaje de aciertos debería aumentar. Sin embargo, se observó que al aumentar el orden del modelo LPC, el desempeño del algoritmo tiene su mejor valor en 8, y en adelante se queda igual. Una posible causa de esto es que si el orden del modelo es muy pequeño, entonces no representa de forma adecuada a la señal; pero si es muy grande representa a la señal con muchos picos que no pertenecen al espectro real de ésta.

Al aumentar el orden de los cepstrales no se observó cambio en la mejora del porcentaje de error.

Se podría intentar mejorar el sistema de verificación realizando lo siguiente:

- Eliminar algunos de los primeros coeficientes cepstrales, debido a que éstos representan al filtro y no a la señal de entrada, y por lo tanto son malos para realizar verificación.

- Utilizar filtros para limitar el ancho de banda de la señal de voz, mejorando la calidad de ésta.

- Utilizar un algoritmo de inicio y fin de frase, para evitar los silencios prolongados al inicio y al final de la frase, mejorando la verificación.

- Utilizar otros tipos de restricciones locales.

Se tiene pensado realizar, entre otros, los siguientes proyectos:

- Una implementación en hardware o en procesadores digitales de señales que permitan el procesamiento en tiempo real.

- El estudio de modelos más complejos del sistema productor de la voz, como un paso fundamental para mejores representaciones de las señales de voz.

- Utilizar modelos distintos a DTW, para la generación y reconocimiento de patrones, como la aproximación estadística clásica o la utilización de técnicas adaptables al fenómeno, como sucede con los modelos escondidos de Markov (HMM).

Referencias

[1] RABINER, Lawrence y SCHAFER, Ronald. *Digital processing of speech signals.* Prentice-Hall, 1978.

[2] MARKEL, JD Y GRAY, A H. Linear prediction of speech . Springer-Verlag, 1976.

[3] O'SHAUGHNESSY, Douglas. Speaker Recognition. IEEE Asssp Magazine, octubre de 1986.

[4] NAIK, Jayant M. Speaker Verification: A tutorial. IEEE Communications Magazine, enero de 1990.

[5] FURUI, Sadaoki. Cepstral Analysis Technique for Automatic Speaker Verification. IEEE Transactions on Accustics Speech and Sígnal Processing, Vol. 29. No 2, abril de 1981.

[6] PROAKIS, John G. *Digital communications.* McGraw-Hill, 1983.

[7] DELLER, JHON; PROAKIS, John y HANSEN, Jhon. *Discrete-time processing of speech signals.* Prentice-Hall, 1987.

[8] IFEACHOR, Emmanuel y JERVIS, Barrie. Digital signals processing: A practical approach. Cap. 1. Addison-Wesley (Eds.), 1993.

[9] KAY, Steven y MARPLE, Stanley. Spectrum analysis- A modern perspective. En: *Proceedings of the IEEE.* Vol. ASSP 26, No. 1, febrero de 1978.

[10] HIROAKI, Sakoe y SEIBI, Chiba. Dynamic Programming algorithm optimization for Spoken Word recognition. IEEE Transactions on Accustics Speech and Signal Processing, Vol. 29, No 2, abril de 1981.

[11] BEDOYA, Mauricio y ROSELLO Alberto. Evaluación del método dtw para reconocimiento de habla continua. Tesis, Universidad Javeriana.

[12] BELLO, Marco y GARCÍA, Pablo. Caracterización de la voz para verificación automática de hablintes. Tesis Universidad Javeriana.

[13] GAGANELIS, D.A, FRANGOULIS, E.D. A novel approach to Speaker Verification. IEEE Transactions on Accustics Speech and Signal Processing, Vol. 42, No 7, abril de 1990.

[14] SAVIC, Michael y GUPTA, Sunil. Variable parameter speaker verification system based on hidden markov modeling. IEEE Transactions on Accustics Speech and Signal Processing, abril de 1990.

[15] G.R. DODDINGTON. A Method of Speaker Verification. J. Acoust. Soc. Amer., Vol. 49, enero de 1971.