Entre Ciencia e Ingeniería, ISSN 1909-8367
Año 7. No. 14 - Segundo Semestre de 2013, página 50 - 57

Uso de la transformación de Fourier de orden fraccional al determinar los coeficientes cepstral en las frecuencias mel para la verificación de locutores

Using the Fourier transformation of fractional order when determining the cepstral coefficients in the mel frequencies for the verification of speakers

E.F. Maldonado, D.D. Bertel, Y. Torres

Recibido Abril 22 de 2103 – Aceptado Noviembre 15 de 2013

Resumen **- La voz es una característica biométrica natural con atributos para la verificación y el reconocimiento de locutores. Si se usa la transformación de Fourier de orden fraccional es posible obtener características de la señal de voz en el espacio tiempo-frecuencia con un grado de libertad agregado, representando así de una manera diferente al locutor, de como se hace tradicionalmente con la transformación de Fourier estándar. En este estudio se compara el desempeño de la transformación de Fourier de orden fraccional en un sistema de verificación de locutores dependiente del texto con respecto a la usual representación usando los coeficientes cepstral en las frecuencias mel, MFCC. Los resultados muestran que para una elección apropiada del orden fraccional de la transformada de orden fraccional se obtiene un mejoramiento en la verificación de locutores.**

Palabras clave **- verificación dependiente del texto, verificación de locutores, transformación de Fourier de orden fraccional, coeficientes MFCC, procesamiento de señales.**

Abstract **- The voice is a natural biometric attribute for the verification and recognition of speakers. Using the Fourier transform of fractional order, it is possible to obtain characteristics of the speech signal in the time-frequency space with an added degree of freedom. The speech is represented in a different way compared to the traditionally Fourier transformation. This paper compares the performance of using the fractional Fourier transformation in the speakers' verification in text-dependent systems, instead of the common representation using the standard Fourier transformation, of the Mel Frequencies Cepstral Coefficients. The results show that, for an appropriate choice of the order of the fractional Fourier transformation, an improvement has been obtained in the verification of the speaker.**

Key Words **- Mel frequency Cespstral coefficients, Fourier transformation of fractional order, Speech feature extraction, text-dependent speaker verification, signal processing.**

[1] Producto derivado del proyecto de Investigación "Centro de excelencia en nuevos materiales – CENM/ Tratamiento de Señales", apoyado por la VIE (Vicerrectoría de Investigación y Extensión de la Universidad Industrial de Santander, Bucaramanga, Colombia) a través del apoyo a la vinculación del GOTS, Grupo de Óptica y Tratamiento de Señales al CENM, Centro de Excelencia de Nuevos Materiales, Unión Temporal legalmente constituida.

E.F. Maldonado Orduz, D.D. Bertel Mendoza son egresados de la Escuela de Ingenierías Eléctrica, Electrónica y Telecomunicaciones, Facultad de Ingenierías Físico-Mecánicas de la Universidad Industrial de Santander, A.A. 678, Bucaramanga, Colombia (correos-e: edgar.maldonado@radiogis.uis.edu.co; david.bertel@radiogis.uis.edu.co).

Y. Torres Moreno, es Profesor Titular, miembro del GOTS, Grupo de Óptica y Tratamiento de Señales, Escuela de Física, Facultad de Ciencias, Universidad Industrial de Santander, A.A. 678, Bucaramanga, Colombia (correo-e: ytorres@uis.edu.co).

I. Introducción

El estudio de los sistemas biométricos bidimensionales que se ha desarrollado hasta el día de hoy, reviste cierta dificultad generalizada en lo concerniente a la representación, análisis y procesamiento de imágenes. Este es el caso, por ejemplo, de la caracterización utilizando el iris (iris, retina), huella dactilar, geometría vascular de la mano, geometría de la cara, escritura, firma, entre otros. La voz es diferente a los sistemas biométricos mencionados anteriormente, pues además de involucrar procesamiento unidimensional, considera una mezcla de características físicas y del comportamiento como la articulación de las palabras, el estado anímico, el contexto y demás variables que se encuentran asociadas al proceso del habla. Usar la voz como patrón de reconocimiento tiene muchas ventajas debido a que su registro puede ser tomado sin contacto directo con el locutor y de manera natural. Esto haría ideal el uso de la voz en gran cantidad de sistemas, pero los métodos actuales para representar una señal de voz no brindan la confiabilidad necesaria para aplicaciones que la requieren al más alto grado, como por ejemplo las transacciones bancarias [1][2]. Es bien conocido que el objetivo principal que persigue la solución al problema de verificación de locutores es aumentar la tasa de aciertos al máximo, reduciendo al mínimo la tasa de falso rechazo y la tasa de falsa aceptación, la cual se realiza para un conjunto cerrado finito de hablantes. El presente trabajo propone, por primera vez, la introducción de un grado de libertad adicional a la solución del problema, por medio del uso de la Transformación de Fourier de orden Fraccional FrFT, como método para la parametrización de la voz mediante los coeficientes cepstral en las frecuencias mel MFCC. La FrFT es una generalización de la transformación de Fourier estándar FT, que brinda la posibilidad de analizar una señal en el dominio tiempo-frecuencia. Con la introducción del orden de la transformación, una nueva variable en el proceso de verificación de locutores, se mejora el proceso de corroborar el hablante [3].

II. Transformación de Fourier de orden fraccional

La FrFT encuentra gran aplicabilidad en la generalización y mejoramiento en las áreas donde la transformación estándar y el concepto del dominio espacio-frecuencia son utilizados. Además, la FrFT es parte importante en el estudio de otros sistemas, permitiendo una generalización de la noción bien establecida de dominio frecuencial, y aumenta así, el conocimiento sobre el producto espacio directo-frecuencia [4].

A. Definición integral

La FrFT de orden *a* es una operación canónica lineal definida por la integral [5]:

$$f_a(u) = \int_{-\infty}^{+\infty} K_a(u,u') f(u') du' \tag{1}$$

Con núcleo de transformación

$$K_a(u,u') = K_\alpha e^{i\pi\left(\cot\left(\alpha u^2\right) - 2\csc(\alpha uu') + \cot\left(\alpha u'^2\right)\right)} \tag{2}$$

Donde $\alpha \equiv \frac{a\pi}{2}$ y $K_\alpha = 1 - i\cot\alpha$ y. Para $a = 0$ y $a = \pm 2$ el núcleo se define como $K_0(u,u') = \delta(u-u')$

y y $K_{\pm 2}(u,u') = \delta(u+u')$ Para $a = 1$ se encuentra que $K_a = 1$ y,

$$f_1(u) = F_1 f(u) = \int_{-\infty}^{+\infty} e^{-i2\pi uu'} f(u') du' \tag{3}$$

Ésta última expresión corresponde a la bien conocida, transformada de Fourier estándar de la señal $f(u)$. De la misma forma $f_{-1}(u)$ es la transformada de Fourier inversa estándar.

B. Algunas propiedades importantes

La FrFT puede ser considerada como un operador que realiza una rotación. Es posible asumir la transformación como una rotación con las siguientes propiedades [6]:

1) Rotación nula $R_\pi^0 = I$
2) Coherencia con la FT $R^2 = F$
3) Adición de Rotaciones $R^\beta R^\alpha = R^{\alpha+\beta}$
4) Rotación de 2π $R^{2\pi} = I$
5) Propiedad de translación:

$$F_\alpha f(x+k) = \exp[-ik\sin\alpha(x + \frac{k}{2}\cos\alpha)] F_\alpha(f)_{[x+k\cos\alpha]} \tag{5}$$

6) Regla de similitud:

$$F_\alpha f(-x) = F_{\alpha-\pi} f(x) \tag{6}$$

7) Propiedad de la convolución[2]:

$$f \overset{a}{*} g = \exp\left\{-ibt^2\right\} \int_{-\infty}^{+\infty} f(\tau) e^{ibt^2} g(t-\tau) e^{ib(t-\tau)^2} d\tau \tag{7}$$

donde, $b = 0{,}5\ \cot(0{,}5\pi a)$.

La distribución de Wigner-Ville es una representación tiempo-frecuencia de la energía de la señal. La transformada fraccionaria de orden *a* de una señal posee una distribución de Wigner igual a la original, sólo que rotada un ángulo de $a\pi/2$ radianes en el plano tiempo-frecuencia. Esto permite que el concepto de "warping" que se realiza en el espacio temporal para aplicaciones de reconocimiento del habla y del locutor (deformación del eje correspondiente al tiempo), se pueda hacer también en el dominio fraccionario. Por otra parte, interferencia, ruido y otras fuentes de variabilidad en la señal podrían ser fácilmente removidos en un dominio fraccionario como se muestra en la Fig. 1. Cuando la señal y el ruido se solapan tanto en el tiempo como en la frecuencia, puede suceder que en un dominio fraccionario las señales lleguen a separarse totalmente [3].

[2] Una definición alterna ha sido formulada recientemente por R. Torres et al. [7].

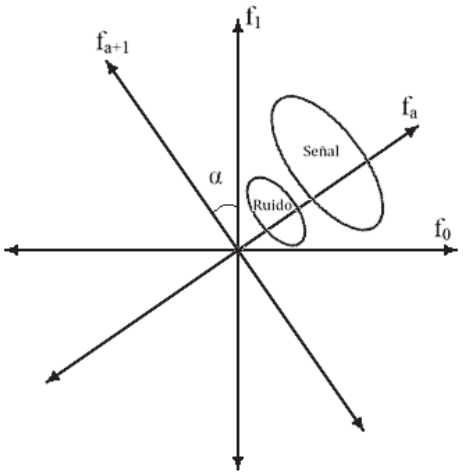

Fig. 1. Rotación en el espacio tiempo-frecuencia del soporte de la distribución de Wigner. Fuente autores.

III. Breve introducción a la verificación actual de locutores

El reconocimiento del locutor es un término genérico para la clasificación de la identidad basándose en una señal acústica. Para la identificación del locutor la persona se clasifica como un integrante de un conjunto finito de locutores, requiriéndose una comparación de una determinada expresión hablada con un conjunto de referencias de cada locutor potencial. Como resultado se determina la identidad de la persona o el evento de no pertenecer al grupo presente en el proceso de entrenamiento. Para el caso de la verificación del locutor se pide una clave (ingresada por medio físico o mediante reconocimiento de otra característica biométrica), y luego con la señal de voz se comprueba si realmente es quien dice ser, clasificándose como poseedora o no de la identidad manifestada [8][9]. El reconocimiento en conjunto abierto consiste en decidir si un locutor pertenece a un conjunto P de locutores conocidos, sin buscar decidir cuál de los P locutores es. La verificación de locutor se reduce al caso particular de la identificación en un conjunto abierto con P = 1. Los sistemas de verificación de locutores se pueden dividir en dos grandes grupos: dependientes e independientes del texto. En los sistemas dependientes del texto se requiere la pronunciación de las mismas palabras usadas en el entrenamiento del sistema, mientras que en los independientes se puede usar cualquier texto, implicando una complejidad muy superior [8]. Generalmente el desempeño de un sistema de verificación de locutores se evalúa de acuerdo a dos tipos de errores:

- Tasa de falsa aceptación (TFA): probabilidad de verificar erróneamente a un impostor.
- Tasa de falso rechazo (TFR): probabilidad de no verificar como válido a un usuario del sistema.

Una forma sencilla de evaluar el desempeño de un sistema de verificación de locutores es utilizar una función de costo dada por,

$$C = c_1 TFA + c_2 TFR \quad (8)$$

Donde c_1 y c_2 corresponden a los pesos acordados a cada uno de estos errores [2]. El valor que se le asignen a los pesos depende de las características del sistema. Por ejemplo, si se desea tener un sistema con alta seguridad se debe dar mayor peso a la tasa de falsa aceptación para rechazar de manera más efectiva a los intrusos. Cuando se da un valor de 0,5 a los dos pesos la función de costo representa la media de las tasas, la cual se conoce como HTER (del inglés, "*Half Total Error Rate"*).

A. Limitaciones de los sistemas biométricos basados en la voz

Gran parte de los sistemas biométricos de reconocimiento de voz que se han desarrollado basan su análisis y representación en la transformación de Fourier estándar; resultados de investigaciones arrojan tasas de error de alrededor del 10%, lo cual no es despreciable en la práctica [8]. En la actualidad, los sistemas más difundidos que utilizan reconocimiento de voz poseen una etapa dedicada al cálculo de la transformada de Fourier, para el análisis referente a las características representadas en el espacio de frecuencias. El problema por el cual la comercialización de los sistemas biométricos de reconocimiento de voz no se ha dado, radica en que no ha sido posible obtener una adecuada caracterización de la señal de voz que permita discernir entre un locutor y otro: la parametrización no satisface el objetivo del sistema. Los sistemas que pretenden alcanzar tasas de error reducidas implican tiempos de ejecución imprácticos; por otro lado, cuando el sistema de reconocimiento ha sido entrenado, la identificación o verificación se deben hacer en las mismas condiciones en las que se ha entrenado el sistema, esto es, con el mismo equipo, micrófono, ubicación, ruido, entre otros. En realidad hay muchos aspectos que aún no han sido entendidos, y muchos otros incluso no se conocen. Actualmente, la capacidad de un sistema de reconocimiento automático del habla es bastante inferior que la de un ser humano; el desempeño decae rápidamente con pequeñas modificaciones tales como el cambio del micrófono que se utiliza o las condiciones del canal entre otros.

Varias son las razones por las que el reconocimiento de la voz es generalmente difícil: Primero, el habla natural es continua; no existen pausas entre las palabras, haciendo difícil determinar sus límites. También los locutores cambian su pensamiento en la mitad de una frase, pronunciando incorrectamente los fonemas o agregando sílabas sostenidas para hacer una pausa (por ejemplo "eee...", "mmm...").

Segundo, el habla natural puede variar su velocidad y la articulación de los fonemas dependiendo del contexto, de las emociones, de la misma forma que la pronunciación de ciertas palabras cambia de una persona a otra. El espectro varía, a menudo dramáticamente, si una de estas modificaciones se presenta incluso con los tamaños de las ventanas que se toman en los sistemas actuales [3]. Tercero, la grabación de la voz varía con la acústica de la habitación, las particularidades del canal, las características del micrófono y el ruido de fondo.

Por ejemplo, usar un micrófono a diferentes grados de inclinación cambia su respuesta en frecuencia e incluso se podrían presentar efectos no deseados como fonemas nasales mucho más fuertes por tener el micrófono cerca de la nariz.

Todos estos factores cambian las características de la señal, una diferencia que los humanos usualmente podemos compensar, pero que los actuales sistemas de reconocimiento no, haciendo de un sistema biométrico un poco más complejo que los demás sistemas conocidos [8]. Los algoritmos para el entrenamiento de sistemas de reconocimiento también deben ser elegidos cuidadosamente, pues grandes tiempos de entrenamiento no son prácticos. Algoritmos que toman demasiado tiempo para ejecutarse pueden ser de un gran interés teórico, pero dado que la mayoría presentan errores no permitirían llevar a cabo un verdadero desarrollo experimental. Pero aún con todas las limitaciones existentes, la investigación de sistemas basados en voz es motivada por el mercado potencial que éstos representan, calculándose que las ganancias y ahorros que se obtendrían de simples aplicaciones telefónicas ascienden a cientos de millones de dólares por año [9].

B. Sistemas actuales de verificación

Para el proceso de reconocimiento, se divide la señal de voz en tramas típicamente de 10 a 30 [ms], creándose un vector de características. Después de obtener una secuencia de vectores se comparan con diferentes modelos previamente almacenados para tratar de determinar quién es el locutor. No obstante, se puede reducir la cantidad de datos por medio de la parametrización, disminuyendo la complejidad computacional del proceso de reconocimiento y transformando la señal de voz en un nuevo espacio de características, donde es más sencillo distinguir al locutor. En este sentido los coeficientes LPC (del inglés, "*Linear Prediction Coding"*) y Cepstrum, con sus asociados respectivos, son las características más usadas en el reconocimiento, siendo los últimos los más estables entre las pronunciaciones repetidas de una misma persona [8].

C. Filtros mel

El comportamiento del oído humano, en cuanto a la percepción de las frecuencias se refiere, es de tipo logarítmico.

Los sistemas convencionales de reconocimiento del habla y del locutor, así como la verificación del locutor, hacen uso de esta propiedad al introducir en sus algoritmos un filtrado, denominado filtrado mel (de melodía) al espectro de la señal de voz y analizar los coeficientes obtenidos. El filtrado mel aproxima el comportamiento del oído a una escala logarítmica de frecuencias representada por la siguiente función,

$$f_{mel} = 2595 \times \log\left(+ \frac{f}{700\,Hz} \right) \qquad (9)$$

La teoría de los filtros mel ha sido ampliamente desarrollada, aunque la obtención de estos filtros se hizo de manera experimental [10].

F. Coeficientes Cepstral en las frecuencias mel

Como producto de un procedimiento de filtrado y posteriormente, una transformación Coseno Discreta (DCT – del inglés, "*Discrete Cosines Transform"*), se obtienen los coeficientes cepstral en las frecuencias mel (MFCC – del inglés, "*Mel Frequency Cepstral Coefficients"*). La transformación Coseno se realiza con el fin de disminuir la extensión de los vectores obtenidos a partir del filtrado mel. El esquema general utilizado para la obtención de los MFCC se observa en la Fig. 2. Aunque existen otras técnicas, éstas no han sido ampliamente difundidas y aceptadas [2].

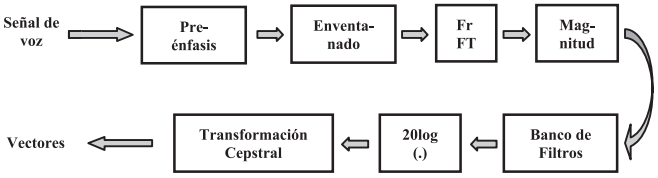

Fig. 2. Obtención de los coeficientes MFCC. Fuente: Autores.

Los coeficientes MFCC son los parámetros más utilizados y aceptados para la extracción de características del habla y del locutor. Los vectores generados por estos coeficientes son utilizados en el entrenamiento y prueba de sistemas de reconocimiento y verificación [2]. El objeto de la técnica propuesta y presentada aquí se centra precisamente en los coeficientes mel y las distintas implicaciones que puede llegar a tener el hecho de que para la verificación del locutor, en lugar de usar su representación en el espacio frecuencial, se use una representación generalizada en el espacio tiempo-frecuencia, la cual permite la introducción de un grado de libertad adicional, que se ha mostrado útil en otras aplicaciones.

G. Modelado estadístico

Entre los modelos más usados para el reconocimiento de locutores están [8]:

- Los paramétricos:
  - Redes neuronales (ANN – del inglés *"Artificial Neural Networks"*).
  - Modelos ocultos de Markov (HMM – del inglés *"Hidden Markov Models"*).
- Los no paramétricos:
  - Cuantificación vectorial (VQ – del inglés *"Vector Quantization"*).
  - Vecino más cercano (NN – del inglés *"Nearest Neighbor"*).
  - Máquinas de vectores de soporte (SVM – del inglés *"Support Vector Machines"*).

El modelo paramétrico presenta la ventaja de necesitar pocos datos para definir la función de densidad de probabilidad. Entre menos datos más limitado es el modelo. Si el modelo es muy restrictivo, es posible que no sea suficientemente ajustado a la realidad que se pretende modelar. El modelo no paramétrico, puesto que es menos restrictivo,

permite un mejor modelado pero requiere un número mayor de vectores de características, especialmente cuando la dimensión de los vectores es elevada. De hecho, la cantidad de datos necesarios para representar las características de la voz de un determinado locutor crece exponencialmente con la dimensión de los vectores. Esto restringe el uso de los modelos no paramétricos y de vectores de características con un número elevado de componentes [8].

La técnica del vecino más cercano, que se usará aquí, calcula la distancia Euclidiana entre los vectores obtenidos durante el entrenamiento con los de la fase de prueba, obteniendo una matriz de distancias para cada uno de los locutores. El vecino más cercano está determinado por la mínima distancia entre el vector de prueba y todos los vectores de entrenamiento para cada locutor. Las distancias mínimas obtenidas para cada locutor se promedian y se comparan para hallar cuál es el locutor que ha proporcionado la menor distancia [8].

IV. Sistema propuesto, basado en la FRFT

A. Descripción del sistema

Las pruebas se hicieron usando EUSTACE, la base de datos de voz en inglés de la Universidad de Edinburgh [11]. La base de datos se compone de las grabaciones de seis locutores. De cada uno de los locutores se usaron catorce grabaciones de la misma palabra. Las grabaciones son tomadas a 16 kHz y cuantificadas a 16 bits. Para el proceso de enventanado se toman 100 ventanas de 16 [ms] cada segundo. Finalmente, cada una de las ventanas es parametrizada por 13 coeficientes. Como resultado del proceso de parametrización se obtiene una matriz de dimensión 13xN, donde N representa el número de ventanas tomadas de la señal en cuestión. El proceso de parametrización se realizó usando los coeficientes MFCC. Para su obtención se utilizó el algoritmo provisto por el Auditory Toolbox de Interval Research Corporation [12], donde fue introducida la FrFT a cambio de la FT estándar. Un banco de cuarenta filtros para modelar el sistema de percepción auditivo humano es usado. Finalmente, se utilizó la técnica del vecino más cercano para evaluar el desempeño del sistema por las ventajas descritas previamente, se trata de comparar los desempeños de la FrFT frente a la FT al momento del cálculo de los MFCC.

B. Uso de la FrFT en los coeficientes MFCC

Es necesario hacer claridad acerca del hecho que la teoría de los bancos de filtros y de los coeficientes MFCC se ha desarrollado para el dominio frecuencial. Se habla de las implicaciones a causa de la inclusión de la FrFT sobre los MFCC y no perder de vista el esquema usualmente utilizado, pero se debe resaltar que al no estar en el espacio inverso o de frecuencias, no se trata de los coeficientes MFCC en dicho espacio, sino en el espacio tiempo-frecuencia a donde la transformación de Fourier de orden fraccional nos conduce.

C. Análisis del mejor dominio fraccionario: Evaluación del error y del desempeño

Para analizar cuál es el mejor dominio fraccionario para el reconocimiento de locutores se evaluó la TFA, la TFR y la HTER, desde el orden 0,1 hasta el orden 1,0 con un paso de 0,1 como se observa en la TABLA I. Se hicieron pruebas con tres radios de aceptación para el modelo de decisión: radio de una desviación estándar, de 1,5 desviaciones estándar y de 2 desviaciones estándar, como se evidencia en la tabla.

TABLA I.
RESULTADOS CON DISTINTOS ÓRDENES FRACCIONALES EN %. (A) EL RADIO DE ACEPTACIÓN ES UNA DESVIACIÓN ESTÁNDAR. (B) EL RADIO DE ACEPTACIÓN ES 1,5 DESVIACIONES ESTÁNDAR. (C) EL RADIO DE ACEPTACIÓN ES

Orden	TFA [%]	TFR [%]	HTER [%]
0,1	41,4	78,57	60,0
0,2	53,3	78,57	66,0
0,3	47,1	69,05	58,1
0,4	50,0	71,43	60,7
0,5	46,7	66,67	56,7
0,6	47,1	64,29	55,7
0,7	44,3	66,67	55,5
0,8	40,0	59,52	49,8
0,□	31,9	57,14	44,5
1,0	21,9	57,14	39,5

□□□

Orden	TFA [%]	TFR [%]	HTER [%]
0,1	63,8	73,81	68,8
0,2	70,5	64,29	67,4
0,3	69,0	59,52	64,3
0,4	63,3	52,38	57,9
0,5	60,0	52,38	56,2
0,6	61,9	40,48	51,2
0,7	62,9	38,10	50,5
0,8	57,6	33,33	45,5
0,□	48,1	30,95	39,5
1,0	40,0	28,57	34,3

□□□

Orden	TFA [%]	TFR [%]	HTER [%]
0,1	77,6	59,52	68,6
0,2	80,5	52,38	66,4
0,3	79,5	35,71	57,6
0,4	74,3	33,33	53,8
0,5	77,6	30,95	54,3
0,6	77,1	23,81	50,5
0,7	76,7	19,05	47,9
0,8	74,3	21,43	47,9
0,□	67,1	7,14	37,1
1,0	55,2	2,38	28,8

□□□

DOS DESVIACIONES ESTÁNDAR.

Como se observa, los resultados mejoran a medida que los órdenes se acercan a la unidad. Esto podría explicarse por el hecho que el filtrado mel es originalmente definido para el espacio frecuencial [10]. Se realizaron pruebas con órdenes fraccionales cercanos a 1, buscando afinar la búsqueda donde se tiene la representación estándar de los MFCC, en pasos más cerrados de 0,01 como se observa en la TABLA II [3]. Las celdas sombreadas señalan los valores más bajos alcanzados.

TABLA II.
RESULTADOS CON ÓRDENES FRACCIONALES CERCANOS A 1 EN %. (A) EL RADIO DE ACEPTACIÓN ES UNA DESVIACIÓN ESTÁNDAR. (B) EL RADIO DE ACEPTACIÓN ES 1,5 DESVIACIONES ESTÁNDAR. (C) EL RADIO DE ACEPTACIÓN ES DOS DESVIACIONES ESTÁNDAR.

Orden	TFA [%]	TFR [%]	HTER [%]
0,91	28,6	59,52	44,0
0,92	28,6	59,52	44,0
0,93	27,6	59,52	43,6
0,94	26,7	59,52	43,1
0,95	25,2	59,52	42,4
0,96	23,3	59,52	41,4
0,97	23,3	57,14	40,2
0,98	22,4	54,76	38,6
0,99	21,9	54,76	38,3
1,00	21,9	57,14	39,5

□□□

Orden	TFA [%]	TFR [%]	HTER [%]
0,91	45,2	30,95	38,1
0,92	44,3	26,19	35,2
0,93	43,8	26,19	35,0
0,94	42,4	26,19	34,3
0,95	41,4	28,57	35,0
0,96	41,0	26,19	33,6
0,97	38,1	28,57	33,3
0,98	38,6	28,57	33,6
0,99	40,5	28,57	34,5
1,00	40,0	28,57	34,3

□□□

Orden	TFA [%]	TFR [%]	HTER [%]
0,91	66,2	7,14	36,7
0,92	63,8	9,52	36,7
0,93	60,5	9,52	35,0
0,94	59,5	9,52	34,5
0,95	59,5	9,52	34,5
0,96	56,7	7,14	31,9
0,97	56,2	4,76	30,5
0,98	54,8	2,38	28,6
0,99	55,7	0,00	27,9
1,00	55,2	2,38	28,8

□□□

Generalmente la comparación de sistemas de verificación de locutores es muy limitada debido a las múltiples condiciones experimentales que se presentan y los diversos entornos de trabajo al que se enfrenta un sistema, de ahí que sea más conveniente usar bases de datos ya establecidas como la Eustace. Sin embargo, con el propósito de brindar una visión del desempeño de sistemas ya en uso se encuentra que la HTER es de alrededor del 3% para sistemas dependientes del texto con grabaciones con muy bajo ruido. Si por ejemplo se usan grabaciones tomadas vía telefónica el desempeño de la HTER puede variar del 2 al 15%. Para sistemas que usan micrófonos de bajo desempeño, lo cual implica mayor ruido, se presentan tasas HTER del 20 al 30% [1][2].

Un análisis de los resultados de las tablas anteriores conduce a que los órdenes con mejores resultados en HTER son 0,99; 0,97 y 0,99, para los radios 1, 1,5 y 2 veces la desviación estándar respectivamente. Puesto que el propósito de un sistema es brindar igual importancia tanto a la TFA y la TFR [2], el radio de decisión de 1,5 veces la desviación estándar es el que brinda los desempeños más próximos en las dos medidas. Por lo tanto, de acuerdo a los resultados, tomar para el orden el intervalo [0,96 – 0,98] brinda el desempeño más equilibrado para el funcionamiento del sistema, de acuerdo con estos resultados. Un sistema de verificación donde no sea importante la falsa aceptación pero sí el falso rechazo, servicio de atención de clientes por ejemplo, tendrá interés en utilizar un orden fraccional de 0,99 donde la tasa de falso rechazo es nula. Para aplicaciones donde se requiera de mayor seguridad el orden fraccional a elegir será también de 0,99, pero con un radio de aceptación menor, una desviación estándar a cambio de dos por ejemplo.

Otra manera de presentar los resultados obtenidos es usar la curva ROC (del inglés, *"Receiver Operating Characteristic"*), *"*), la Fig. 3. El mejor método posible de verificación se situaría en un punto en la esquina superior izquierda, o coordenada (0,100) del espacio ROC (no accesible en la figura donde ella se ha ampliado para efecto de una mejor visualización de los resultados), representando un 100% de sensibilidad (ningún falso negativo) y un 100% también de especificidad (ningún falso positivo). Como un buen criterio de desempeño, la distancia del punto resultado a dicha coordenada "ideal", el mejor orden fraccional para el caso de un radio de aceptación de una desviación estándar, sería 0,96. La diferencia entre las distancias correspondientes al orden 1,00 o FT frente al orden 0,96 es fácilmente calculable y es 1,42.

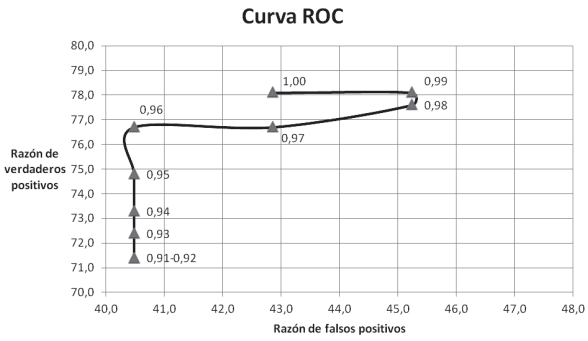

Fig. 3. Curva ROC para el caso de obtención de los coeficientes MFCC cuando el radio de aceptación es una desviación estándar. Se evalúa el desempeño utilizando sólo falsos positivos y los verdaderos positivos de la observación en la TABLA II (A). Fuente: Autores.

D. Costo computacional

Dado que el único cambio que se propone, frente a las técnicas de hoy en día, es la utilización de la FrFT en reemplazo de la FFT que dispone Matlab; se evalúa el costo computacional de dicho cambio. Cabe aclarar que el tiempo de ejecución depende de las características del equipo utilizado para realizar las pruebas y las condiciones a las que esté sometido. El equipo utilizado presenta un procesador AMD Athlon 64 X2 de 1,8 GHz y 960 MB de memoria RAM. Sin ejecutar otro programa distinto a Matlab el tiempo de ejecución empleado con el método convencional es en promedio de 293,21 [s]. El tiempo de procesamiento usando la FrFT en las mismas condiciones es en promedio de 320,57

[s], presentándose un aumento en el tiempo de procesamiento de 27,36 [s], lo cual representa un incremento del 9,33 % del tiempo total que toma el sistema para el cálculo de las distintas tasas de error.

V. Análisis de resultados y conclusiones

El presente trabajo propone modificar la etapa de parametrización en un sistema de verificación de locutores con la introducción de la FrFT en substitución de la FFT.

Se estudiaron los coeficientes cepstral en las frecuencias mel (MFCC) por ser el método de parametrización más usado por los sistemas actuales de reconocimiento. El método estadístico de decisión por el vecino más cercano, es la técnica no paramétrica que mejor se ajusta a las condiciones del presente trabajo. Los resultados obtenidos muestran que a aunque los filtros mel están diseñados para ser usados en el dominio frecuencial, los menores porcentajes de HTER se obtienen con una FrFT de orden fraccional próximo a la unidad. Al realizar pruebas con mayor resolución, centésimas de orden, en la proximidad del orden uno se encontró que con un orden en el intervalo [0,96 – 0,98] se obtiene una mejora de más del 1,4% con respecto a un sistema basado en la transformada de Fourier estándar. Puesto que el banco de filtros mel fue obtenido de manera experimental, la FrFT permite realizar una sintonización que adapta los filtros a la señal de voz del locutor en cuestión en un espacio de representación tiempo-frecuencia cercano al puramente frecuencial. Aunque la mejora es pequeña en números, es significativa cuando se trate el problema en volumen, y abre la posibilidad para un rediseño de los bancos de filtros en combinación con un modelo estadístico más elaborado que permita obtener mejores resultados en el dominio fraccionario.

En la búsqueda de mejorar los resultados aquí obtenidos se proponen las siguientes acciones: Uso de la fase para la obtención de los coeficientes MFCC; la mayoría de los sistemas de verificación del locutor como se evidenció anteriormente utilizan exclusivamente la magnitud de los coeficientes que representan el espectro de la señal, dejando de lado la información que contiene la fase. Estudios recientes muestran el potencial del uso de la información que contiene la fase en la verificación [13]. Uso de la convolución fraccionaria invariante por traslación; la definición usual de la convolución fraccionaria exhibe sólo parcialmente propiedades de invariancia que no permiten su uso en varias aplicaciones de procesamiento de señales. Una nueva definición, realizada por R. Torres et al. [7], podría llegar a ser útil para mejorar los resultados de la presente investigación. Pruebas con bases de datos independientes del texto, basados en los resultados prometedores obtenidos que aplican únicamente para sistemas texto-dependientes. Los sistemas texto-independientes no fueron revisados, por lo tanto sería pertinente analizar los resultados para la inclusión de la FrFT. Pruebas con LPC; los coeficientes LPC junto con los coeficientes mel son los más usados en el reconocimiento de voz. Cada uno de estos coeficientes representa características distintas, por lo tanto una parametrización que contenga una combinación de estos dos puede ser muy útil teniendo en cuenta los efectos que puede llegar a tener la inclusión de la FrFT. Dado que en la actualidad se requieren varias capas de procesamiento de la señal para obtener un resultado que aún no satisface las necesidades reales de un sistema de verificación del locutor, queda la duda si las técnicas empleadas son las adecuadas, por lo que se considera que después de muchos años de investigación la verificación de locutor es aún un problema abierto sin resolver [2]. Esto motiva a que se pruebe el sistema basado en la FrFT con otras capas de procesamiento como por ejemplo la normalización de los datos entre otras.

Agradecimientos

Los dos primeros autores agradecen al grupo GOTS por el apoyo brindado. De la misma manera agradecen al profesor Jaime Guillermo Barrero Pérez de la E3T y a los ingenieros Euclides Alfonso Rueda Díaz e Idriss Tyler Sandoval Villamizar. Su apoyo fue muy valioso en la realización del trabajo de investigación cuyos resultados se presentan aquí.

Referencias

[1] Reynolds D.A. *An Overview of Automatic Speaker Recognition Technology. IEEE ICASSP 2002*. 2002, vol. IV, pp. 4072-4075.

[2] Bimbot, F.; Bonastre, J.F.; Fredouille, C.; Gravier, G.; Magrin-Chagnolleau, I.; Meignier, S.; Merlin, T.; Ortega-García, J.; Petrovska-Delacrétaz, D. and Reynolds, D.A. A Tutorial on Text-Independent Speaker Verification. *EURASIP 2004*. 2004, vol. 4, pp. 430-451.

[3] Srikaya, R.; Gao, Y. and Saon, G. *Fractional Fourier Transform features for speech recognition. IEEE ICASSP 2004*. 2004, vol. I, pp. 529-532.

[4] Ozaktas, H.M.; Zalevsky, Z. and Kutay, M.A. *The Fractional Fourier Transform: with applications in optics and signal processing.* Chichester: John Wiley & Sons, 2001. Wiley Series in Pure and Applied Optics Series , #39, 513pp. ISBN: 978-0471963462.

[5] Namias, V. The fractional order Fourier transform and its application to quantum mechanics. *J. Inst. Math. Appl.,* 1980, vol. 25, pp. 241-265.

[6] Almeida, L.B. The Fractional Fourier Transform and Time-Frequency Representations. *IEEE Transactions on signal processing.* 1994, vol. 42, núm. 11, pp. 3084-3091.

[7] Torres, R.; Pellat-Finet P. and Torres Y. Fractional convolution, fractional correlation and their translation invariance properties. *Signal processing.* 2010, vol. 90, núm. 6, pp. 1976-1984.

[8] Faúndez Z., M. *Tratamiento digital de voz e imagen y aplicación a la multimedia.* México: Marcombo, 2000. 288pp. ISBN: 978-8426712448.

[9] Gold B. and Morgan N. *Speech and audio signal processing.* New York: John Wiley & Sons, first edition, 1999. 537pp. ISBN: 978-8126508228.

[10] Stevens, S.S.; Volkmann, J. and E. B. Newman, E.B. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America.* 1937, vol. 8, núm. 3, pp. 185–190.

[11] White, L.S. and King, S. *The EUSTACE speech corpus.* Centre for Speech Technology Research, University of Edinburgh. 2003. [web online]. <http://www.cstr.ed.ac.uk/projects/eustace>. [Consulta: 01-4-2011]

[12] Malcolm Slaney. *Auditory Toolbox version 2*. Interval Research Corporation. 1998. [web online]. <http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/>. [Consulta: 01-4-2011]

[13] Wang, N.; Ching P.C. and Lee, T. Robust Speaker Verification Using Phase Information of Speech. National Cheng Kung University. *The Proceedings of ISCLSP 2010, The 7th International Symposium on Chinese Spoken Language Processing.* Tainan & Sun Moon Lake, Taiwan, november 29 to december 3 de 2010. IEEE Conference Publications, pp. 483-487.

Edgar F. Maldonado Orduz recibió su título de Ingeniero Electrónico de la Universidad Industrial de Santander, Colombia, en el año 2011.

Master en Comunicaciones Móbiles de Télécom ParisTech, Francia, actualmente es consultor de telecomunicaciones en Paris, Francia.

Su área de interés actual son los sistemas de telecomunicaciones y sistemas OFDM.

David Daniel Bertel Mendoza recibió su grado de Ingeniero Electrónico de la Universidad Industrial de Santander en el año 2011.

Ha trabajado como asistente de investigación en el campo de la Radiopropagación y Servicios de localización en el grupo de investigación RadioGIS, Bucaramanga, Colombia, y en el área de tratamiento de tratamiento imágenes en Fraunhofer IPT, Aachen, Alemania.

Actualmente realiza tesis de Maestría en la RWTH Aachen University, Aachen, Alemania, en Carácterisacion de antenas activas.

Yezid Torres Moreno recibió su grado de Doctor en óptica y tratamiento de la señal en la Universidad de Franche Comté, Besançon, Francia en 1983. Se vinculó a la Escuela de Física de la Universidad Industrial de Santander, Bucaramanga, Colombia en 1984 donde es Profesor Titular de física.

Ha realizado varias estancias postdoctorales, en el Laboratoire d'Optique P.M. Duffieux, Besançon, France, le Centre d'Optique Photonique et laser COPL, Quebec, Canada, Laboratorio de Procesado de Imagenes, Terrassa, España, École Normale Supérieure de Télècommunications de Bretagne, Brest, Francia y la Florida Atlantic University, Boca Raton, USA.

Su campo de interés actual, en donde orienta su investigación, es el de los haces con momento angular orbital y las aplicaciones de la técnica de la transformada de Fourier de tiempo promedio.