ANÁLISIS DE UN SISTEMA PARA LA VERIFICACIÓN DE LOCUTORES, UTILIZANDO LA TRANSFORMACIÓN DE FOURIER DE ORDEN FRACCIONAL

ÉDGAR FERNANDO MALDONADO ORDUZ

DAVID DANIEL BERTEL MENDOZA

UNIVERSIDAD INDUSTRIAL DE SANTANDER

FACULTAD DE INGENIERÍAS FÍSICO MECÁNICAS

ESCUELA DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

BUCARAMANGA

2011

ANÁLISIS DE UN SISTEMA PARA LA VERIFICACIÓN DE LOCUTORES, UTILIZANDO LA TRANSFORMACIÓN DE FOURIER DE ORDEN FRACCIONAL

ÉDGAR FERNANDO MALDONADO ORDUZ

DAVID DANIEL BERTEL MENDOZA

Trabajo de Investigación para optar al título de Ingeniero Electrónico

Director:

Ph.D. Yezid Torres Moreno

Codirector:

M.Sc. Jaime Guillermo Barrero Pérez

UNIVERSIDAD INDUSTRIAL DE SANTANDER

FACULTAD DE INGENIERÍAS FÍSICO MECÁNICAS

ESCUELA DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

BUCARAMANGA

2011

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

TÍTULO: ANÁLISIS DE UN SISTEMA PARA LA VERIFICACIÓN DE LOCUTORES, UTILIZANDO LA TRANSFORMACIÓN DE FOURIER DE ORDEN FRACCIONAL[1]

AUTORES: Édgar Fernando Maldonado Orduz y David Daniel Bertel Mendoza.[2]

PALABRAS CLAVE: Espacio tiempo-frecuencia, MFCC, Representación de la voz, Tratamiento de señales, Transformación de Fourier fraccionaria, Verificación de locutores.

DESCRIPCIÓN

Dentro del ámbito del reconocimiento de personas por medio de características biométricas, la voz es sin duda la más natural de todas. Por otro lado, la implementación de aplicaciones que hacen uso del reconocimiento de voz, no necesita de infraestructura diferente, o por lo menos, no muy diferente a la que se tiene actualmente en telefonía y en sistemas de comunicación.

La extracción de características de la voz representadas por los coeficientes cepstrales de Mel y los coeficientes LPC, o una combinación de estas dos representaciones, es una técnica que ha producido buenos resultados en el reconocimiento y verificación del habla y del hablante, pero la caracterización aún no satisface los objetivos de un sistema de verificación seguro.

La transformación de Fourier fraccionaria (FrFT) es una generalización de la transformación de Fourier estándar (FT), donde la señal es representada en un espacio tiempo-frecuencia.

En este trabajo se analiza el efecto de utilizar la transformada de Fourier de fraccionaria como método de la caracterización de la voz en un sistema de verificación de locutores dependiente del texto basado en Coeficientes Cepstrales en las Frecuencias de Mel (MFCC). Para ello, la señal de voz es representada en el dominio fraccionario, tiempo-frecuencia, teniéndose la posibilidad de encontrar un dominio fraccionario que mejore la representación del locutor. La factibilidad para la verificación de un individuo por su voz es abordada en este espacio de representación.

[1] Proyecto de grado

[2] Facultad de Ingenierías Físico Mecánicas. Escuela de Ingeniería Eléctrica, Electrónica y Telecomunicaciones. Director: Ph.D. Yezid Torres Moreno. Codirector: M.Sc. Jaime Guillermo Barrero Pérez.

ABSTRACT

TITLE: ANALYSIS OF A SPEAKER VERIFICATION SYSTEM, USING FRACTIONAL FOURIER TRANSFORM[3]

AUTHORS: Édgar Fernando Maldonado Orduz and David Daniel Bertel Mendoza.[4]

KEYWORDS: Time-frequency space, MFCC, Speech representation, Signal traitment, Fractional Fourier Transform, Speaker verification.

DESCRIPTION

In the field of pattern recognition speech is undoubtedly the most natural signal. Besides, speech recognition-based applications do not need different networks because actual communication systems provide a network for speech signal processing.

The extraction of speech characteristics through Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) coefficients, or a combination of these representations, is a well-known technique that had produced good results in speech and speaker recognition and verification, but speech characterization does not satisfy secure verification system objectives yet.

The Fractional Fourier Transform (FrFT) is a generalization of the Fourier Transform (FT), which characterizes a signal in a time-frequency space.

.

In this work is analyzes the effect of use Fractional Fourier Transform as a speech parameterization method inside a text-dependent MFCC-based speaker verification system. For that, speech signal is represented in a time-frequency fractional domain, with the possibility of find a fractional domain where speech representation is better. The feasibility to verify a person through his voice is treated in this representation space.

[3] Degree project

[4] Faculty of Physics Mechanics Engineering. School of Electrics Engineering, Electronics Engineering and Telecommunications. Director: Ph.D. Yezid Torres Moreno. Codirector: M.Sc. Jaime Guillermo Barrero Pérez.

I. Introducción

El estudio de los sistemas biométricos dos dimensionales que se ha desarrollado hasta el día de hoy reviste cierta dificultad, en lo general correspondiente a la representación, análisis y procesamiento de imágenes. Este es el caso 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 no involucrar procesamiento bidimensional, considera una mezcla de características físicas y del comportamiento como la articulación de las palabras, el contexto y demás variables que se encuentran involucradas en el 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 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 en la verificación para un conjunto cerrado finito de hablantes. El presente artículo presenta 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 fraccionaria FrFT, como método para la caracterización de la voz mediante los MFCC. La FrFT es una generalización de la transformación de Fourier estándar que brinda la posibilidad de analizar una señal en un espacio tiempo-frecuencia. Con la introducción del orden de la transformación, una nueva variable en el proceso de verificación, es posible mejorar el proceso de corroborar el hablante [3].

II. Transformación de Fourier fraccionaria

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

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

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

con núcleo

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

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

Para $a = 1$ se encuentra que $K_a = 1$ y

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

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

A) Algunas propiedades

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

- Rotacion nula $R^0 = I$
- Coherencia con FT $R^{\pi/2} = F$
- Adición de Rotación $R^\beta R^\alpha = R^{\beta+\alpha}$
- Rotación 2π $R^{2\pi} = I$

- Propiedad de translación:

$$F_\alpha f(x+k) = \exp[-iksin\alpha(x + \tfrac{k}{2}cos\alpha)]\ F_\alpha(f)_{[x+kcos\alpha]} \quad (5)$$

- Regla de similitud

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

- Propiedad de la convolución[5]:

$$f^{a} * g = \exp(-ibt^{2}) \int_{-\infty}^{\infty} f(\tau)e^{ibt^{2}} g(t-)e^{ib(t-\tau)^{2}} d\tau, \qquad (7)$$

donde b=0.5 cot(0.5πa).

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 está rotada un ángulo de *aπ/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 observa en la Figura 1. La señal y el ruido se solapan tanto en el tiempo como en frecuencia, pero en un dominio fraccionario las señales podrían llegar a separarse totalmente [3].

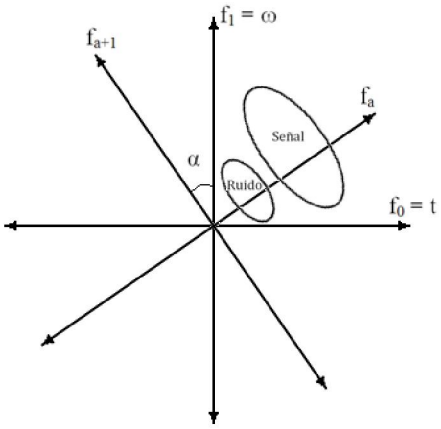

Figura 1. Rotación en un espacio tiempo-frecuencia. Fuente:[3].

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

III. Verificación de locutores

El reconocimiento del locutor es un término genérico para la clasificación de la identidad basándose en una señal acústica. Cuando se refiere a 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 la no pertenencia al grupo presente en el proceso de entrenamiento.

Para el caso de 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 buscarse decidir cuál de los P locutores es. La verificación de locutor es un caso particular de la identificación en un conjunto abierto con P = 1.

Los sistemas de verificación de locutores se pueden dividir en dos grupos: dependientes e independientes del texto. En los sistemas dependientes del texto se requiere que se pronuncien las mismas palabras usadas en el entrenamiento del sistema, mientras que en los independientes se puede usar cualquier texto, implicando una complejidad superior [8].

Generalmente el desempeño de un sistema de verificación de locutores se evalúa de acuerdo a dos tipos de tasas:

- 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 de evaluar el desempeño de un sistema de verificación de locutores con respecto a otros es utilizar una función de costo dada por

$$C = C_1 T_{FA} + C_2 T_{FR}, \qquad (8)$$

donde T_{FA} y T_{FR} corresponden a la tasa de falsa aceptación y a la tasa de falso rechazo; C1

y C2 corresponden a los pesos acordados a cada uno de estos parámetros [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 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 de 10%, lo cual no es despreciable en la práctica [8].

En la actualidad, los sistemas más difundidos que utilizan reconocimiento de voz poseen un bloque dedicado al cálculo de la transformación 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, espacio, 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.

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, pronunciándose 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 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 pueden compensar, pero que los actuales sistemas de reconocimiento no, haciéndolo 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 está 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 de verificación actuales

Para el proceso de reconocimiento, se divide la señal de voz en tramas 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 un proceso llamado 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 (*Linear Prediction Coding*) y Cepstrum, con sus respectivos derivados, son las características más usadas en el reconocimiento, siendo los últimos los más estables entre las pronunciaciones repetidas de una misma persona [8].

1) Filtros de 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 de Mel (de melodía) al espectro de la señal de voz y analizar los coeficientes obtenidos.

El filtrado de Mel aproxima el comportamiento del oído a una escala logarítmica de frecuencias representada por la siguiente función

$$f_{mel} = 2595 * \log(1 + \frac{f}{7\ 00\ Hz}). \quad (9)$$

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

2) Coeficientes Cepstrales en las frecuencias de Mel

Como producto de un procedimiento de filtrado (banco de filtros), y posteriormente una transformación Coseno Discreta DCT, se obtienen los coeficientes cepstrales en las frecuencias de Mel (MFCC – *Mel Frequency Cepstral Coefficients*). La transformación Coseno se realiza con el fin de disminuir la extensión de los vectores obtenidos a partir del filtrado de Mel. El esquema general utilizado para la obtención de los MFCC se observa en la 0Aunque existen otras técnicas, éstas no han sido ampliamente difundidas [2].

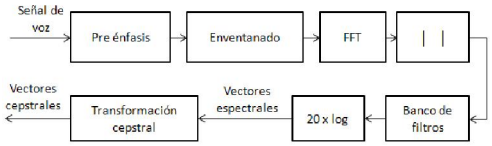

Figura 2. Obtención de los MFCC. Fuente:[2].

Los MFCC son los parámetros más utilizados y aceptados en lo referente a extracción de características del habla y del locutor. Los vectores generados por los coeficientes de Mel son utilizados en el entrenamiento y prueba de sistemas de reconocimiento y verificación [2].

El objeto de esta investigación se centra precisamente en los coeficientes de 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 introduce un grado de libertad adicional que la hace útil en otras aplicaciones.

3) Modelado estadístico.

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

- Paramétricos:
  - Redes neuronales (ANN – *"Artificial Neural Networks"*).
  - Modelos ocultos de Markov (HMM – *"Hidden Markov Models"*).
- No paramétricos:
  - Cuantificación vectorial (VQ – *"Vector Quantization"*).
  - Vecino más cercano (NN – *"Nearest neighbor"*).
  - Máquinas de vectores de soporte (SVM – *"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 pretende modelar.

El modelo no paramétrico, al ser menos restrictivo, puede permitir un mejor modelo 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].

4) *Vecino más cercano (Nearest Neighbor - NN)*

El método de vecino más cercano calcula la distancia 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 euclidiana 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 compara cuál es el locutor que ha proporcionado la menor distancia [8].

IV. Sistema basado en la FrFT

A) *Descripción del sistema*

Las pruebas se hicieron usando EUSTACE, 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 por muestra. Para el proceso de enventanado se toman 100 ventanas de 16 ms cada segundo. Finalmente cada una de las ventanas estará 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 MFCC. Para su obtención se utilizó el algoritmo provisto por el *Auditory Toolbox* de *Interval Research Corporation* [12], donde fue introducida la FrFT. Se utiliza un banco de cuarenta filtros para modelar el sistema de percepción auditivo humano.

Se utilizó la técnica de Vecino más cercano para evaluar el desempeño del sistema por las ventajas descritas en la sección de modelado estadístico.

B) *Uso de la FrFT en los MFCC*

Es necesario hacer claridad acerca de que la teoría de los bancos de filtros y de los MFCC se encuentra definida en el dominio de la frecuencia. Se habla de las implicaciones a causa de la inclusión de la FrFT sobre los MFCC para no perder de vista el esquema inicial, pero se debe resaltar que al no encontrarsen en un espacio de frecuencia, no se trata de los MFCC en dicho espacio, sino en el espacio tiempo-frecuencia.

C) *Análisis del mejor dominio fraccionario: Evaluación del error y 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 paso de 0,1 como se observa en la TABLA I. 0Se hicieron pruebas con tres radios de aceptación para el modelo de decisión: la desviación estándar, 1,5 desviaciones estándar y 2 desviaciones estándar, como se evidencia de la TABLA I a la TABLA III.

TABLA I. Resultados con distintos órdenes. R = Desv

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
0,9	31,9	57,14	44,5
1,0	21,9	57,14	39,5

TABLA II. Resultados con distintos órdenes. R = 1,5*Desv

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
0,9	48,1	30,95	39,5
1,0	40,0	28,57	34,3

TABLA III. Resultados con distintos órdenes. R = 2*Desv

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
0,9	67,1	7,14	37,1
1,0	55,2	2,38	28,8

Como se observa los mejores resultados se obtienen para órdenes cercanos a 1. Este resultado se debe a que el filtrado de Mel está definido para un espacio en frecuencia, y de acuerdo con el trabajo realizado por Sarikaya et al. en [3], la FrFT aplicada a sistemas de reconocimiento del habla presenta mejores resultados en las inmediaciones del orden 1. Por lo tanto se realizaron pruebas en órdenes cercanos a 1 como se observa de la 0TABLA IV a la TABLA VI.

TABLA IV. Resultados con órdenes cercanos a 1. R = Desv

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

TABLA V. Resultados con órdenes cercanos a 1. R = 1,5*Desv

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

TABLA VI. Resultados con órdenes cercanos a 1. R = 2*Desv

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

Tomando los resultados de las tablas se observa 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. Debido a que el propósito de un sistema es brindar importancia tanto a la TFA y la TFR [2], el radio de decisión de 1,5 veces la desviación estándar es el que brinda desempeños similares en las dos medidas. Por lo tanto, de acuerdo a los resultados, tomar para el orden el intervalo [0,95 – 0,99] brinda el desempeño más adecuado para el funcionamiento del sistema.

D) Desempeño de otros sistemas

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. 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 poco 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].

E) Costo computacional

Dado que el único cambio que se presenta es la utilización de la FrFT con respecto a la FFT que dispone Matlab, se evalúa el costo computacional de la una con respecto a la otra. Cabe aclarar que el tiempo de ejecución depende de las características del equipo utilizado para realizar las pruebas y las condiciones a las que este 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 tradicional (con FFT) 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] para la transformación, 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 propósito de la presente investigación fue modificar la etapa de parametrización en un sistema de verificación de locutores con la introducción de la FrFT en reemplazo de la FFT. Se estudiaron los Coeficientes Cepstrales en las Frecuencias de Mel (MFCC) por ser el método de parametrización más usado para sistemas de reconocimiento. El método estadístico de decisión utilizado fue el de Vecino más cercano, que al ser un modelo no paramétrico se ajusta a las condiciones del presente trabajo.

Los resultados obtenidos muestran que debido a que los filtros de Mel están diseñados para ser usados en frecuencia, los menores porcentajes de HTER se obtienen con una FrFT de orden cercano a 1. Al realizar pruebas más detalladas en la proximidad del orden 1 se encontró que con un orden en el intervalo [0,95 – 0,99] se obtiene una mejora de alrededor del 1% con respecto a un sistema basado en FFT.

Debido a que el banco de filtros de Mel es obtenido de manera experimental, la FrFT permite realizar una sintonización que adapta los filtros a la señal de voz del locutor en cuestión.

Aunque la mejora es pequeña, 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.

A) Trabajo futuro

En búsqueda de mejorar los resultados aquí obtenidos se proponen las siguientes acciones:

1) Uso de la fase para la obtención de los MFCC

La mayoría de los sistemas de verificación del locutor como se evidenció anteriormente hacen uso exclusivamente de 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 el trabajo de verificación [13].

2) 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], puede ser útil para mejorar los resultados de la presente investigación.

3) Pruebas con texto independiente

Los resultados obtenidos aplican únicamente para sistemas texto-dependientes. Los sistemas texto-independientes no fueron revisados, por lo tanto sería pertinente analizar los resultados de utilizar la FrFT.

4) Pruebas con LPC

Los coeficientes LPC junto con los coeficientes de 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 pueda 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 sin resolver [2]. Esto motiva a que se pruebe el sistema basado en la FrFT con otras capas de procesamiento como la normalización de los datos.

VI. Agradecimientos

Agradecemos al grupo GOTS por el apoyo brindado, en especial al profesor Yezid Torres Moreno. De la misma manera agradecemos al profesor Jaime Barrero de la E3T y a los ingenieros Alfonso Rueda e Idriss Sandoval. Su apoyo fue muy valioso en la realización del trabajo de investigación.

VII. Referencias

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

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

[3] R. Srikaya, Y. Gao y G. Saon, "Fractional Fourier Transform features for speech recognition", IEEE ICASSP 2004 , pp. I: 529-532.

[4] H. Ozaktas, M. Kutay y Z. Zalevsky, "The Fractional Fourier Transform: with applications in optics and signal processing". Chichester: John Wiley & Sons, 2001.

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

[6] Luís B. Almeida, "The Fractional Fourier Transform and Time-Frequency Representations". IEEE Transactions on signal processing, vol 42, no. 11, pp. 3084-3091, 1994.

[7] R. Torres, P. Pellat-Finet y Yezid Torres, "Fractional convolution, fractional correlation and their translation invariance properties". Elsevier, Signal processing, vol. 90 (6), pp. 1976-1984, 2010.

[8] M. Faúndez, "Tratamiento digital de voz e imagén y aplicación a la multimedia". México: Marcombo, 2000.

[9] B. Gold y N. Morgan, "Speech and audio signal processing". New York: John Wiley & Sons, 2000.

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

[11] *The EUSTACE speech corpus*, L.S. White and S. King, Centre for Speech Technology Research, University of Edinburgh. 2003. [Online]. Disponible en: http://www.cstr.ed.ac.uk/projects/eustace

[12] *Auditory Toolbox version 2*, Malcolm Slaney, Interval Research Corporation. 1998. [Online]. Disponible en : http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/

[13] N. Wang, P.C. Ching y T. Lee, "Robust Speaker Verification Using Phase Information of Speech". 7th ISCSLP 2010, pp. 483-487.

VIII. Biografías

Edgar F. Maldonado Orduz. Recibió el título de Bachiller con especialidad en electrónica en el Instituto Técnico Superior Damaso Zapata, Bucaramanga, Santander, Colombia, en el año 2005.
Ingeniero Electrónico de la Universidad Industrial de Santander, Colombia, forma parte del Grupo de investigación en Óptica y Tratamiento de Señal GOTS, de la UIS.

David D. Bertel Mendoza. Recibió el título de Bachiller en el Colegio Gimnasio Cerromar, Riohacha, La Guajira, Colombia, en el año 2004.

Ingeniero Electrónico de la Universidad Industrial de Santander, Colombia, forma parte del Grupo de investigación en Óptica y Tratamiento de Señal GOTS, de la UIS.