Revista Cubana de Ciencias Informáticas
Vol. 12, No. 3, Julio-Septiembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 140-151
http://rcci.uci.cu

Tipo de artículo: Artículo de revisión
Temática: Reconocimiento de patrones
Recibido: 21/11/2017 | Aceptado: 19/03/2018

Método de extracción de rasgos robusto para un sistema de diarización

Method of robust feature extraction for a diarization system

Edward L. Campbell Hernández*, Gabriel Hernández Sierra, José R. Calvo de Lara

[1]Empresa DATYS, Calle 7a A # 21406 e/ 214 y 216, Playa, Ciudad Habana, CUBA

***Autor para correspondencia**: ecampbell@cenatav.co.cu

Resumen
Los Sistemas Automáticos de Reconocimiento de Locutores, son sistemas biométricos que permiten realizar la identificación y verificación de personas, empleando la voz como rasgo discriminatorio. Uno de los desafíos a superar durante el proceso de reconocimiento, ocurre cuando el flujo de audio a procesar presenta varios locutores, ya que es necesario tener conocimiento de la ubicación temporal de los segmentos de audio relativos a cada locutor, para poder comparar directamente dichos segmentos con las muestras de locutores almacenadas en la base de datos de enrolamiento. Los sistemas de diarización permiten ubicar temporalmente los segmentos de audio relativos a cada locutor, dando solución, al problema mencionado en el reconocedor. En este artículo se propone el empleo de una técnica de extracción de rasgos robusta como subconjunto del sistema de diarización, denominada Respuesta sin Distorsión de Variación Mínima Perceptiva, la cual demostró mayor robustez ante ruido que la técnica dominante en el estado del arte, los Coeficientes Cepstrales en las Frecuencias de Mel. Experimentalmente se demostró como el rasgo propuesto presenta un menor nivel de varianza con respecto a los rasgos mel, entre tramas limpias y sucias, sometiendo el audio a una relación señal ruido de 6 dB y 8dB respectivamente.

Palabras claves: diarización, rasgo robusto, respuesta sin distorsión de variación mínima perceptiva.

Abstract
Abstract: Automatic Speakers Recognition Systems are biometric systems that allow the identification and verification of people, using voice as a discriminatory feature. One of the challenges to overcome during the recognition process is when the audio flow to be processed has several speakers, since it's necessary to have knowledge of the temporal location of the audio segments relative to each speaker, in order to be able to directly compare those segments with the speaker samples stored in the enrollment database. The diarization system allow to define the audio regions that are associated to a same speaker, solving, the mentioned problem in the recognition process. In this article is proposes a robust feature extraction technique as subsystem of the diarization system, called Perceptive Minimum Variance Distortionless Response, which demonstrated greater robustness to noise than the dominant technique in state-of-the-art, Mel Frequency Cepstral Coefficients. Experimentally is demostrated as the feature proposed present a level less of variance compared with the mel feature, between clean and noisy frame, subjecting the audio to a signal noisy relation of 6 dB and 8 dB respectively.

Revista Cubana de Ciencias Informáticas
Vol. x, No. x, Mes x-y, z
ISSN: 2227-1899 | RNPS: 2301
Pág. x-y
http://rcci.uci.cu

Keywords: *diarization, perceptive minimum variance distortionless response, robust feature.*

Introducción

El proceso de diarización responde la pregunta " ¿Quién habla y cuándo? "; las aplicaciones del mismo son diversas, desde empleos comerciales como indexación de audio y transcripción rica, hasta aplicaciones de búsqueda criminal como el análisis forense. Dicho proceso consta de las siguientes etapas: pre-procesado, extracción de rasgos, segmentación, agrupamiento y etiquetado; existiendo tres campos de investigación: ambiente telefónico, radio difusión y grabación de reuniones (Hernández, 2016). El sistema tratado se perfila a telefonía, con ambiente no controlado; constituyendo una etapa de pre-procesamiento de un sistema automático de reconocimiento de locutores, con el objetivo general de definir las regiones de audio que pertenecen a un mismo locutor, para concluir con la identificación o verificación de los mismos empleando el sistema de reconocimiento. Como objetivo específico, se plantea el diseño e implementación de una técnica de extracción de rasgos robusta ante ruido.

Materiales y método

En la propuesta de subsistema, se plantea el empleo de una técnica de extracción de rasgos robustas denominada Respuesta sin Distorsión de Variación Mínima Perceptiva, la cual resalta las regiones de frecuencia asociadas a los formantes F1 y F2. El experimento se realizó sobre la herramienta Matlab 2015, y para la comprobación de robustez y simulación de ruido se empleó el toolkit FaNT. En la figura 1 se muestra el

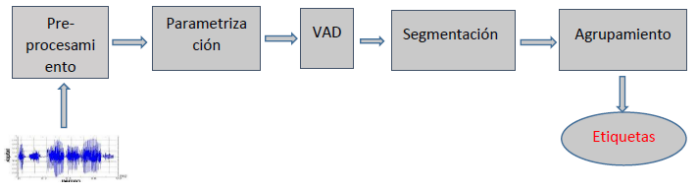


Figura 1. Diagrama funcional de un sistema de diarización.

diagrama funcional de un sistema de diarización. Antes de agrupar los segmentos de voz de locutor, se recomienda la extracción del i-vector de cada segmento, como modelos representativos. La metodología i-vector, permite obtener un vector de bajas dimensiones que aumenta la eficiencia de las técnicas de compensación de variabilidad de sesión (Hernández, 2014), ya que estas emplean una extensa metadata, que de utilizar vectores de altas dimensiones terminaría comprometiendo los recursos computacionales disponibles, algo que sucede frecuentemente al emplear un súper-vector, obtenido a partir de la concatenación de las medias de cada una

de las componentes del modelo de mezclas gaussianas obtenido a partir del segmento de locutor. Observación: el empleo de los i-vector también aumenta la eficiencia de la etapa de agrupamiento.

Pre-procesamiento

La etapa de pre-procesamiento se divide en pre-énfasis, enventanado y solapamiento. El pre-énfasis, se basa en procesar la señal de audio empleando un filtro de Respuesta Finita al Impulso (FIR) todo cero de primer orden (Woelfel, 2003), incurriendo en una amplificación de las muestras espectrales ubicadas en la región de altas frecuencias; este paso se realiza con el objetivo de disminuir el efecto de inclinación espectral[1] de las señales de voz, logrando así un espectro más plano que aumenta el poder discriminativo del proceso de extracción de rasgos, ya que en la región de altas frecuencias se encuentra la mayoría de la información relativa a la configuración del tracto vocal del locutor (Story, 2003), información que es amplificada a través del filtro. Cabe aclarar, que la configuración del tracto vocal de 2 locutores cualesquiera nunca va a coincidir, por eso dicha información es usada comúnmente para identificar o verificar locutores.
Debido a que una señal de voz es un proceso no estacionario, es necesario, para realizar un correcto procesamiento estadístico de la misma, acotarla a longitudes lo suficientemente pequeñas como para ser considerada un proceso cuasi-estacionario y que a su vez contenga información útil; generalmente se escogen ventanas de longitudes entre 20 y 30 milisegundos para el acotamiento. Sin embargo, este proceso no es suficiente, ya que entre ventanas consecutivas, existen discontinuidades como consecuencia del efecto de atenuación definido por la función transferencial de dichas ventanas, por lo que para evitarlas se solapan las ventanas contiguas una longitud de entre 20 y 10 milisegundos (Kondoz, 2004).

Extracción de rasgos

Para que un rasgo sea considerado como tal, este debe cumplir con las siguientes propiedades (Castro, 2010):

- Universalidad.
- Distintividad.
- Evaluabilidad.
- Estabilidad.

La voz tiene 6 niveles de información: espectral, prosódico, fonético, ideolectal, dialógica y semántica; de estos, el nivel espectral es el básico, obteniéndose a partir de él la ubicación de los formantes de la señal de voz, y por

[1] Disminución de la potencia espectral de la señal de voz a las altas frecuencias en relación a las bajas.

consiguiente, información sobre la configuración del tracto vocal del locutor. Este tiene la ventaja de emplear ventanas de menor longitud que los restantes niveles para poder extraer información útil de la voz (Ribas, 2016). Dentro de este nivel, las técnicas de extracción de rasgos más empleadas se basan en la transformación del dominio temporal de la señal, al dominio cepstral, poseyendo este último dominio las siguientes ventajas (Calvo et al., 2008):

- Cuando la ganancia de la señal varía, la forma de onda del espectro se preserva y solo se desplaza en amplitud.
- Un filtrado lineal causado por la acústica del local o por variaciones en la línea telefónica, tiene efectos convolucionales en la forma de onda y multiplicativos en el espectro de potencia, reflejándose como adiciones en el logaritmo del espectro de potencia, lo que trae como consecuencia un menor nivel de deformación de las características de la señal en el dominio cepstral.
- La distribución estadística del espectro en el dominio logarítmico tiene propiedades no presentes en el espectro de potencia lineal, que son convenientes en el reconocimiento del locutor y del habla.

En el sistema propuesto empleamos una técnica cepstral, denominada Respuesta sin Distorsión de Variación Mínima Perceptiva (PMVDR), la cual emplea como núcleo, una técnica de estimación de envolvente espectral robusta ante ruido denominada Respuesta sin Distorsión de Variación Mínima (MVDR).

Estimación espectral MVDR

La envolvente de predicción lineal, es uno de los método de estimación de envolvente más difundidos en la comunidad científica, sin embargo, tiene la desventaja de ofrecer una ineficaz parametrización de señales de voz de tonos medios y altos, debido a que sobrestima la potencia espectral de susodichas señales; dicha sobrestimación puede ser erradicada, empleando como técnica de estimación de envolvente la denominada Respuesta sin Distorsión de Varianza Mínima (MVDR), trayendo como consecuencia una mayor eficacia durante el proceso de parametrización de la señal (Murthi and Rao, 2000).
MVDR se basa en el diseño de un banco de filtros, sujeto a la condición de mínima distorsión, en la cual se establece una respuesta unitaria del filtro centrado en la frecuencia de interés, mientras que en las restantes frecuencias evita el paso, vea ecuación 1 (Woelfel, 2003).

$$H(e^{jw_{foi}}) = \sum_{k=0}^{M} h^*(k) e^{-jw_{foi}k} = 1, \tag{1}$$

Donde:

- h(k): respuesta al impulso del filtro sobre la muestra k.
- foi: frecuencia de interés.

Este método es sumamente trabajoso, debido a que requiere el diseño de un filtro para cada frecuencia de interés. Una alternativa a este es el empleo de un método paramétrico para definir la envolvente del Espectro de Varianza Mínima (o MVDR), definida por la siguiente expresión (Dharanipragada and Rao, 2001):

$$P_{MV}(w) = \frac{1}{\sum_{k=-M}^{M} u(k)e^{-jwk}}, \tag{2}$$

Donde:

- M: orden de estimación.
- u(k): parámetro definido a partir de los coeficientes de predicción lineal de la ventana de la señal, y se define a partir de la siguiente ecuación:

$$u(k) = \begin{cases} \frac{1}{P_e} \sum_{i=0}^{M-k} (M+1-k-2i) a_i a_{i+k}^*, & \text{para } k = 0, ..., M. \\ u^*(-k), & \text{para } k = -M, ..., -1, \end{cases} \tag{3}$$

Donde:

- P_e: error de varianza de predicción lineal.
- a_k: coeficiente de predicción lineal k.

MVDR tiene 3 propiedades básicas: banco de filtro, envolvente espectral y conexión directa con la predicción lineal. En función del orden de estimación espectral que se emplee, se puede variar el nivel de distorsión del espectro de potencia estimado, variando desde la representación del espectro en sí, hasta la representación de su envolvente (Murthi and Rao, 2000).

Revista Cubana de Ciencias Informáticas
Vol. x, No. x, Mes x-y, z
ISSN: 2227-1899 | RNPS: 2301
Pág. x-y
http://rcci.uci.cu

Técnica de extracción de rasgos propuesta, PMVDR

La técnica de extracción de rasgos más empleada en el estado del arte, son los Coeficientes Cepstrales en las Frecuencias de Mel (MFCC), la cual se basa en estimar el espectro de potencia de la ventana de la señal, para posteriormente transformar dicho espectro a la escala de frecuencias de Mel, finalizando en una transformación al dominio cepstral (Calvo et al., 2008); la desventaja de este método, es que el banco de filtros Mel empleado para distorsionar la escala de frecuencia lineal de la señal, tiene la propiedad de aumentar la separación de los filtros ubicados a las altas frecuencias, disminuyendo así la resolución y la eficacia de la parametrización (Ghosh et al., 2012), ya que se pierde parte importante de la información relativa a la configuración del tracto vocal del locutor, ubicada en las altas frecuencias.
Como solución, se propone el empleo de un método de extracción de rasgos que aplica una técnica de distorsión perceptiva vía interpolación, que en comparación con el banco de filtros mel, posee una mayor resolución a las altas frecuencias, ya que las muestras espectrales estimadas, se encuentran uniformemente espaciadas sobre todo el espectro (Yapanel and Hansen, 2008), además, emplea una técnica de estimación espectral robusta (MVDR). La figura 2 muestra el diagrama funcional del método propuesto, denominado Respuesta sin Distorsión de Variación Mínima Perceptiva (PMVDR).

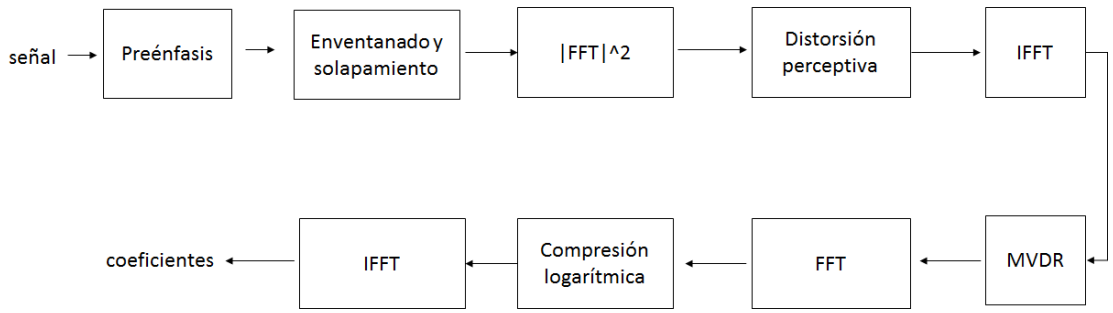


Figura 2. Diagrama funcional de la técnica PMVDR.

Distorsión directa (Yapanel and Hansen, 2008):

El objetivo del método de distorsión directa vía interpolación (distorsión perceptiva) empleado se basa en, obtener a partir del espectro de potencia de la ventana espaciado linealmente (w), el espectro de potencia distorsionado (w_d), cumpliendo la siguiente relación:

$$w_d = \tan^{-1} \frac{(1-\alpha^2)sin(w)}{(1+\alpha^2)cos(w) - 2\alpha}. \quad (4)$$

Dicha relación se puede garantizar con el empleo de un sistema pasa todo de primer orden que cumpla con la siguiente condición:

$$H(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \qquad |\alpha| < 1, \tag{5}$$

la variable alfa controla el grado de distorsión; para señales muestreadas a 16 khz se recomienda el empleo entre el rango de 0.42 y 0.55, mientras que para 8 khz se recomienda entre 0.31 y 0.42.
¿Cómo obtener el espectro distorionado?

1. Obtener el espectro de potencia de la ventana de entrada de longitud N a través de FFT, N debe ser seleccionada como la potencia más cercana posible de 2, lo cual provee N puntos espectrales en un espacio de potencia espectral lineal.

2. Calcular N puntos espectrales linealmente espaciados sobre el espacio de frecuencia deformada con una separación de enteros de 2π entre puntos adyacentes:

$$w_d[i] = \frac{2i\pi}{N}, \qquad i = 0, ..., N-1. \tag{6}$$

3. Hallar las frecuencias lineales y el índice de las FFT correspondientes al espectro deformado usando:

$$w[i] = tan^{-1}\frac{(1-\alpha^2)sin(w_d[i])}{1+\alpha^2 cos(w_d[i]) + 2\alpha}, \qquad i = 0, ..., N-1. \tag{7}$$

$$k_d[i] = \frac{w[i]N}{2\pi}, \qquad i = 0, ...N-1. \tag{8}$$

4. Interpolar los valores espectrales lineales más cercanos para obtener el valor espectral deformado:

$$k_l[i] = min(N-2, k_d[i]), \qquad i = 0, ..., N-1. \tag{9}$$

$$k_u[i] = max(1, k_l[i] + 1), \qquad i = 0, ..., N-1. \tag{10}$$

$$S_d[i] = (k_u[i] - k_d[i])S[k_l[i]] + (k_d[i] - k_l[i])S[k_u[i]] \tag{11}$$

Ajustes de PMVDR: factor de distorsión de 0.57 utilizando los primeros 12 coeficientes cepstrales, excluyendo el de orden cero. Desplazamiento entre tramas de 10 ms, con una longitud de ventana de 20 ms, y filtro de pre-énfasis con cero de 0.95.

Detección de la actividad de la voz

La función del detector de la actividad de la voz (VAD), como bien se infiere por su nombre, es la detección de los segmentos de voz; siendo la voz, el acto individual del ejercicio del lenguaje, producido al elegir determinados signos, entre los que ofrece la lengua mediante su realización oral (Huijbregts, 2008). El VAD es un elemento crucial del pre-procesamiento en el marco de los sistemas de diarización, pues aumenta la eficacia de la etapa de segmentación al garantizar el procesamiento de un flujo continuo de voz (Hernández, 2016).
Las tramas eliminadas, clasificadas como no voz, pueden estar compuestas por silencio, música o ruido. Los clasificadores de máxima-verosimilitud son el enfoque más usado en la detección de la voz, empleando modelos de mezclas gaussianas (GMM) pre-entrenados a partir de tramas de habla y no habla para modelar clases acústicas (Hernández, 2016), dichas clases varían de un sistema a otro, pudiéndose emplear clases que representen la voz, ruido, silencio y música respectivamente. Los detectores de energía son elementos que se pueden emplear como VAD, pero su eficacia depende del nivel de energía de las tramas de no voz (Huijbregts, 2008).

Segmentación

Durante el proceso de segmentación, se determinan los puntos de cambios de locutor; representando puntos consecutivos, segmentos de audio relativos a un único locutor. Los métodos más empleados en el estado del arte se basan en métricas como el Coeficiente de Información Bayesiana (BIC), o en modelos como los Modelos de Mezclas Gaussianas (Hernández, 2016).

Agrupamiento

Posterior a la segmentación, luego de haber sido determinados los segmentos relativos a un único locutor, y definidos a su vez los segmentos de voz mediante la aplicación del VAD, se agrupan los segmentos de habla pertenecientes a un mismo locutor en un mismo grupo, respondiendo el proceso descrito al nombre de agrupamiento. Entre los métodos más empleados se encuentran los basados en Máquina de Vectores de Soporte, como clasificadores; y el agrupamiento jerárquico (Hernández, 2016).

Revista Cubana de Ciencias Informáticas
Vol. x, No. x, Mes x-y, z
ISSN: 2227-1899 | RNPS: 2301
Pág. x-y
http://rcci.uci.cu

Resultados y discusión

En este epígrafe se llevará a cabo un análisis de los resultados obtenidos a través del empleo de la técnica de extracción de rasgos propuestas, haciendo énfasis en el comportamiento de la misma en ambientes ruidosos.

Envolvente espectral

Como habíamos precisado previamente, la estimación de la envolvente espectral de la señal de habla, mediante el método de predicción lineal, genera una sobrestimación de potencia espectral, que se puede compensar empleando la envolvente de varianza mínima; fenómeno que se evidencia en la figura 3, a partir de un experimento realizado en (Murthi and Rao, 2000), en donde se compararon las envolventes espectrales obtenidas a partir del método de Varianza Mínima y el de predicción lineal, de orden 19 respectivamente. Comprobamos la vera-

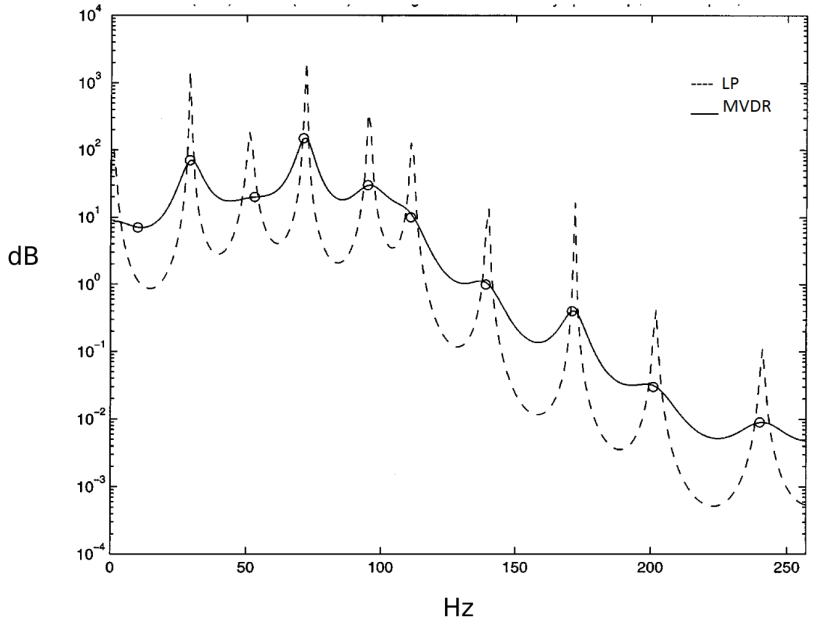


Figura 3. Estimación espectral empleando MVDR y LPC.

cidad de este planteamiento a partir del análisis de 100 segmentos de habla de 25 señales de la base de datos Fisher [2], concluyendo, cómo la propuesta de empleo de la técnica de Varianza Mínima incide directamente en la disminución de los picos sobrestimados de potencia, obtenidos a partir de señales de voz de tonos altos y medios, debido al efecto de suavizado del método MVDR sobre la envolvente del espectro de la señal (vea figura 4).

[2]para informarse sobre Fisher visite https://catalog.ldc.upenn.edu/LDC2004S13

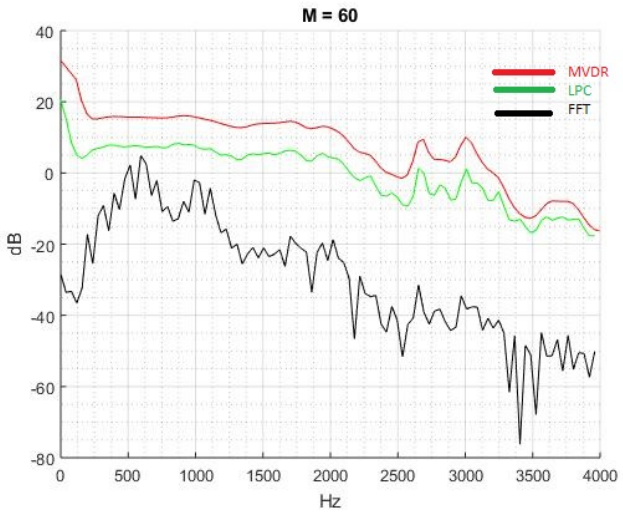
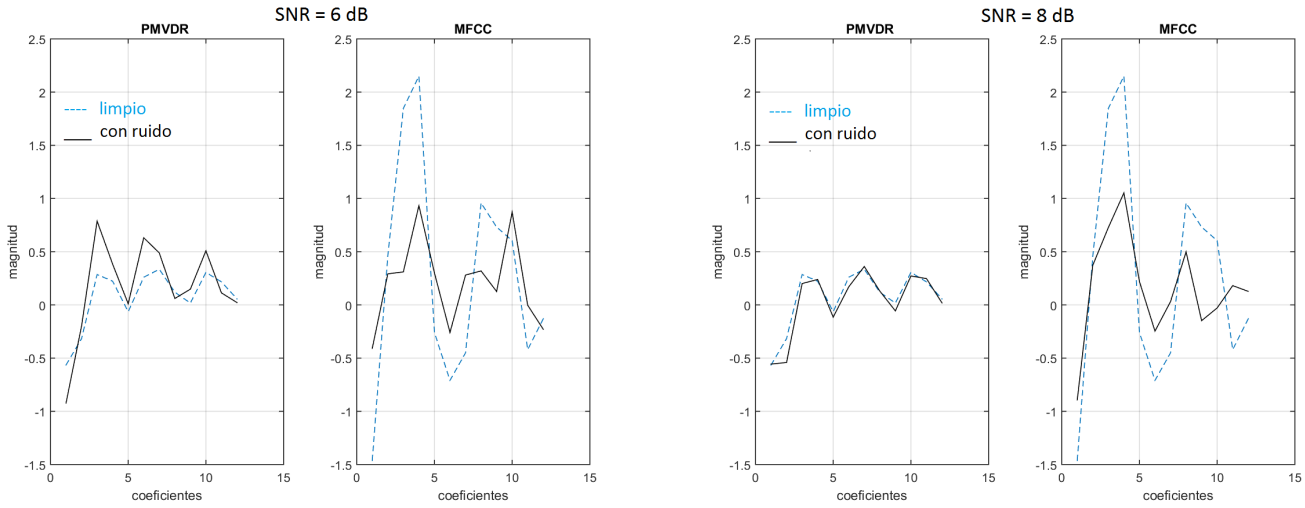


Figura 4. Comprobación de efecto de suavizado de MVDR. .

Robustez ante ruido

Un factor de impacto de los rasgos acústicos es la variabilidad de estos ante el ruido, debido a las múltiples condiciones acústicas no controlables a las que en la práctica se exponen los sistemas de diarización en ambiente telefónico. En el experimento realizado (figura 5), PMVDR demostró poseer mayor robustez que MFCC ante ruido.

La figura 5 muestra la variación de los primeros 12 coeficientes cepstrales de PMVDR y MFCC extraídos a partir de la ventana de una señal limpia telefónica, luego de agregar ruido blanco, estableciendo una relación señal ruido de 8 y 6 dB respectivamente en cada experimento, evidenciándose un nivel inferior de variación de los coeficientes de PMVDR ante ruido blanco en comparación con MFCC.



((a)) SNR=6dB ((b)) SNR=8dB

Figura 5. Comparación de robustez

Conclusiones

El método de extracción de rasgos, PMVDR, es un método propuesto a partir de la necesidad de suprimir las deficiencias sobre señales de voz de mediano y alto tono, que presentaba la predicción lineal (LPC), demostrándose experimentalmente, como se lograba una disminución de los picos sobre-estimados de potencia al aplicar el espectro de varianza mínima (MVDR) sobre la predicción lineal. Debido a que la investigación se perfiló a ambiente telefónico no controlado, era necesario que los rasgos propuestos fueran estables bajo estas condiciones, por lo que a la señal de audio se le adicionó ruido blanco, y se comparó el nivel de varianza alcanzado entre PMVDR y los rasgos mel bajo estas condiciones; manifestándose como la variación entre las tramas limpias y sucias de PMVDR, era inferior a la de los rasgos mel. A partir del análisis hecho, se puede concluir, como PMVDR es un rasgo robusto; y específicamente, MVDR, es una técnica que permite obtener una buena estimación de los formantes, no solo de señales de voz de tono bajo, sino también de tonos medios y altos.

Dado lo expuesto en este trabajo, recomendamos extender la investigación sobre las restantes etapas del sistema de diarización, y comprobar la compatibilidad, de los métodos empleados en conjunto, para la confección del sistema.

Referencias

J. R Calvo, R. Fernández, and G. Hernández. Métodos de extracción, selección y clasificación de rasgos acústicos para el reconocimiento del locutor. Technical Report RT_08, Serie Azul: Reconocimiento de Patrones, CENATAV-DATYS, Siboney, Playa, La Habana, Cuba, February 2008.

A. H. Castro. Fiabilidad en sistemas forenses de reconocimiento automático de locutor explotando la calidad de la señal de voz, 2010.

S. Dharanipragada and B. D. Rao. MVDR based feature extraction for robust speech recognition. pages 309–312, Salt Palace Convention Center, Salt Lake City, Utah, USA, May 7-11 2001. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2001.

D. Ghosh, D. S. Debnath, and S. Bose. A comparative study of performance of FPGF based mel filter bank bark filter bank. *CoRR*, 2012.

Gabriel Hernández. *Métodos de representación y verificación del locutor con independencia del texto.* PhD thesis, Instituto Superior Tecnológico José Antonio Echeverría, Ciudad de La Habana, 2014.

Gabriel Hernández. Diarización de locutores sobre señales telefónicas. Technical Report RT_081, Serie Azul: Reconocimiento de Patrones, CENATAV-DATYS, Siboney, Playa, La Habana, Cuba, February 2016.

M. Huijbregts. *Segmentation, diarization and speech transcription: surprise data unraveled.* PhD thesis, University of Twente, Enschede, Netherlands, 2008.

A.M. Kondoz. *Digital Speech. Coding for Low Bit Rate Communication Systems.* University of Surrey, UK, 2nd edition, 2004.

M. N Murthi and B. D. Rao. All-pole modeling of speech based on the minimum variance distortionless response spectrum. *IEEE Trans. Speech and Audio Processing*, (3):221–239, 2000.

D. Ribas. *Reconocimiento robusto de locutores en ambientes no controlados.* PhD thesis, Instituto Superior Politécnico José Antonio Echeverría, Facultad de Ingeniería Eléctrica, La Habana, Cuba, 2016.

B. H. Story. Using imaging and modeling techniques to understand the relation between vocal tract shape to acoustic characteristics. Stockholm, Sweden, August 6-9 2003. Stockholm Music Acoustics Conference.

M. Woelfel. Minimun variance distorsionless response spectral estimation and subtraction for robust speech recognition, 2003.

U. H. Yapanel and J. H. L. Hansen. A new perceptually motivated mvdr based acoustic front-end (pmvdr) for robust automatic speech recognition. *Speech Communication*, (2):142–152, 2008.