RNPS No. 2142
ISSN 2072-6287
Versión Digital

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO
Reconocimiento de Patrones

SERIE AZUL

Utilización de nuevos rasgos dinámicos del habla para el reconocimiento del locutor

José R. Calvo de Lara, Rafael Fernández Torres,
Gabriel Hernández Sierra, y
Dayana Ribas González

RT_040 **febrero 2011**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO
Reconocimiento de Patrones

Utilización de nuevos rasgos dinámicos del habla para el reconocimiento del locutor

José R. Calvo de Lara, Rafael Fernández Torres,
Gabriel Hernández Sierra, y
Dayana Ribas González

RT_040 **febrero 2011**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Tabla de contenido

Utilización de nuevos rasgos dinámicos del habla para el reconocimiento del locutor

José R. Calvo de Lara, Rafael Fernández Torres, Gabriel Hernández Sierra, y Dayana Ribas González

Dpto. Reconocimiento de Patrones, Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV), Ciudad de La Habana, Cuba
{jcalvo,gsierra,dribas}@cenatav.co.cu

RT_040, Serie Azul, CENATAV
Aprobado: noviembre de 2010

Resumen. Este reporte introduce el uso del rasgo cepstral delta desplazado (SDC: "shifted delta cepstral") de la señal del habla en aplicaciones de reconocimiento del locutor, como una alternativa al coeficiente cepstral en escala Mel (MFCC: "Mel- frequency cepstrum coefficient"), debido a su robustez a las desigualdades del canal, robustez al ruido y a su carácter seudo-prosódico. Otra ventaja del rasgo SDC es su eficiencia computacional: los bloques SDC se seleccionan y concatenan directamente a partir del rasgo delta "Δ " del coeficiente MFCC, sin costo computacional adicional. En el trabajo se recogen los resultados experimentales obtenidos por los autores, en la evaluación del rasgo SDC en el reconocimiento del locutor, que ya han sido publicados.

Palabras clave: rasgo cepstral delta desplazado, SDC, reconocimiento del locutor.

Abstract. This technical report introduces the use of shifted delta cepstral feature (SDC) of speech signal in speaker recognition applications as an alternative of Mel-frequency Cepstrum coefficients (MFCC) due its robustness to channel mismatch and noise, and its pseudo-prosodic behavior. Another advantage of SDC feature is its computational efficiency; the SDC feature blocks are selected and concatenated directly from Δ feature under the control of SDC parameters, without additional computational cost. This work contains experimental results obtained by the authors in the evaluation of the SDC in speaker recognition, which have been published.

Keywords: shifted delta cepstral feature, SDC, speaker recognition.

1 Introducción

La señal del habla conlleva varios niveles de información. A nivel primario, el habla lleva un mensaje semántico vía las palabras expresadas, a otro nivel, el habla contiene información referida al idioma en que se habla, y a las emociones, el estado de salud, el género y la identidad del locutor.

Mientras el reconocimiento del habla se dirige a reconocer las palabras habladas y el reconocimiento del lenguaje a identificar el idioma o dialecto utilizado, la meta de los sistemas de reconocimiento automático del locutor es extraer, caracterizar, y reconocer la información en la señal del habla que conlleva la identidad del locutor.

El reconocimiento del locutor, como método de reconocimiento de patrones, se lleva a cabo en dos etapas: una previa donde se entrenan los modelos de los locutores objetivos y una etapa de prueba,

donde se establece la similitud entre la muestra de origen desconocido y los modelos previamente entrenados.

El reconocimiento del locutor se divide en dos grandes áreas, en dependencia de la aplicación: la identificación del locutor, donde se establece una comparación de 1:N y se brinda como resultado una lista de los locutores más probables de haber expresado la muestra de habla de origen desconocido, este es el caso de la identificación forense del locutor; y la verificación del locutor, donde se establece una comparación 1:1 entre una muestra de habla de un locutor que clama ser reconocido y su modelo, aceptándose si la similitud sobrepasa un umbral, como en la verificación biométrica del locutor.

La variabilidad del locutor, constituye el principal reto a enfrentar en su reconocimiento, dicha variabilidad, según Acero [1] puede clasificarse en:

- Variabilidad propia intra-locutor: determinada por la variabilidad entre sesiones debido a cambios en la manera de hablar, en el entorno comunicativo o en las condiciones emocionales y de salud, envejecimiento, ocurrencias de variaciones dialectales, intentos de disfrazar la voz, etc.
- Variabilidad forzada intra-locutor: causada por efecto lombard (hablar en condiciones acústicas desfavorables), hablar bajo estrés provocado por influencias externas o por efecto cocktail-party (habla mezclada de varios locutores con ruido ambiente).
- Variabilidad por influencias externas dependientes del canal: tipo de micrófono, reducción del ancho de banda y rango dinámico de la voz, ruido eléctrico y acústico, reverberación y distorsión.

Las principales investigaciones que se llevan a cabo se dirigen a extraer rasgos del habla, que sean robustos ante la variabilidad del locutor. Desde el punto de vista perceptual, los rasgos cepstrales MFCC [2] reflejan mejor el comportamiento del sistema auditivo humano al tener en cuenta la naturaleza no lineal de la percepción del tono fundamental, así como la percepción no lineal de la intensidad, lo que los hace más adecuados para el reconocimiento del habla que otros como Este comportamiento, unido a su cálculo robusto y efectivo (ver Anexo 1), los ha convertido en un estándar para dichas investigaciones.

Los rasgos prosódicos, como la frecuencia del tono fundamental y la energía, tanto en sus valores estáticos y dinámicos, brindan información útil acerca de los hábitos del habla dependientes del locutor, siendo usados en el reconocimiento del locutor por Adami et al. [3]. No obstante la estimación de la frecuencia del tono fundamental aun adolece de dificultades en entornos reales.

Reynolds et al [4] reportan la utilización de nuevos rasgos de información de alto nivel, incluyendo los prosódicos, fonéticos, idiolectales, semánticos y lexicológicos, aplicando diversos métodos de clasificación para cada rasgo y la fusión posterior de los resultados de los clasificadores, los resultados obtenidos con los nuevos rasgos de alto nivel superan los obtenidos hasta el momento con rasgos cepstrales. Estos rasgos son robustos ante el ruido y las distorsiones del canal pero requieren del uso de métodos de reconocimiento del habla para obtener las secuencias de fonemas y /o palabras, y de un volumen apreciable de muestras de voz para estimar los modelos fonéticos y de lenguaje, lo cual incorpora complejidad a los sistemas de reconocimiento del locutor.

Diferentes estudios se han hecho para utilizar la información dinámica contenida en la señal de habla. La extracción conveniente de rasgos dinámicos del habla tiene un efecto significativo en la clasificación de un sistema de reconocimiento del locutor. La solución más popular consiste en obtener las derivadas en el tiempo de cada coeficiente cepstral, conocidas como rasgo delta Δ y delta -delta "$\Delta\Delta$", y añadirlas al vector cepstral. Furui [5] propuso la combinación del rasgo cepstral y sus rasgos Δ y $\Delta\Delta$ para la verificación de locutores y estableció la efectividad de combinar rasgos temporales y dinámicos. Estudios posteriores de Bernasconi [6] y Soong y otros [7] confirmaron que la unión de la información estática y dinámica cepstral puede llevar a una mejora en la robustez de la verificación del locutor.

Los autores proponen y evalúan en este reporte el uso del rasgo cepstral delta desplazado "SDC: shifted delta cepstral" como un rasgo dinámico de larga duración, en el reconocimiento del locutor siendo, al menos en nuestro conocimiento, el primer intento de su uso en este campo. El rasgo SDC se obtiene concatenando rasgos Δ computados en varias tramas del habla y ha sido reportado por Torres-

Carrasquillo y otros [8, 9, 10, 11], que posee una efectividad de reconocimiento superior al rasgo Δ, en identificación de idioma y dialecto con clasificadores generativos y discriminativos. Más recientemente, otros autores [12, 13, 14], informan resultados similares en el mismo campo de aplicación.

Este reporte recoge los resultados obtenidos por los autores en la aplicación del rasgo SDC, como un nuevo rasgo dinámico en el reconocimiento del locutor:

- evaluando su comportamiento ante la variabilidad del canal y de la sesión, ante el modo de hablar y las desigualdades en el canal y el teléfono.
- evaluando su carácter seudo-prosódico.
- evaluando una selección de combinaciones de los mismos, a partir de su información mutua con la identidad del locutor.

2 Dinámica del rasgo cepstral

Estudios dedicados a la explotación de la información dinámica del habla en los sistemas de reconocimiento del locutor [5, 6,7, 15, 16, 17, 18] han demostrado la importancia de utilizar dicha información dinámica como otra fuente de identidad del locutor, obteniendo un comportamiento mejor al combinarlo con la información estática, siendo más robustos en ambientes ruidosos y más resistentes a la variabilidad del canal.

El uso del rasgo cepstral dinámico es ampliamente aceptado para extraer la estructura temporal dinámica del habla. Estos rasgos pueden interpretarse como estimaciones de la primera y segunda derivada temporal -rasgo delta "Δ" y delta -delta "ΔΔ" - de la trayectoria temporal del coeficiente cepstral estático.

La trayectoria temporal de cada coeficiente cepstral normalmente no tiene una expresión analítica y su derivada en el tiempo sólo puede aproximarse por una diferencia finita, pero la diferencia finita de primer orden es intrínsecamente ruidosa. Para superar esta dificultad, Furui [5] sugiere el uso de un ajuste polinómico ortogonal de la trayectoria temporal de cada coeficiente cepstral sobre una ventana de tiempo *h* finita. El coeficiente de orden cero o término constante de este polinomio es:

$$c(t) = \frac{\sum_{d=-D}^{D} h_d c(t+d)}{\sum_{d=-D}^{D} h_d} \tag{1}$$

donde h es una ventana simétrica de longitud 2D + 1 tramas temporales del coeficiente cepstral.

Entonces el coeficiente polinómico ortogonal de primer orden, o la pendiente espectral generalizada en el tiempo, se denota como:

$$\Delta c(t) = \frac{\sum_{d=-D}^{D} d h_d c(t+d)}{\sum_{d=-D}^{D} h_d d^2} \tag{2}$$

Furui [1] demuestra que la caracterización polinómica de primer orden de los cambios espectrales es adecuada para una representación eficaz de la trayectoria dinámica del coeficiente cepstral sobre un segmento corto del habla. Debe usarse una ventana rectangular (h_d =1) de una longitud razonable para asegurar un ajuste suave a los valores entre una trama y la próxima [7]. Los rasgos Δ y ΔΔ normalmente

han sido calculados usando la ec. (2), utilizando entre 5 y 11 tramas temporales de cada coeficiente cepstral, ($2 \leq D \leq 5$) dependiendo de la longitud de tiempo de la trama [5, 7, 18].

2.1 El rasgo cepstral delta desplazado "SDC"

Propuesto por Bielefeld [19], SDC es un rasgo dinámico cepstral de larga duración, obtenido concatenando en un vector varios vectores de rasgos Δ a través de múltiples tramas de información cepstral del habla y agregándolo al vector cepstral. El vector de rasgos SDC se especifica por 4 parámetros, (*N, D, P, k*) donde:

- *N*: dimensión del vector de coeficientes cepstrales en cada trama
- *D*: adelanto y atraso en tiempo para cómputo del rasgo Δ
- *P*: separación entre vectores Δ a concatenar.
- *k*: número de vectores Δ que se concatenan para formar el vector SDC.

El cálculo del vector SDC es un procedimiento relativamente simple [11, 12] y se ilustra en la Fig. 1:

- se computa el vector cepstral N-dimensional en cada trama de habla *t.*
- se obtiene el vector Δ en cada trama *t* utilizando t ± *D* tramas.
- se concatenan *k* vectores Δ, espaciados *P* tramas, para formar el vector SDC en cada trama *t.*

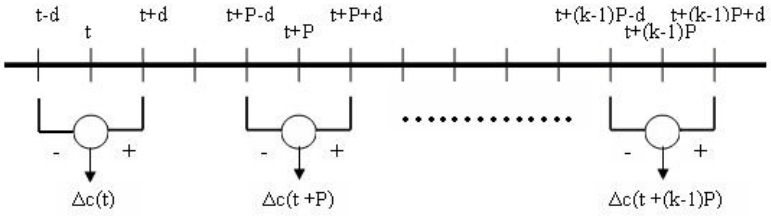


Fig. 1. Computación del vector SDC

Para el caso mostrado en la Fig.1 el rasgo SDC para cada coeficiente cepstral *c* en la trama *t* se obtiene de la concatenación de *i* = 0 a *k*-1 de todos los Δ*c* (*t* + *iP*):

$$\Delta c(t+iP) = \frac{\sum_{d=-D}^{D} dc(t+iP+d)}{\sum_{d=-D}^{D} d^2} \qquad (3)$$

La ec. (3) es una generalización del coeficiente polinómico ortogonal de primer orden definido por Furui [5]en la ec. (2), con $h_d = 1$ e incluyendo el desplazamiento en tiempo *iP* entre bloques, donde *i* = 0 a *k*-1.

Se requieren kN parámetros para cada vector SDC, que comparado con 2N parámetros para el vector de rasgos cepstral y Δ, implica un crecimiento de la dimensionalidad del vector. No obstante, se observa que para obtener el vector SDC no se requiere costo computacional adicional del requerido para obtener el vector Δ.

2.2 Extracción de rasgos cepstrales y dinámicos: obtención del vector SDC

El método más común para representar el espectro del habla a corto término lo constituyen los rasgos cepstrales normalizados derivados de un banco de filtros de frecuencias Mel (MFCC) [2], el proceso de su obtención se describe en el Anexo 1. Los rasgos Δ pueden obtenerse usando la ec. (2) con $h_d = 1$, para cada coeficiente cepstral, conformándose un vector de rasgos Δ de dimensión N.

Como los rasgos SDC pueden obtenerse por la ec. (3), que es una generalización de la ec. (2), no se requiere ningún cómputo adicional si se usa la misma D para ambos rasgos. Sólo es necesario seleccionar k-1 vectores de rasgos Δ adicionales, separados P tramas y concatenarlos con el vector de rasgos Δ de la trama; para cada coeficiente cepstral c, su correspondiente rasgo SDC se conformar ía:

$$SDC(c, D, P, k) = \Delta\, c(t),\ \Delta\, c(t+P),....., \Delta\, c(t+(k-1)P)\ . \qquad (4)$$

3 La base de datos Ahumada

Para evaluar el comportamiento de los rasgos SDC en el reconocimiento del locutor se utiliza la base de datos en idioma español Ahumada [22], utilizada en la evaluación NIST 2001. Ahumada es una base de datos de habla de 104 locutores masculinos españoles, diseñada y adquirida bajo condiciones controladas, para la caracterización e identificación del locutor, que incorpora varios factores de variabilidad del habla. Cada locutor en la base de datos expresa cinco tipos de expresiones (sucesiones de dígitos, frases balanceadas, texto fonológicamente balanceado y aleatorio, y habla espontánea) en siete sesiones de micrófono y tres sesiones de teléfono, con un intervalo de tiempo entre ellas.

Las frases fonológica y silábicamente balanceadas, de 8 a 10 palabras cada una, garantizan una buena representación del idioma español, porque estas frases tienen un coeficiente de correlación fonético de 0.9966 y un coeficiente de correlación silábico de 0.9963, ambos respecto al modelo del idioma español [22]. Estas frases son las mismas para cada uno de los 104 locutores, lo que representa un desafío adicional porque cualquier experimento de reconocimiento de locutores utilizando las frases de dicha base, requerirá reconocer la identidad de un locutor objetivo entre todos los impostores, usando las mismas frases para todos los locutores.

4 Evaluación de los resultados con Curva DET: EER y mínimo de DCF

La evaluación de los resultados del reconocimiento del locutor se lleva a cabo utilizando la curva de Comportamiento del Error de Detección (Detection Error Tradeoff: "DET") propuesta por el NIST en 1997 [23]. Esta curva se obtiene agrupando todos los resultados obtenidos de la clasificación de los objetivos y de los impostores, y evaluando la probabilidad de Falsa Aceptación (FA) y de Falso Rechazo (FR), respecto a un umbral de clasificación variable.

Se usan dos indicadores para evaluar el comportamiento: la razón de igual error ("EER: Equal Error Rate"), donde coinciden FA y FR, y el mínimo de la función de costo de detección ("DCF: Detection Cost Function"), que se obtiene:

$$DCF = CFR * FR * P_{Target} + CFA * FA * P_{NonTarget}\ . \qquad (5)$$

donde:
CFR: costo de una detección perdida (de un falso rechazo)
CFA: costo de una falsa alarma (de una falsa aceptación)
P_{Target}: probabilidad a priori locutor objetivo

$P_{NonTarget:}$: probabilidad a priori locutor impostor

Siguiendo también la recomendación de NIST [23], estos parámetros son: *CFR = 10, CFA = 1, PTarget = 0.01 y PNonTarget = 0.99*

5 Experimentos para evaluar robustez de los rasgos SDC ante la variabilidad del canal y la sesión, así como ante la manera de hablar

Los primeros experimentos realizados en el año 2007, fueron dirigidos a evaluar el rasgo SDC en el reconocimiento remoto (por teléfono) del locutor ante la variabilidad del canal y la sesión, así como ante la manera de hablar. Se llevaron a cabo dos experimentos:

1. Experimento con expresiones cortas: usa una misma secuencia de diez dígitos (D) y una misma frase fonológicamente equilibrada de 8 a 12 palabras de longitud (P). Este experimento usa un clasificador con cuantificación vectorial "VQ: vectorial quantization" con el algoritmo LBG (Linde, Buzo,Gray) con 64 centroides [24].
2. Experimento con expresiones largas: usa un mismo texto fonológicamente balanceado de aproximadamente 180 palabras (R) y más de un minuto de habla espontánea descriptiva (S). Este experimento usa un clasificador de mezclas gaussianas "GMM: Gausssian Mixture Models" con 64 mezclas [25].

Cada uno de los experimentos hace una evaluación del rasgo SDC ante:

- Variabilidad de sesión para el mismo canal: el mismo micrófono en dos sesiones.
- Variabilidad de canal para la misma sesión: habla simultánea en dos canales diferentes, micrófono en el entrenamiento y teléfono para la prueba.
- Variabilidad de canal y de sesión: diferentes canales y sesiones, micrófono para el entrenamiento y teléfono para la prueba.

Una fuente adicional de variabilidad, referida a la independencia del texto, ocurre en el segundo experimento, cuando se entrena con texto balanceado (R) y se prueba con habla espontánea (S).

Se evaluaron tres variantes de rasgos:

- Lineabase:12 coeficientes MFCC + Δ: MFCC-D
- Rasgos SDC(12, 2, 2, 5) "en lugar de" los 12 coeficientes MFCC: SDC
- Rasgos SDC(12, 2, 2, 2) "añadidos a" los 12 coeficientes MFCC: MFC-SDC

Los vectores de rasgos cepstrales normalizados y los vectores SDC, fueron implementados según se explicó en el epígrafe 2.2.

5.1 Experimento 1 con expresiones cortas

Los 100 locutores de la base Ahumada se usaron como objetivos para sus modelos correspondientes y como impostores para el resto de los modelos, lo que representan en total 100 objetivos y 9900 impostores. Tanto la secuencia de dígitos como la frase balanceada tienen una duración aproximada de 5 segundos, tanto en entrenamiento como en la prueba.

Las curvas DET con el comportamiento del rasgo SDC ante la variabilidad de sesión, de canal y de sesión y canal, se muestran respectivamente en las figuras 2, 3 y 4.

Los resultados de EER y del mínimo DCF se muestran en las tablas 1, 2 y 3 respectivamente, para cada factor de variabilidad.

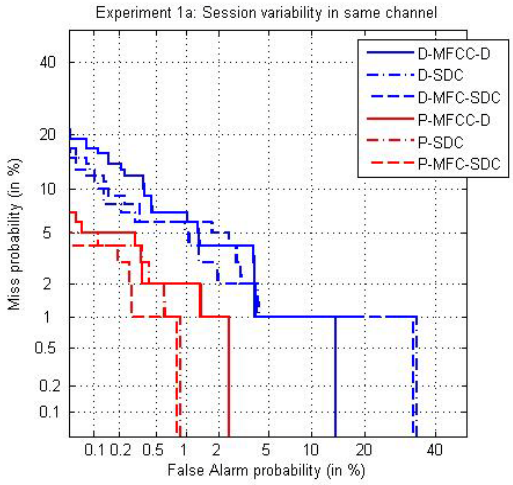


Fig. 2. Variabilidad de sesión en el mismo canal *D: dígitos. P: frases*

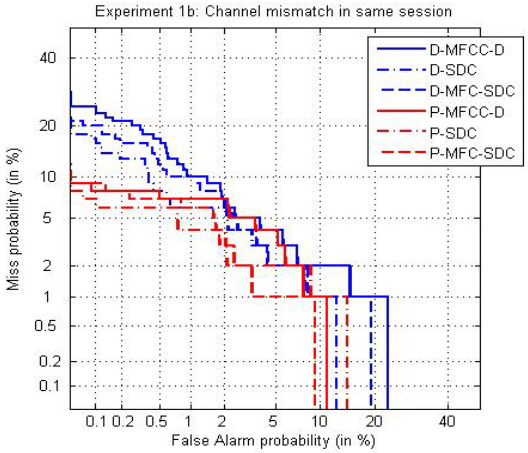


Fig. 3. Variabilidad de canal en la misma sesión *D: dígitos. P: frases*

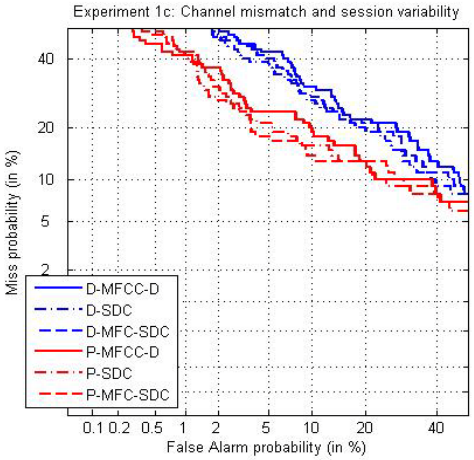


Fig. 4. Variabilidad de canal y de la sesión *D: dígitos. P: frases*

Tabla1. EER y DCF del experimento 1: variabilidad de sesión

Expresiones	Rasgos	EER %	DCF
D: dígitos	MFCC-D	3.83	0.0114
	SDC	2	0.009
	MFC-SDC	3	0.0093
P: frases	MFCC-D	1.36	0.0055
	SDC	0.87	0.0051
	MFC-SDC	0.80	0.0037

Tabla 2. EER y DCF del experimento 1: variabilidad de canal

Expresiones	Rasgos	EER %	DCF
D: dígitos	MFCC-D	4	0.019
	SDC	3.35	0.0123
	MFC-SDC	3.63	0.0159
P: frases	MFCC-D	4	0.0089
	SDC	2.09	0.007
	MFC-SDC	2.4	0.0093

Tabla 3. EER y DCF del experimento 1: variabilidad de sesión y canal

Expresiones	Rasgos	EER %	DCF
D: dígitos	MFCC-D	21.4	0.0666
	SDC	21	0.061
	MFC-SDC	20	0.0618
P: frases	MFCC-D	16	0.0479
	SDC	14	0.0445
	MFC-SDC	13	0.0492

Conclusiones del experimento 1:

- Se comprueba que la mejor efectividad de reconocimiento del locutor se logra ante la variabilidad de la sesión, empeora ante la variabilidad del canal, y es mucho más mala cuando están presentes ambas variabilidades, lo cual se ajusta a lo esperado.
- Se observa además para cualquier tipo de variabilidad, que el clasificador VQ brinda mejor efectividad de reconocimiento del locutor con frases balanceadas que con dígitos. Esto se debe a que se logra con las frases balanceadas una mejor representación fonética de cada locutor en los 64 clústeres del clasificador.
- El experimento demuestra que el rasgo SDC presenta una mejor efectividad de reconocimiento de locutores con expresiones cortas que el rasgo MFCC + Δ, en ambas variantes de utilización del rasgo SDC, "en lugar de" y "añadidos a" los rasgos MFCC.
- Debe notarse que, sin embargo no son apreciables las diferencias en la efectividad de reconocimiento entre ambas variantes del SDC, a pesar de que la variante "en lugar de", tiene una dimensión de 60 y la variante "añadidos a", tiene una menor dimensión de 36.

Estos resultados indican que el rasgo SDC puede ser una nueva alternativa al uso del rasgo MFCC+ Δ en aplicaciones robustas de reconocimiento del locutor, con expresiones cortas para el entrenamiento y la prueba, enfrentadas a la variabilidad de sesión y de canal. El EER refleja disminuciones que van desde 0.5 hasta 3% en algunos casos y el mínimo de la DCF también refleja en la mayoría de los casos una disminución.

5.2 Experimento 2 con expresiones de mayor duración

Como en el experimento anterior, los 100 locutores se usaron como objetivos para sus modelos correspondientes y como impostores para el resto de los modelos. Tanto el texto leído balanceado usado en el entrenamiento y en la primera prueba como la expresión espontánea usada en la segunda prueba, tienen una duración de alrededor de un minuto.

Las curvas DET con el comportamiento del rasgo SDC ante la variabilidad de sesión, de canal y de ambas en conjunto, se muestran en las figuras 5, 6 y 7, respectivamente. Los resultados de EER y mínimo DCF se muestran en las tablas 4, 5 y 6 respectivamente.

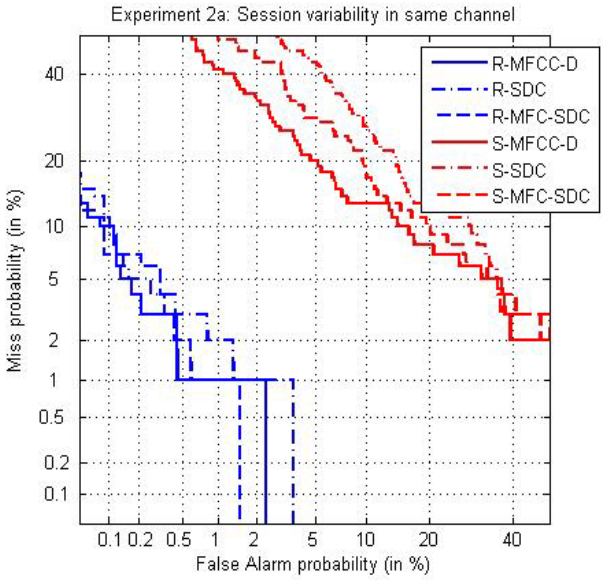


Fig. 5. Variabilidad de sesión en el mismo canal. *R: texto leído. S: entrenamiento: leído, prueba: espontáneo*

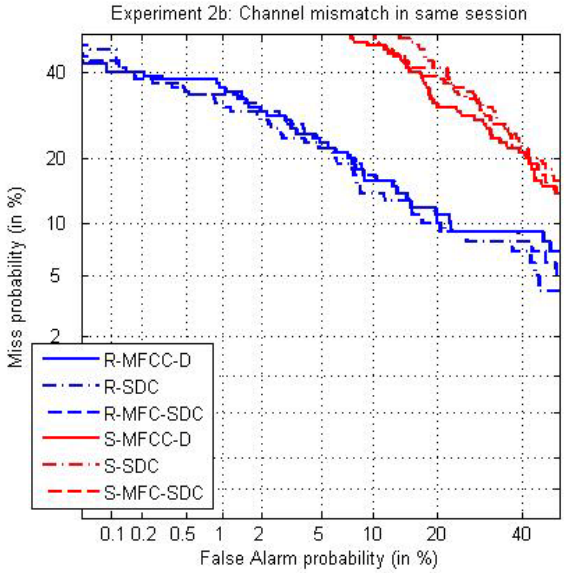


Fig. 6. Variabilidad de canal en la misma sesión. *R: texto leído. S: entrenamiento: leído, prueba: espontáneo*

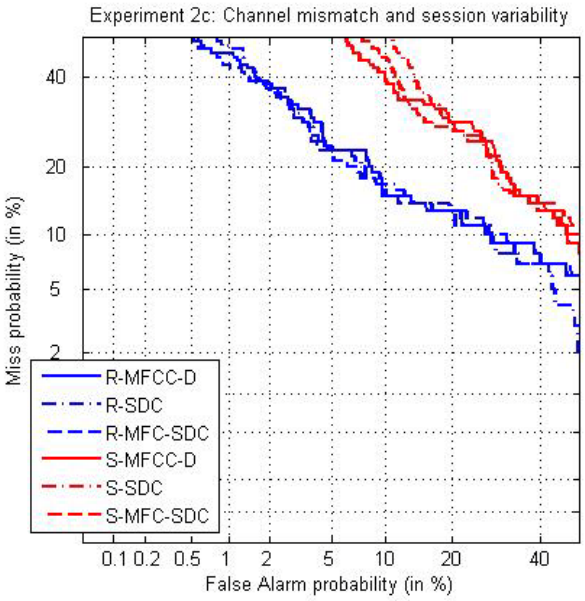


Fig. 7. Variabilidad de canal y de la sesión. *R: texto leído. S: entrenamiento: leído, prueba: espontáneo*

Tabla 4. EER y DCF del experimento 2a: variabilidad de sesión

Expresiones	Rasgos	EER %	DCF
R:texto leído	MFCC-D	1	0.0051
	SDC	1.3	0.0066
	MFC-SDC	1	0.0062
S:leído y espontáneo	MFCC-D	13	0.050
	SDC	15.67	0.097
	MFC-SDC	13	0.0574

Tabla 5. EER y DCF del experimento 2b: variabilidad de canal

Expresiones	Rasgos	EER %	DCF
R:texto leído	MFCC-D	14	0.0406
	SDC	13	0.0393
	MFC-SDC	14	0.0395
S:leído y espontáneo	MFCC-D	28.18	0.0809
	SDC	31	0.0905
	MFC-SDC	30	0.084

Tabla 6. EER y DCF del experimento 2c: variabilidad de sesión y canal

Expresiones	Rasgos	EER %	DCF
R:texto leído	MFCC-D	14	0.0533
	SDC	14	0.0557
	MFC-SDC	14	0.0531
S:leído y espontáneo	MFCC-D	25.85	0.081
	SDC	25	0.091
	MFC-SDC	26	0.0811

Conclusiones del experimento 2:

- Se comprueba que la efectividad del reconocimiento del locutor es mejor ante la variabilidad de la sesión, y peor ante la variabilidad del canal y ante ambas variabilidades evaluadas de forma conjunta, lo que indica que la variabilidad del canal tiene una afectación muy importante en la efectividad, cuando se entrena con expresiones de mayor duración.
- Se observa además que el clasificador GMM es más efectivo cuando se entrena y prueba con texto leído que cuando se entrena con texto leído y se prueba con la frase espontánea. Este elemento de variabilidad adicional incorporado, relacionado con la forma de hablar, provoca un mayor error en la clasificación en todos los casos, debido probablemente a la no correspondencia fonética entre ambas expresiones y a diferencias en la dinámica espectral entre ambas expresiones (leído-espontáneo).
- Se demuestra que el rasgo SDC tiene peor efectividad que el MFCC + Δ en reconocimiento del locutor con expresiones de mayor longitud ante la variabilidad de sesión, sin embargo tiene una efectividad similar ante la variabilidad del canal y de sesión y canal, solo cuando se entrena y prueba con texto leído.
- Como en el experimento anterior debe notarse que no son apreciables las diferencias en los resultados entre ambas variantes de SDC, a pesar de la diferencia en la dimensionalidad de ambas.

Los resultados obtenidos en ambos experimentos, con expresiones cortas y de mayor longitud, indican que el rasgo SDC puede ser una alternativa o un complemento al rasgo MFCC + Δ en aplicaciones robustas de reconocimiento remoto del locutor, aunque se requerirá mayor experimentación en condiciones controladas de variabilidad, para evaluar mejor su desempeño ante diversas condiciones de desigualdad en el canal.

Una parte de estos resultados fueron presentados y publicados en las memorias del 8avo Congreso Interspeech 2007, celebrado en Bélgica. [26]

6 La verificación biométrica del locutor

Como modalidad de verificación biométrica de la persona, el habla es una característica cuya obtención no es considerada amenazante o intrusiva. El teléfono es la modalidad principal de verificación biométrica de personas por el habla, dado que es una vía de comunicación oral ubicua y no intrusiva.

Como modalidad biométrica, el habla es una combinación de características fisiológicas y conductuales. Los rasgos del habla de un individuo se basan en características fisiológicas relativamente invariantes, como la forma y tamaño de su tracto vocal, las cavidades nasales y los labios, usados en la producción de la voz. Además, el habla es clasificada como un elemento biométrico conductual, porque la manera en que un individuo habla, dependen de su formación social y cultural.

Por otra parte, el habla de la persona cambia con la edad, las condiciones de salud, el estado emocional, las razones medioambientales, etc., por lo que no es muy distintiva y no es apropiada para la identificación biométrica en gran escala. Sin embargo, en una aplicación biométrica tele-bancaria o de comercio electrónico, las técnicas basadas en el habla, combinadas con otros métodos de autenticación de usuario, son muy utilizadas ya que pueden integrarse transparentemente durante el proceso de autentificación sobre el sistema telefónico.

La verificación biométrica del locutor sufre considerablemente cualquier variación que ocurra en el teléfono y en el canal de transmisión, deteriorándose apreciablemente su comportamiento cuando las condiciones de entrenamiento y de explotación no son iguales. El ruido del fondo también puede ser un problema importante, así como las variaciones en la voz debido a enfermedades, emociones o el envejecimiento. Muchos de estos problemas han recibido poca atención aún.

6.1 Categorías de verificación del locutor

Los sistemas de verificación de locutores se categorizan dependiendo del grado de libertad en el habla del usuario. La siguiente secuencia de métodos, está basada en tareas que aumentan su complejidad, correspondiéndose con la sofisticación de los algoritmos usados para reducir la vulnerabilidad ante ataques por repetición, y el progreso del estado del arte en el tiempo [27]:

- Texto fijo: El locutor dice una palabra predeterminada o frase que fue grabada durante el entrenamiento. La palabra actúa como una contraseña secreta, pero una vez grabada, un ataque por repetición es fácil, y es necesario un re-entrenamiento para cambiar la contraseña.
- Texto inducido: El locutor es inducido por el sistema para decir una expresión específica, por lo que no tiene que memorizar la contraseña. El sistema compara lo pronunciado con el texto entrenado para determinar al usuario. Para esto, el entrenamiento es normalmente más largo, pero el texto inducido puede cambiarse a voluntad, lo que dificulta un ataque por repetición de grabaciones del habla. Las expresiones como las secuencias de dígitos son más vulnerables que las frases a ataques por repetición basados en empalmes del texto.
- Texto independiente: El sistema procesa cualquier expresión del locutor. Aquí el habla puede orientarse a una tarea, por lo que se dificulta adquirir muestras de habla para además llevar a cabo la meta de un impostor. La supervisión puede ser continua, mientras más se diga, mayor es la confianza del sistema en la identidad del usuario. El advenimiento de la síntesis del habla entrenable, podría habilitar los ataques en este acercamiento. Tales sistemas pueden identificar a una persona incluso cuando cambia el idioma.
- Combinada con verificación de la expresión [28]: El sistema presenta al usuario una serie de frases aleatorias para repetir, y no sólo verifica la coincidencia de la voz sino también la coincidencia del contenido de las frases. Adicionalmente, es posible usar formas de verificación automática de conocimiento, donde una persona se verifica comparando el contenido de su habla contra la información almacenada en su perfil personal.

6.2 Enfrentamiento a las desigualdades del canal telefónico en la verificación del locutor

Los sistemas de verificación remota del locutor que utilizan el teléfono como vía de comunicación, enfrentan desafíos significativos causados por condiciones acústicas adversas como la limitación en el ancho de banda, el nivel de ruido en el canal y la variabilidad en cuanto a sensibilidad y ancho de banda del micrófono telefónico. La degradación en el comportamiento de los sistemas debido a dichas desigualdades constituye uno de los desafíos principales al despliegue actual de las tecnologías de reconocimiento de locutor.

Se han propuesto varias técnicas para enfrentar este problema:

- extraer rasgos del habla menos sensibles a los efectos del canal que el rasgo cepstral tradicional [29].
- reducir el efecto de las desigualdades removiendo la media cepstral [30,31].
- transformar los modelos del locutor para compensar las desigualdades [32,33].
- normalizar el resultado de la clasificación del locutor con relación a las desigualdades [34].

Los autores llevaron a cabo durante el año 2007 los siguientes experimentos para evaluar la robustez del rasgo SDC ante las desigualdades en el canal y en el teléfono, simulando un sistema biométrico de verificación remota del locutor con texto inducido.

7 Experimentos para evaluar la robustez ante las desigualdades en el canal y el teléfono

Teniendo en cuenta los resultados de los experimentos descritos en el epígrafe 5, se propusieron nuevos experimentos, utilizando el rasgo SDC con la combinación de parámetros (*N, D, P, k)* = (12, 2, 2, 2), lo que posibilitó:

a. Mantener una dimensión propuesta como estándar N=12 del vector de coeficientes cepstrales MFCC [35]
b. Obtener una buena estimación de la dinámica de las transiciones espectrales: la selección de $k = 2$ tramas, con $P = 2$ saltos entre tramas y $D = \pm 2$ tramas para el cómputo delΔ, brinda una duración total del rasgo SDC de 7 tramas correspondiente a 147 ms., para una duración de trama de 30 ms. con solapamiento de 30%. Esta duración del rasgo dinámico es la adecuada, según Furui [5] y Soong y otros [7], para estimar la dinámica espectral.
c. Reducir el costo computacional para obtener el rasgo SDC a partir de los rasgos Δ, asegurando el valor mínimo recomendado de $D = \pm 2$ para obtener una adecuada pendiente espectral [18].
d. Comparar el comportamiento del vector de rasgos SDC y el vector de coeficientes cepstrales + Δ, en una dimensionalidad similar.
e. Evitar “la maldición de la dimensionalidad" [36], teniendo en cuenta que los experimentos realizados anteriormente (epígrafe 5) con una dimensión mayor del vector de rasgos SDC (*kN=60*), brindaron resultados de efectividad muy similares a los obtenidos con la dimensión propuesta, *kN=24*.

Los nuevos experimentos evaluaron el comportamiento del rasgo SDC respecto al rasgo MFCC + Δ, al igual que en el experimento anterior (epígrafe 5), en dos acercamientos: "añadidos a" los coeficientes cepstrales y "en lugar de" los coeficientes cepstrales. Se evaluaron tres variantes de rasgos:

- Lineabase:12 coeficientes MFCC + Δ: MFCC +D
- Rasgos SDC (12, 2, 2, 2) “en lugar de” los 12 coeficientes MFCC: SDC
- Rasgos SDC (12, 2, 2, 2) “añadidos a” los 12 coeficientes MFCC: MFCC+SDC

Los vectores de rasgos cepstrales normalizados y los vectores SDC fueron implementados según se explicó en el epígrafe 2.2

7.1 Implementación de los experimentos

El sistema remoto de verificación biométrica del locutor con texto inducido se simuló con las diez frases fonológicamente y silábicamente balanceadas de las tres sesiones telefónicas de Ahumada utilizando una línea telefónica convencional. El sistema de verificación del locutor se evaluó con el clasificador GMM adaptado al Modelo Universal de Fondo ("UBM: Universal Background Model") con 256 mezclas, propuesto por Reynolds y otros [37].

Se dividió la base de datos Ahumada en dos subconjuntos de 50 locutores: el sistema se entrenó y probó con cada una de las diez frases balanceadas del primer subconjunto y se entrenó el UBM con las frases del segundo subconjunto. En la prueba, cada uno de los 50 locutores del primer subconjunto se usó como objetivo para sus correspondientes modelos y como impostor para el resto de modelos, obteniéndose 500 objetivos y 24500 impostores en cada experimento.

7.2 Sesiones de entrenamiento y prueba

El juego de muestras de entrenamiento se obtiene de la sesión T1, concatenando las diez frases balanceadas (aproximadamente 40 segundos de habla) de cada uno de los 50 locutores. En dicha sesión cada locutor llama desde el mismo teléfono, en una llamada interna.

Para evaluar el comportamiento ante las desigualdades del canal y del teléfono, las muestras de prueba se obtienen de las sesiones T2 y T3, con cada una de las frases de los mismos locutores (aproximadamente 5 segundos de habla cada una). Los locutores en la sesión T2 realizan una llamada desde el teléfono de su casa, en un ambiente silencioso, pero las características del canal y del teléfono son completamente desconocidas y no se posee ningún control sobre ellas. En la sesión T3, los locutores realizan una llamada local en un cuarto silencioso, usando 9 modelos de teléfonos seleccionados al azar, cada locutor usa uno de los 9 teléfonos. En este caso, para cada modelo de teléfono, tres características están reportadas [22]:

- sensibilidad del micrófono del teléfono
- respuesta de frecuencias del micrófono del teléfono
- rangos de relación señal a ruido en el canal asociado al teléfono.

La Tabla 7 resume para la sesión T3, la distribución de los 50 locutores por modelo de teléfono, así como las características de cada teléfono y de su canal asociado.

Tabla 7. Distribución de locutores por modelo de teléfono, características de cada teléfono y de su canal asociado.

	Modelo de teléfono	Número de identificación del locutor	Características del teléfono		Relación s/n (dB) del canal asociado		
			Sensibilidad (mV/P)	Nivel atenuación (dB)	Min	Max	media
1	Teide	14,18,19,24,27,31,34,37,38	3.2	15	36	39.5	37.8
2	Teide	15,21,23,26,30,36,44	3.2	15	32.4	41	36.7
3	Heraldo	6,9,43,50	0.9	45	20.6	27.9	24.2
4	Heraldo	7,8,10,22,42,49	3	45	31.6	33.9	32.5
5	Góndola	11,12,35,39,47	0.8	50	26.6	32	29.8
6	Complet	1,3,13,17,20,46,104	bajo	40	36	38.5	37.8
7	Teide	2,5,28,33	2.8	15	34	36.5	35.2
8	Heraldo	4,25,29,45	bajo	50	29	32.2	30.6
9	GE	16,32,40,48	bajo	35	33.1	38	35.6

Para evaluar la robustez del rasgo SDC ante las desigualdades del canal y el teléfono, los locutores en la sesión T3 se agruparon en dos clases, para cada una de las tres características medidas:

- Baja sensibilidad (<1 mV./P) y alta sensibilidad (> 2.5 mV./P) del micrófono del teléfono.
- Bajo nivel de atenuación (<20 dB) y alto nivel de atenuación (> 35 dB) de la respuesta de frecuencia del micrófono del teléfono.
- Baja y alta relación señal ruido media (umbral: 35 dB) en el canal asociado.

El teléfono utilizado en la sesión T1 es un modelo Teide, con alta sensibilidad y bajo nivel de atenuación en el micrófono y con alta relación señal ruido media en el canal asociado.

7.2.1 Experimento 1: Evaluación ante la desigualdad del canal en condiciones no controladas sesiones T1-T2

Se entrenó con las diez frases balanceadas de 50 locutores en la sesión T1 y se probó con cada una de las frases balanceadas de los mismos locutores, en la sesión T2. La curva DET con el comportamiento ante la desigualdad del canal en condiciones no controladas, se muestra en la figura 8.

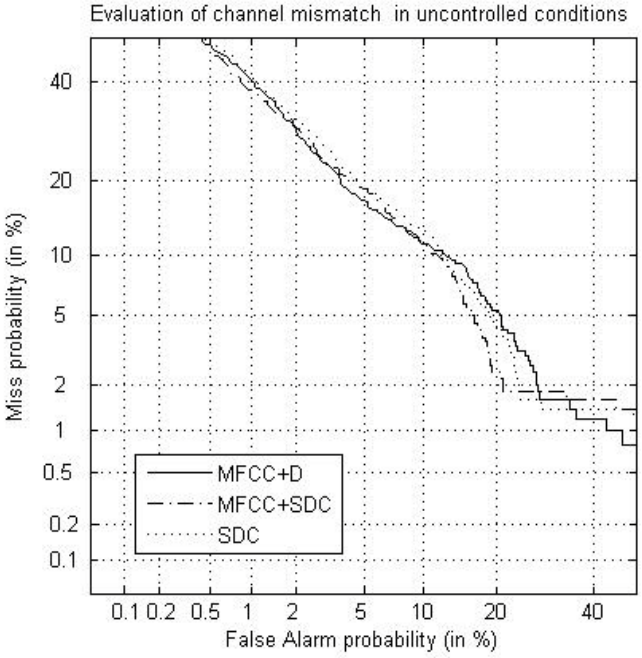


Fig. 8. Experimento 1: Condiciones no controladas, sesiones T1 y T2

La curva DET refleja una efectividad muy similar de ambas variantes del rasgo SDC y del rasgo MFCC+ Δ ante la desigualdad del canal. No obstante en la zona de Conveniencia Alta (por debajo del punto EER) ambas variantes de SDC superan en efectividad al rasgo MFCC+ Δ y en la zona de Seguridad Alta (por encima de EER) ambas variantes del rasgo SDC tienen similar efectividad que el rasgo MFCC+ Δ. Este primer resultado en condiciones del canal y del ente no controladas, constituyó una buena razón para continuar la evaluación del comportamiento de ambas variantes del rasgo SDC bajo condiciones de desigualdad controladas, en experimentos con la sesión T3.

7.2.2 Experimento 2: Evaluación ante la desigualdad del canal bajo condiciones controladas T1-T3

Se entrenó con las diez frases balanceadas de 50 locutores en la sesión T1 y se probó con cada una de las frases balanceadas de los mismos locutores en la sesión T3.

Experimento 2a) desigualdad debido a la sensibilidad del micrófono del teléfono.

Se agruparon los locutores de la sesión T3 en dos clases: baja sensibilidad (24 locutores) y alta sensibilidad (26 locutores). La curva DET con el comportamiento ante la sensibilidad del micrófono del teléfono, se muestra en la figura 9.

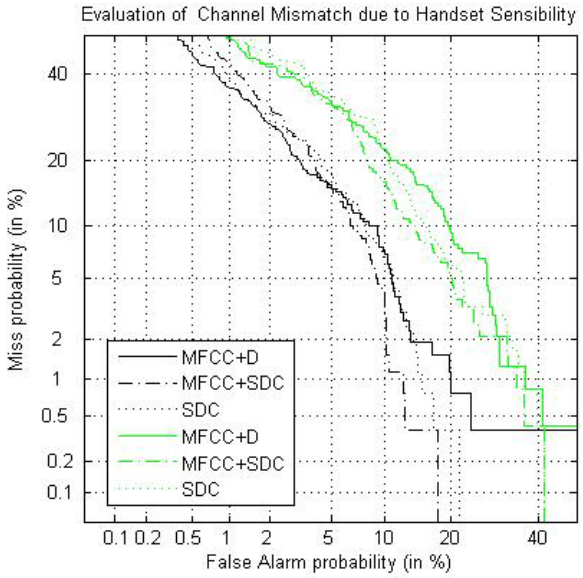


Fig. 9. Experimento 2a: *negro: alta sensibilidad, verde: baja sensibilidad*

Experimento 2b) desigualdad debido a la respuesta de frecuencia del teléfono.

Se agruparon los locutores de la sesión T3 en dos clases: Bajo nivel de atenuación (30 locutores) y Alto nivel de atenuación (20 locutores). La curva DET con el comportamiento ante la respuesta de frecuencia del teléfono, se muestra en la figura 10.

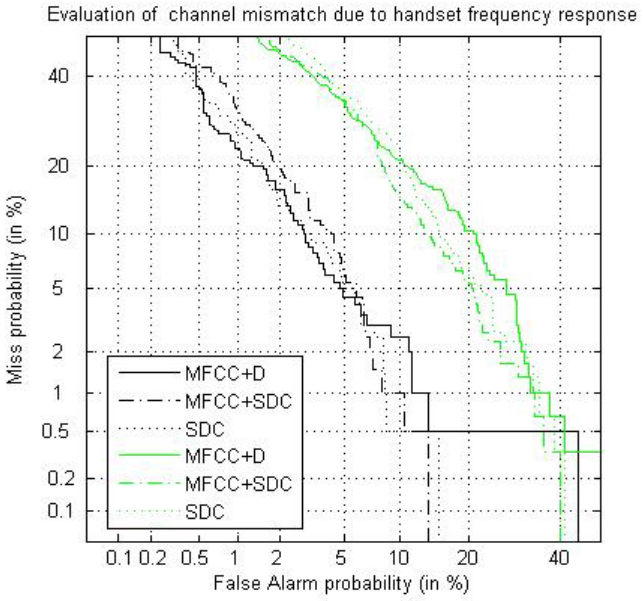


Fig. 10. Experimento 2b. *negro: baja atenuación, verde: alta atenuación*

Experimento 2c) desigualdad debido a la razón señal a ruido promedio del canal.

Se agruparon los locutores de la sesión T3 en dos clases: Baja (19 locutores) y Alta (31 locutores) relación señal /ruido promedio. La curva DET con el comportamiento ante la razón señal a ruido del canal, se muestra en la figura 11.

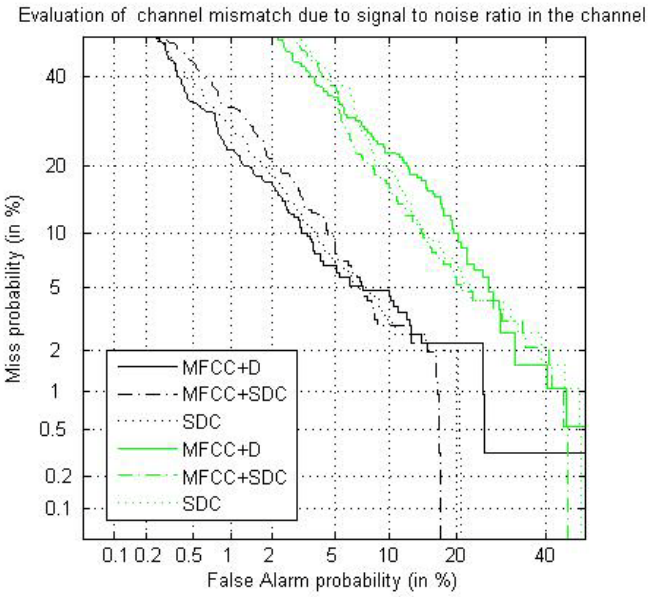


Fig. 11. Experimento 2c: *negro: alta relación s/n, verde: baja relación s/n*

Los resultados del experimento 2 reflejan que:

- cuando existe poca desigualdad entre el entrenamiento y la prueba (curvas en color negro), la efectividad de la verificación es mucho mejor que cuando existe desigualdad (curva verde), para cualquier tipo de desigualdad y variante de rasgos utilizada, lo que era de esperar.
- cuando existe poca desigualdad, el comportamiento de ambas variantes de rasgos SDC es muy similar a la observada bajo condiciones no controladas del canal y el teléfono (experimento 1).
- sin embargo se observa una mayor efectividad de reconocimiento de ambas variantes de rasgos SDC "añadidos a" y "en lugar de", respecto al coeficiente MFCC+ las peores condiciones de desigualdad: sensibilidad baja del teléfono, alta atenuación en la respuesta de frecuencia del teléfono y baja relación señal /ruido promedio en el canal.

Los resultados del comportamiento del EER para las tres variantes de rasgos propuestas en los experimentos 1 y 2 (a, b y c), se muestran en la tabla 8.

Tabla 8. Valores de EER en %

Tipo de rasgos	Exp1	Experimento 2a		Experimento 2b		Experimento 2c	
		Baja sensibilidad	Alta sensibilidad	Alta atenuación	Baja atenuación	Baja s/n	Alta s/n
MFCC + D	10.7	15.4	9.1	15.4	4.9	15.7	5.8
MFCC + SDC	10.2	11.9	8.0	12	5.5	12.1	6.3
SDC	11.1	13.2	8.4	13.3	5	12.7	6.1

La tabla 8 muestra que los rasgos SDC “añadidos a” tienen un mejor EER que los rasgos MFCC +Δ bajo condiciones de desigualdad no controladas en el experimento 1, y bajo las peores condiciones de desigualdad en el experimento 2. Debe observarse que los rasgos SDC “en lugar de” tienen un mejor EER que los rasgos MFCC +Δ, bajo las peores condiciones de desigualdad en el experimento 2.

La tabla 9 refleja la reducción relativa de EER en %, para ambas variantes de rasgos SDC respecto a los rasgos MFCC +Δ, en la evaluación bajo las peores condiciones de desigualdad (experimento 2).

Tabla 9. Reducción relativa del EER en %

Condición de desigualdad	MFCC + SDC	SDC
Baja sensibilidad del teléfono	22	14
Alta atenuación del teléfono	22	13
Baja relación señal/ruido en el canal	23	19

Se confirma de los resultados del experimento 2, que ambas variantes de rasgos SDC muestran una mejora en la efectividad de reconocimiento respecto a la variante con coeficientes MFCC +Δ. La variante de los rasgos SDC "añadidos a" reflejan una mejora entre el 22 y el 23 %, mientras que la variante de los rasgos SDC "en lugar de" muestran una mejora en la efectividad entre el 13 y el 19%. Este resultado confirma la robustez del rasgo SDC ante las desigualdades del canal.

Los resultados obtenidos confirman lo predicho al finalizar el epígrafe 5, el rasgo SDC puede considerarse como nueva alternativa o complemento robusto de los coeficientes cepstrales MFCC, con la finalidad de reducir los efectos de las desigualdades del canal y el teléfono en aplicaciones de reconocimiento remoto del locutor. Los rasgos SDC "añadidos a" los rasgos MFCC muestran la mayor robustez, aunque los rasgos SDC "en lugar de" los rasgos MFCC +Δ son robustos también, con la misma dimensionalidad del vector. Debe observarse que no se incrementa el costo computacional, ya que el rasgo SDC es una concatenación de rasgos Δ.

El trabajo a continuación se dirigió a determinar la posible causa de la robustez del rasgo SDC en su carácter seudo-prosódico, evaluando su relación con la dinámica temporal de los rasgos prosódicos del habla.

Estos resultados fueron presentados y publicados en las memorias del 12mo CIARP 2007, celebrado en Chile. [38]

8 Los rasgos SDC y su carácter seudo prosódico

Muchos estudios [39, 40, 41, 42, 43] han demostrado que la estructura prosódica del habla conlleva información de gran valor sobre la identidad del locutor. La prosodia del habla se caracteriza por la dinámica y el comportamiento estadístico del tono fundamental y la energía de la voz (entonación), así como por el ritmo del habla (razón habla-pausa). Dichos rasgos prosódicos, combinados adecuadamente con los rasgos cepstrales y dinámicos han elevado apreciablemente la efectividad de los métodos de reconocimiento del locutor al contener información de la estructura de las cuerdas vocales que diferencia a las personas, así como han elevado su robustez ante las desigualdades y el ruido [40,42].

La información de la prosodia puede obtenerse de varias maneras a partir de los rasgos acústicos. Se puede calcular la estadística global de los rasgos, como la media y la desviación estándar del tono fundamental y la energía, pero esta aproximación no captura la información dinámica temporal de la secuencia de los rasgos prosódicos. Otra forma de extracción consiste en comparar la trayectoria temporal del contorno del tono fundamental y la energía [39], lo cual no es muy eficiente. La obtención de la función de la derivada del tono fundamental y la energía para describir su dinámica es el método que ha demostrado ser más efectivo [40] para el reconocimiento del locutor. El problema práctico de la utilización de los rasgos prosódicos es el gran volumen de datos necesarios para un reconocimiento adecuado y su alto costo computacional [42,43]

Los rasgos cepstrales dinámicos del habla -Δ y ΔΔ- obtenidos sobre intervalos de tiempo del habla relativamente extensos (intentando abarcar un probable comportamiento seudo-prosódico), han sido usados en verificación del locutor [1, 5, 7, 11,12] como se explicó en el epígrafe 2. Furui [5] recomienda un intervalo de tiempo de 90 ms. para su cálculo, lo que permite conservar la información asociada a la transición espectral entre fonemas. Soong y Rosemberg [7] recomiendan un intervalo de tiempo algo mayor, entre 100 y 160 ms., lo que permite obtener una buena estimación de la tendencia de las transiciones espectrales entre las sílabas.

Como ejemplo, el cálculo de los rasgos Δ y ΔΔ utilizando la ec. (2), con tramas de 20 ms. solapadas en un 50 %, requiere fijar *D=5* para cubrir un intervalo dinámico de 110 ms.; si las tramas son de 30 ms. y se solapan un 30 % (como en nuestros experimentos), se requiere fijar *D=3*, para lograr cubrir un intervalo de 127 ms. Ambas propuestas provocan un incremento del costo computacional, debido al intervalo temporal que se requiere cubrir, con el consiguiente incremento de *D*.

Recientemente (2006) J. Lareau [13] en su tesis expresa pero no defiende, la siguiente idea: "The use of the shifted delta cepstral feature vectors allows for a pseudo-prosodic feature vector to be computed without having to explicitly find or model the prosodic structure of the speech signal."

Esta idea de que el rasgo SDC puede tener un carácter seudo-prosódico, puede tener su base en el hecho de que el rasgo SDC, como un rasgo dinámico espectral de larga duración, pero que se obtiene concatenando varios rasgosΔ obtenidos con *D* =1 o 2, debe reflejar la dinámica espectral del habla mucho mejor que el rasgoΔ de larga duración [5, 7]. En cada trama cepstral, el rasgo SDC contendrá información sobre la dinámica espectral de los formantes del habla (que son un reflejo de la dinámica de los articuladores vocales y nasales) si se asegura que el intervalo de tiempo evaluado por el rasgo SDC a lo largo de las siguientes tramas cepstrales, incluya las transiciones espectrales entre fonemas y sílabas, como se recomienda en [5] y [7]. Esta información dinámica espectral de larga duración, pudiera ser considerada como un reflejo indirecto (seudo) del comportamiento prosódico del habla, obtenida ahora a un bajo costo computacional, sin necesidad de calcular los rasgos prosódicos del habla ni de utilizar altos valores de *D* en el cálculo de los rasgos Δ.

Los parámetros *N, D, P y k* del rasgo SDC, influyen de diferente manera en la dimensionalidad del vector, el costo computacional y el probable comportamiento seudo-prosódico del rasgo SDC:

- *N:* numero de coeficientes en cada vector cepstral, en conjunto con k, determina la dimensionalidad del vector SDC
- *D:* duración de la ventana para el cálculo del Δ en la ec. (2), determina el costo computacional
- *P:* desplazamiento entre tramas a concatenar, en conjunto con D, regula el solapamiento entre tramas y la redundancia en el vector SDC, no afecta el costo computacional o la dimensionalidad del vector.
- *k:* numero de tramas cuyos vectores de rasgos Δ se concatenan para formar el vector SDC, determinan el carácter seudo-prosódico del rasgo SDC y en conjunto con N, su dimensionalidad.

9 Experimentos para evaluar el carácter seudo-prosódico de los rasgos SDC

Teniendo en cuenta la simplicidad computacional del cálculo del rasgo SDC, los autores evaluaron su probable carácter seudo-prosódico, midiendo la relación lineal entre el rasgo SDC y la dinámica de dos rasgos prosódicos [40], el tono fundamental y la energía, y evaluando la robustez de los rasgos SDC más correlacionados en un experimento de verificación de locutores. Dicha evaluación fue realizada con las mismas frases de la base NIST 2001 Ahumada [22] y bajo las condiciones de desigualdad del canal y el teléfono explicadas en el epígrafe 7, se utilizó como clasificador el modelo GMM adaptado al UBM [37] de 256 mezclas, con similar distribución de objetivos e impostores.

El vector de rasgos SDC utiliza los parámetros (*N, d, P, k)* = (12, 2, 2, 2) cubriendo un intervalo de tiempo de 7 tramas de 21 ms., equivalente a 147 ms. que se corresponde con el intervalo sugerido por Soong y Rosemberg [7], lo que posibilita estimar la dinámica espectral entre las silabas.

9.1 Correlación temporal del rasgo SDC con la dinámica del tono fundamental y la energía

Para evaluar la relación lineal que pudiera existir entre el rasgo SDC y los rasgos prosódicos se utilizó la correlación temporal entre la secuencia temporal del rasgo SDC para cada coeficiente cepstral y la

secuencia temporal de la dinámica del tono fundamental y la energía. La cros-correlación entre dos secuencias de longitud N, "*x*" y "*y*", posibilita una comparación estadística de ambas secuencias como función del desplazamiento temporal *m*, e indica la fortaleza y dirección de la relación lineal entre ambas.

$$\Phi_{xy}[m] = \frac{1}{2N+1}\sum_{t=1}^{N-m} x[t]y[t+m] \tag{6}$$

Si "*x*" e "*y*" están estandarizadas, los límites de la cros-correlación son $-1 \leq {}^{\Phi_{xy}[m]} \leq 1$. Las fronteras ±1 indican una máxima correlación, y 0 indica no correlación. Una correlación alta pero negativa, indica una relación lineal inversa.

Para calcular la cros-correlación existente entre ambas secuencias temporales, se seleccionaron dos expresiones (una leída y otra espontánea) de 30 locutores de la base Ahumada, representando en total alrededor de 90 minutos de habla telefónica. Se obtuvieron en cada trama de habla los 12 coeficientes MFCC y sus correspondientes rasgos SDC, como se explicó en el epígrafe 2.2. La normalización de la media y la varianza de los rasgos se aplicaron como se explica en el Anexo 1 para los rasgos cepstrales. Los valores de la derivada del tono fundamental y la energía, Δpitch y Δenergía, fueron calculados utilizando la ec.(2) con D=2 en cada trama de habla.

Los resultados de la cros-correlación entre cada uno de los 12 rasgos SDC con la Δenergía (rojo) y el Δpitch (azul), se muestran en la figura 12 y en la tabla 10.

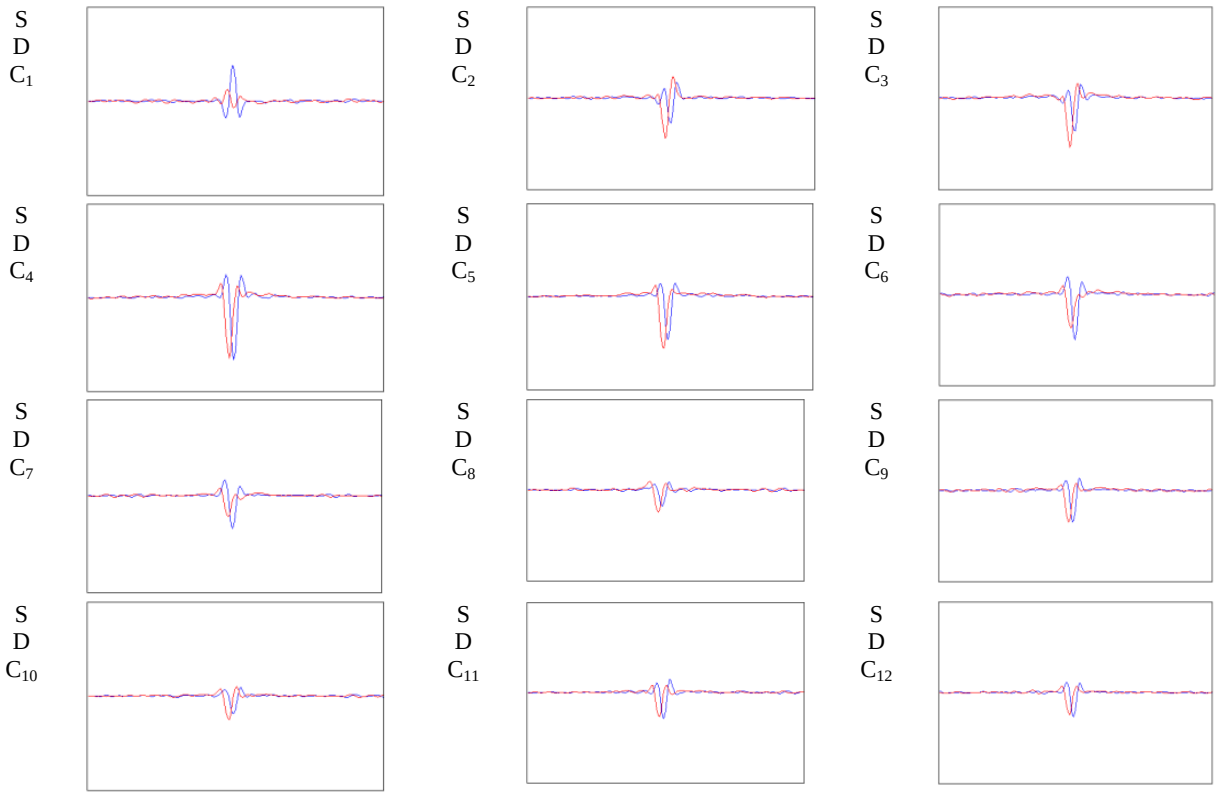


Fig. 12. Cros-correlación entre cada rasgo SDC y el Δpitch (azul), y cada rasgo SDC y Δenergía (rojo) (escala vertical ± 1 , escala horizontal ± 200)

Tabla 10. Cros-correlación de Rasgos SDC con el Δpitch y Δenergía (en rojo los seis mayores)

	Cros- correlación con Δenergía	Cros- correlación con Δpitch
SDC1	+0.12	+0.35
SDC2	-0.45	-0.30
SDC3	-0.55	-0.37
SDC4	-0.65	-0.67
SDC5	-0.56	-0.48
SDC6	-0.37	-0.50
SDC7	-0.22	-0.35
SDC8	-0.25	-0.18
SDC9	-0.35	-0.35
SDC10	-0.25	-0.20
SDC11	-0.27	-0.30
SDC12	-0.25	-0.27

La mayor cros-correlación de los rasgos SDC se obtuvo con respecto a laΔenergía. En general, los picos de correlación son negativos, reflejando una relación lineal inversa. Aunque los valores de los picos de correlación no son muy impresionantes, lo que indica que la relación lineal entre ambas secuencias temporales no es muy fuerte, hay algunos rasgos SDC más correlacionados con la dinámica de los rasgos prosódicos que otros. Estos fueron seleccionados para concatenarlos al vector MFCC y evaluar su robustez en un experimento de verificación de locutores similar al explicado en el epígrafe 7. La Tabla 11 refleja los rasgos SDC organizados en orden decreciente de correlación con ambos rasgos prosódicos.

Tabla 11. Rasgos SDC organizados en orden decreciente de cros-correlación

Orden	1	2	3	4	5	6	7	8	9	10	11	12
Δenergía	SDC_4	SDC_5	SDC_3	SDC_2	SDC_6	SDC_9	SDC_{11}	SDC_8	SDC_{10}	SDC_{12}	SDC_7	SDC_1
Δpitch	SDC_4	SDC_6	SDC_5	SDC_3	SDC_9	SDC_7	SDC_1	SDC_2	SDC_{11}	SDC_{12}	SDC_{10}	SDC_8

Entonces, dos vectores de seis rasgos SDC fueron creados y concatenados al vector MFCC: el primer vector, más correlacionado con la Δenergía, compuesto por SDC_2 , SDC_3, SDC_4, SDC_5, SDC_6 y SDC_9 , y un segundo vector, más correlacionado con el Δpitch, compuesto por SDC_3, SDC_4, SDC_5, SDC_6, SDC_7 y SDC_9. Ambos vectores tienen la misma dimensionalidad que el vector Δ.

El experimento evaluó la efectividad de reconocimiento de un sistema de verificación de locutor, como el explicado en el epígrafe 7, pero con dos nuevas variantes de rasgos: los dos vectores de rasgos SDC seleccionados como los más correlacionados a laΔenergía y al Δpitch, concatenados al vector de rasgos MFCC, manteniendo la misma dimensionalidad que el vector de rasgos que se utiliza como línea base:

- Lineabase:12 coeficientes MFCC + Δ: M-D
- 12 MFCC + 6 SDC más correlacionados con Δenergía: M-SDC-E
- 12 MFCC + 6 SDC más correlacionados con Δpitch: M-SDC-P

9.2 Resultados experimentales de verificación de locutor con rasgos SDC mejor correlacionados con la dinámica de los rasgos prosódicos.

Los resultados del experimento de verificación de locutor para las peores condiciones de desigualdad en el canal y el teléfono (semejante al experimento del epígrafe 7.2.2), se reflejan en las curvas DET de las figuras 13, 14 y 15:

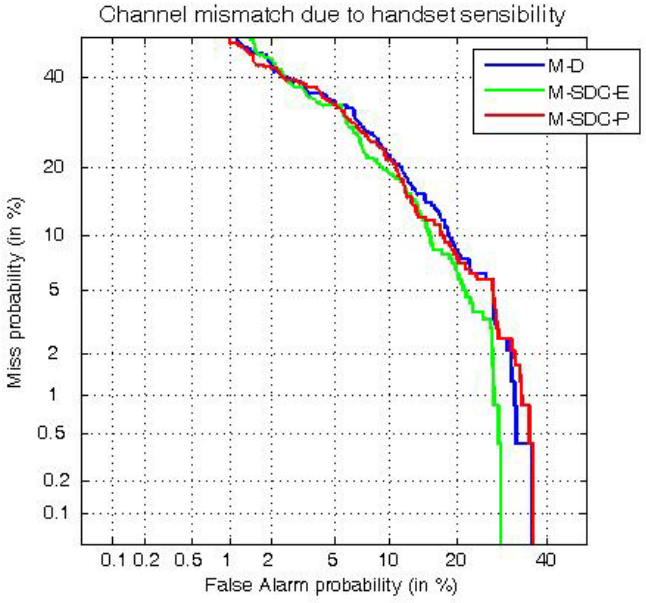


Fig. 13. Experimento con desigualdad por *baja sensibilidad del micrófono del teléfono* (< 1 mV/P)

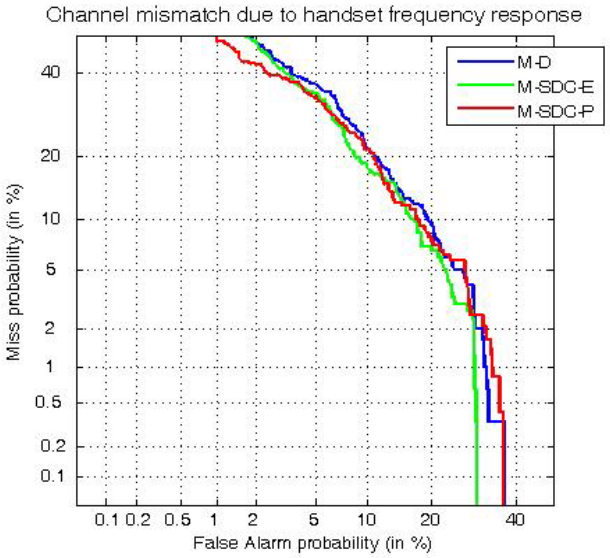


Fig. 14. Experimento con desigualdad por baja amplitud en respuesta de frecuencia del micrófono (< 20 dB)

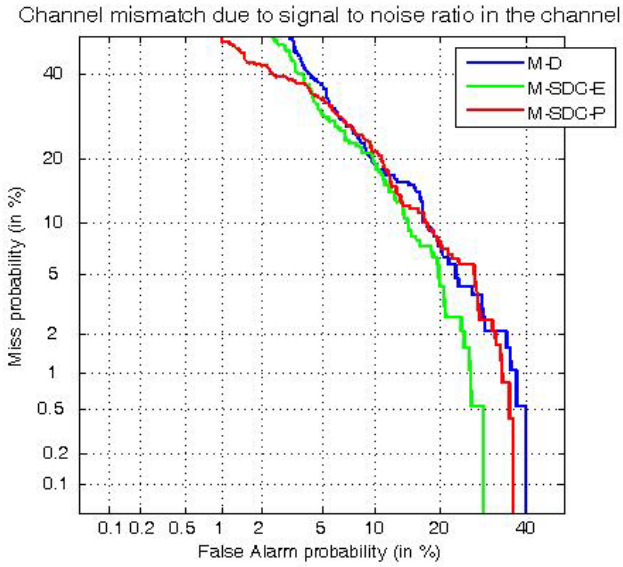


Fig. 15. Experimento con desigualdad por *baja relación señal/ruido en el canal*(< 30 dB)

Los valores de EER (%) y DCF de los experimentos se muestran en la Tabla 12.

Tabla 12. EER (%) y DCF de la línea base y de las dos variantes de rasgos MFCC+ SDC.

Variantes de rasgos	MFCC +Δ (líneabase)		MFCC+ SDC correlacionado con Δenergía		MFCC+ SDC correlacionado con Δpitch	
Condición de desigualdad	EER	DCF	EER	DCF	EER	DCF
Baja sensibilidad del teléfono	14.7	0.06	13.7	0.061	13.2	0.068
Alta atenuación del teléfono	14.2	0.065	13.9	0.066	13.4	0.062
Baja relación señal/ruido en el canal	15.3	0.068	12.6	0.069	13.4	0.058

La Tabla 13 refleja la reducción relativa en % de EER, para ambas variantes de rasgos MFCC+ SDC respecto a MFCC +Δ.

Tabla 13 Reducción en % del EER para ambas variantes de rasgos MFCC+ SDC respecto a la línea base

Condición de desigualdad	MFCC+ SDC correlacionado con Δenergía	MFCC+SDC correlacionado con Δpitch
Baja sensibilidad del teléfono	6.8	8.8
Alta atenuación del teléfono	2.1	5.6
Baja relación señal/ruido en el canal	17.6	13.7

Los resultados experimentales de las curvas DET y las tablas muestran:

- Una efectividad de reconocimiento superior de ambas variantes de rasgos MFCC+ SDC respecto al rasgo MFCC + Δ (Tabla 13).
- Una mejor efectividad de reconocimiento respecto al rasgo MFCC + Δ, del rasgo SDC más correlacionado con la Δenergía (curva verde en las figuras 13,14 y 15). Este resultado es consistente con la mayor correlación del rasgo SDC con Δenergía.
- Una robustez superior de ambas variantes de rasgos SDC, principalmente ante la baja relación señal-ruido en el canal, consistente con la robustez ante el ruido, propia de los rasgos prosódicos (Tabla 13).

Los resultados experimentales confirman que el rasgo SDC puede considerarse como un nuevo rasgo dinámico robusto con cierto carácter seudo-prosódico, como alternativa a los rasgos MFCC +Δ, lo que posibilita reducir los efectos de desigualdad del canal y el teléfono en la verificación del locutor, teniendo en cuenta que:

- El intervalo de tiempo que puede abarcar el rasgo SDC permite evaluar la dinámica espectral del habla, incluso las transiciones entre fonemas y entre sílabas.
- El costo de cómputo del rasgo SDC es casi el mismo que el de Δ, sólo requiere la concatenación adecuada de los rasgos Δ para conformar el rasgo SDC.

Estos resultados fueron presentados y publicados en las memorias del 13 CIARP 2008, celebrado en La Habana [42].

10 Selección de la combinación más efectiva de parámetros del rasgo SDC, utilizando la Información Mutua con la identidad del locutor

Los resultados obtenidos hasta aquí, han mostrado que el rasgo SDC presenta una mejor robustez que el rasgo MFCC, ante las desigualdades del canal y el micrófono del teléfono y cierta relación lineal con la dinámica del pitch y la energía, reflejando un comportamiento seudo-prosódico.

Sin embargo no se ha obtenido una medida objetiva de las mejores combinaciones de parámetros del rasgo SDC para el reconocimiento del locutor, por lo que se propuso un método para estimar la Información Mutua entre el rasgo SDC, con diferentes combinaciones de parámetros, y la identidad del locutor.

10.1 Información Mutua

Si X es una variable aleatoria discreta, la entropía incondicional de X [43,44] es:

$$H(X) = -\sum_k p(X = k)\log p(X = k) = -\sum_k p(x)\log p(x) \quad (7)$$

donde *p(x)* es la función de densidad de probabilidad marginal de *x*.

Como $p(X = k) \le 1 \forall k \Rightarrow \log p(X = k) \le 0 \Rightarrow H(X) \ge 0$, la entropía incondicional de X es siempre positiva.

Si se observa una segunda variable aleatoria discreta *Y*, su valor en general alterará la distribución de los posibles valores de *X*, $p(X = k / Y = y)$ y la entropía condicional de *X* dado el valor de *Y* es:

$$H(X / Y = y) = -\sum_k p(X = k / Y = y)\log p(X = k / Y = y) = -\sum_k p(x / y)\log p(x / y) \quad (8)$$

donde *p(x/y)* es la función de densidad de probabilidad condicional de *x* dado *y*.

La entropía condicional media es un promedio de las entropías condicionales respecto a la función de densidad de probabilidad condicional de *X* dado *Y*:

$$H(X / Y) = \sum_y p(Y = y)H(X / Y = y) = -\sum_y p(y)\sum_k p(x / y)\log p(x / y) \quad (9)$$

Las ec. 7 y 9 reflejan la entropía de Shannon de *X* antes y después de observar *Y*.

Conocer el valor de *Y* puede, como promedio, solo reducir la incertidumbre acerca de *X*, por lo que la entropía condicional *H(X/Y)* es siempre menor o igual que la entropía incondicional *H(X)*. La diferencia entre ambas es una medida de cuánto conocer *Y* reduce la incertidumbre acerca de *X*, y se conoce como Información Mutua entre *Y* e *X*, denotándose como *MI*:

$$MI(X;Y) = MI(Y;X) = H(X) - H(X / Y) = H(Y) - H(Y / X) \quad (10)$$

La *MI* mide la información que comparten *X* e *Y*, establece cuánto del conocimiento de una de las variables reduce la incertidumbre acerca de la otra. La *MI* es simétrica, debido al resultado intuitivo de que la información que *Y* brinda acerca de *X* es la misma cantidad de información que, conociendo *X*, nos brinda de *Y*.

Si *X* e *Y* fuesen completamente independientes, entonces el conocimiento de *X* no brindaría información acerca de *Y*, y viceversa, entonces *MI(X;Y) = 0: H(X) = H(Y) = H(X/Y)=H(Y/X)*

Si *X* e *Y* fuesen completamente dependientes, conociendo *X* se determina *Y*, y viceversa, entonces *MI* es igual a la incertidumbre de *X* o de *Y*, y *H(X/Y) =H(Y/X)= 0: MI(X;Y)= H(X)=H(Y)*

Entonces, los límites de la Información Mutua son: $0 \le MI(X;Y) \le H(X)$

10.2 Información Mutua entre la identidad de un locutor y los rasgos del habla

En reconocimiento del locutor se asume un modelo estocástico para el habla generada por un locutor *S*, seleccionado aleatoriamente con probabilidad uniforme de un conjunto de *N* locutores, entonces *p(S) = 1/N*.

Una expresión de habla generada por el locutor *S* -de acuerdo con su función de densidad de probabilidad condicional- se usa para determinar su identidad: para cada trama de habla de la expresión, se obtiene un vector independiente de rasgos x_i, utilizando un método de extracción de rasgos; la secuencia de rasgos $X = \{x_1, x_2, x_3, ...x_T\}$ alimenta un clasificador para obtener un modelo λ_X que clasifica al locutor *S*.

La Información Mutua entre un locutor *S* –representado por su modelo λ_X - y su secuencia de vectores de rasgos *X*, obtenida de su habla, brinda una medida de cuánto el conocimiento de *X* reduce la

incertidumbre acerca de la identidad del locutor S:

$$MI(S;X) = H(S) - H(S/X) \tag{11}$$

donde la entropía incondicional del conjunto de locutores es:

$$H(S) = -\sum_{S} p(S)\log p(S) \tag{12}$$

si $p(S) = 1/N \Rightarrow H(S) = -\sum_{S=1}^{N} \frac{1}{N}\log\frac{1}{N} = \log N$

La entropía condicional del conjunto de locutores dado una secuencia de vectores de rasgos X_i es:

$$H(S/X = X_i) = -\sum_{S=1}^{N} p(S/X_i)\log p(S/X_i) \tag{13}$$

y la entropía condicional media del conjunto de locutores dados su conjunto de secuencias de vectores de rasgos es:

$$H(S/X) = \sum_{i=1}^{N} p(X = X_i)H(S/X = X_i) = -\sum_{i=1}^{N} p(X)\sum_{S=1}^{N} p(S/X_i)\log p(S/X_i) \tag{14}$$

La ec. 12 es la entropía incondicional del conjunto de locutores, solo depende de las clases *S* y no depende del conjunto de vectores de rasgos *X*. En este caso, como el conjunto de locutores posee una distribución uniforme, cada locutor tiene la misma probabilidad p*(S) =1/N. H(S)* es el límite superior de la Información Mutua *MI (S; X).*

La ec. 14 es la entropía condicional del conjunto de locutores dado el conjunto de vectores de rasgos *X*, y puede ser interpretado como una disminución de la incertidumbre de la identidad de los locutores. Esto es, si existe una gran interdependencia entre el vector de rasgos *X* y el locutor *S*, el conocimiento de *X* reduce la incertidumbre acerca de la identidad del locutor *S*, entonces $H(S/X) \Rightarrow 0$, y habrá una gran certeza en la clasificación del locutor dado sus conjunto de vectores de rasgos, $MI(S;X) \Rightarrow H(S)$.

La Información Mutua *MI(S; X)* es máxima e igual a *H(S)* cuando el conjunto de vectores de rasgos *X* y el conjunto de locutores *S* son totalmente dependientes. Entonces, una medida de la *MI(S; X)* para diferentes vectores de rasgos, reflejará cuál de ellos contendrá más información acerca de la identidad del locutor.

10.3 La función de densidad de probabilidad "*pdf*" condicional de *S* dado *X*

El problema es obtener una buena representación de la función de densidad de probabilidad "*pdf: probability density function"* condicional del locutor *S* dado el correspondiente vector de rasgos *X*. Se propone para eso el uso del modelo de mezclas gausianas "*GMM: Gaussian Mixture Model*" [25].

El modelo GMM está bien establecido para la clasificación de locutores. Durante el entrenamiento del clasificador, la *pdf* condicional de un vector de rasgos $\vec{x}$ extraído de una expresión de habla dicha de un locutor conocido *S* puede modelarse por una mezcla de densidades gaussianas definida por:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) . \tag{15}$$

donde $\vec{x}$ son los vectores de rasgos, p_i son los pesos de las mezclas, *M* es el número de mezclas y $b_i(\vec{x})$ son las densidades gaussianas, que se parametrizan por sus pesos, el vector de medias y la matriz de covarianza $\lambda = \{p_i, \mu_i, \sum_i\}$, *i=1 ,..., M*.

La meta de la clasificación es estimar el parámetro λ del modelo GMM, que mejor se corresponda con la *pdf* del vector de rasgos, usando el método de estimación de máxima similaridad, [45] "MLE: maximum likelihood estimation" cuyo objetivo es encontrar el parámetro λ del modelo que maximice la probabilidad de generar el vector de rasgos. Para una secuencia de *T* vectores de rasgos $X = (\vec{x}_1, \vec{x}_2, \vec{x}_3, ..., \vec{x}_T)$, la probabilidad de la secuencia con respecto a λ se define como la función de probabilidad $p(X|\lambda)$; si $\vec{x}_1 ..... \vec{x}_T$ variables aleatorias independientes, $p(X|\lambda)$ puede escribirse como:

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda) \tag{16}$$

El parámetro λ del modelo que maximice la probabilidad se estima usando el algoritmo iterativo de maximización de la expectación ("EM: Expectation-Maximization") [45]. La idea del algoritmo EM es, comenzando con un modelo inicial λ, estimar un nuevo modelo λ^{new} tal que $p(X \mid \lambda^{new}) \geq p(X \mid \lambda)$. El nuevo modelo se convierte en el modelo inicial para la próxima iteración y el proceso se repite hasta que se alcance un umbral de convergencia. Cuando el método EM finaliza, el parámetro λ del modelo GMM es la mejor representación del vector de rasgos *X* obtenido de la expresión de habla dicha por un locutor conocido *S*. Entonces el parámetro λ del modelo GMM "clasifica" al locutor *S* en una clase.

Durante la prueba de reconocimiento del locutor, la intención es determinar cuál de los N modelos del conjunto de locutores mejor se corresponde con una expresión de habla de origen desconocido. Para un conjunto de *N* locutores conocidos representados por sus modelos GMMs $\lambda_1, \lambda_2, ... \lambda_N$, el primer paso es obtener la *probabilidad a posteriori* de ocurrencia de cada locutor conocido representado por su modelo λ_k, con respecto a la secuencia de vectores de rasgos *X* obtenida de la expresión de habla de origen desconocido (C*lasificador de Bayes*):

$$p(\lambda_k \mid X) = \frac{p(X \mid \lambda_k)}{p(X)} p(\lambda_k) \tag{17}$$

El objetivo final es hallar el modelo del locutor $\hat{s}$ que tenga la máxima *probabilidad a posteriori* de ocurrencia aplicando la regla de decisión de Bayes de *mínima razón de error*:

$$\hat{s} = \arg \max_{1 \leq k \leq S} P(\lambda_k \mid X) = \arg \max_{1 \leq k \leq S} \frac{p(X \mid \lambda_k)}{p(X)} P(\lambda_k) \tag{18}$$

La ec. 17 representa la *pdf* condicional de un locutor conocido S_k, representado por su modelo λ_k, dada la secuencia de vectores de rasgos $X = (\vec{x}_1, \vec{x}_2, \vec{x}_3, ..., \vec{x}_T)$ de la expresión de habla de origen desconocido. Para obtener la máxima *pdf* condicional de cada locutor S_k, la secuencia de vectores de rasgos *X* de entrenamiento, utilizada para clasificar al locutor S_k, debe ser la misma usada como secuencia de vectores de rasgos *X* de prueba. Esto asegura la incertidumbre mínima de *S* dado *X*, y la entropía mínima *H(S/X)*, entonces $MI(S;X) \Rightarrow H(S)$, la mayor información mutua *MI* posible.

En este reporte la *MI* entre la identidad del locutor y el conjunto de rasgos SDC, obtenido de su

respectiva expresión de habla, será evaluada en un experimento de reconocimiento de locutor usando la misma secuencia de vectores de rasgos en entrenamiento y prueba; para diferentes combinaciones de parámetros de los rasgos SDC.

10.4 Información Mutua del conjunto de rasgos cepstrales en reconocimiento de locutor

Una expresión de habla correspondiente a un locutor contiene la máxima información posible sobre su identidad, y cada paso del proceso de extracción de rasgos generalmente disminuye la información de la identidad del locutor. Una función invertible no reducirá la información, mientras que las funciones no-invertibles sí lo hacen [46]. El proceso de extracción de rasgos puede entonces no proveer nueva información sobre el locutor, pero si reducirá la complejidad del clasificador.

La contribución del rasgo cepstral y el rasgo Δ a la información del locutor es:

- Coeficientes cepstrales: los MFCC [2] son el conjunto estándar de rasgos para el reconocimiento del locutor. En el proceso de computar los MFCC, el espectro de magnitud, la deformación en frecuencia Mel, y la obtención del vector cepstral, son funciones no-invertibles y por tanto reducirán la información.
- Rasgo cepstral Δ o rasgo SDC: cualquier rasgo cepstral Δ puede computarse del propio rasgo cepstral. Entonces ellos no contienen más información de la que ya contiene el rasgo cepstral, y no se debe alcanzar una ganancia adicional, al utilizarlos en conjunto.

Estas conclusiones son válidas para reconocimiento del locutor solo desde un punto de vista teórico. La teoría nos recuerda también que cualquier rasgo obtenido en un sistema práctico no es óptimo en relación con la información del locutor, con respecto a la información contenida en la expresión de habla original [47]. Cualquier sistema práctico puede tener métodos de pos procesamiento como de supresión de ruidos, normalización de rasgos, combinación de rasgos MFCC con Δ o SDC, etc., para incrementar el rendimiento de la clasificación.

11 Experimento para evaluar la Información Mutua entre el rasgo SDC y la identidad del locutor

Para evaluar la *MI* entre el rasgo SDC y la identidad del locutor fue implementado un experimento con las frases de la base NIST 2001 Ahumada [22]. La dimensionalidad del vector de rasgos MFCC + Δ será de 24, y la dimensionalidad del vector de rasgos SDC será determinada por *kN*, siendo 24, 36 y 48 para k = 2, 3 y 4, respectivamente.

Las combinaciones de vectores de rasgos evaluadas en el experimento fueron:

1. Vector de rasgos MFCC + Δ con *D=1,2 and 3*, como línea base.
2. Vector de rasgos SDC*(12, D, P, k)* con las combinaciones:
   - *P=1, 2, 3, 4, y 5,*
   - *D=1, 2, y 3,*
   - *k=2, 3, y 4.*

11.1 La Información Mutua del vector SDC

La *MI* entre la identidad del locutor y una combinación de rasgos SDC fue evaluada en un experimento de reconocimiento de locutor utilizando las mismas expresiones para entrenamiento y prueba. Las muestras de habla fueron obtenidas de un conjunto de locutores S_k , *k=1,...,50,* en la sesión T1, concatenando las 10 frases balanceadas de cada locutor. Los vectores de rasgos cepstrales normalizados MFCC, Δ y SDC se obtuvieron como se explica en el epígrafe 2.2, constituyendo el vector X_i, *i=1,...,50* para cada locutor; posteriormente se obtienen los modelos GMMs $\lambda_1, \lambda_2, \ldots \lambda_{50}$ para cada locutor como

se explica en el epígrafe 10.3.

Entonces, usando los mismos vectores de rasgos del entrenamiento pero ahora como pertenecientes a expresiones desconocidas, se obtienen la *probabilidad a posteriori* de ocurrencia de cada uno de los *N* locutores representadas por sus modelos λ_k, *k=1,...,50* como se explica en el epígrafe 10.3. Este experimento fue repetido para cada uno de las combinaciones de vectores de rasgos mencionadas anteriormente.

Se obtiene una matriz de 50 modelos $\lambda_1, \lambda_2, \ldots \lambda_{50}$ frente a 50 secuencias desconocidas de vectores de rasgos Xi, *i=1,..., 50*, cada uno de los elementos de la matriz contiene la *pdf* condicional de cada locutor S_k respecto a cada secuencia de vectores de rasgos X_i como se muestra en la Tabla 14.

Tabla 14. Matriz con la *pdf condicional* de cada locutor *Sk* respecto a cada secuencia de vectores de rasgos *Xi*

Speaker	*model*	X_1	X_2	...	...	...	X_{50}
$\boldsymbol{S_1}$	$\boldsymbol{\lambda_1}$	$p(S_1/X_1)$	$p(S_1/X_2)$				$p(S_1/X_{50})$
$\boldsymbol{S_2}$	$\boldsymbol{\lambda_2}$	$p(S_2/X_1)$	$p(S_2/X_2)$				$p(S_2/X_{50})$
...	...						
...	...						
...	...						
$\boldsymbol{S_{50}}$	$\boldsymbol{\lambda_{50}}$	$p(S_{50}/X_1)$	$p(S_{50}/X_2)$				$p(S_{50}/X_{50})$

Para garantizar que $\sum_{k=1}^{50}\left(p(S_k / X_i) \Big/ \sum_{k=1}^{50} p(S_k / X_i) \right) = 1$, para cada secuencia de vectores de rasgos X_i, cada *pdf* condicional se normaliza respecto a la suma de cada columna.

La entropía condicional de 50 locutores dado una secuencia de vectores de rasgos X_i se obtiene evaluando la ec. (13) en cada columna de la matriz:

$$H(S_k / X = X_i) = -\sum_{k=1}^{50} p(S_k / X_i) \log p(S_k / X_i) \tag{19}$$

y la entropía condicional media de los 50 locutores dado el conjunto de secuencias de vectores de rasgos Xk se obtiene evaluando la ec. (15) para todas las columnas:

$$H(S/X) = \sum_{k=1}^{50} p(X = X_k) H(S_k / X = X_i) = -\sum_{k=1}^{50} p(X_k) \sum_{k=1}^{50} p(S_k / X_i) \log p(S_k / X_i) \tag{20}$$

Considerando una distribución normal de ocurrencia de cada secuencia de vectores de rasgos, entonces $p(X_k) = 1/50$.

Teniendo en cuenta que cada locutor S_k, es seleccionado aleatoriamente con una probabilidad uniforme, de un conjunto de 50 locutores, entonces:

$$p(S_k) = 1/50 \Rightarrow H(S) = -\sum_{k=1}^{50} \frac{1}{50} \log \frac{1}{50} = \log 50 = 3.910 \tag{21}$$

La evaluación de la *MI* (ec. 12) para diferentes combinaciones de vectores de rasgos, se muestra en la Tabla 15, recordando que el parámetro *k=1* representa el rasgo MFCC + Δ, la línea base:

Tabla 15. Información Mutua del vector SDC para cada combinación de *D, P y k*

D	K	1	2	3	4
1	P=1	0.38	0.15	0.32	0.54
	P=2	0.38	0.15	0.26	0.39
	P=3	0.38	0.13	0.22	0.31
	P=4	0.38	0.12	0.19	0.25
	P=5	0.38	0.11	0.17	0.22
2	P=1	0.48	0.27	0.62	1.18
	P=2	0.48	0.26	0.61	1.04
	P=3	0.48	0.28	0.61	0.99
	P=4	0.48	0.27	0.54	0.86
	P=5	0.48	0.25	0.46	0.65
3	P=1	0.54	0.44	0.87	1.43
	P=2	0.54	0.38	0.94	1.57
	P=3	0.54	0.40	0.84	1.42
	P=4	0.54	0.41	0.84	1.31
	P=5	0.54	0.40	0.84	1.25

Conclusiones previas pueden obtenerse de estos resultados:

- El rasgo SDC con *k=2* tiene peor *MI* que el correspondiente rasgo MFCC + Δ: el parámetro *k=2* no aporta suficiente información sobre el locutor.
- En general, los rasgos MFCC +Δ y SDC con *D=1* tienen la peor *MI*: este valor de *D* no aporta tampoco suficiente información sobre el locutor.
- *La MI* esta directamente correlacionada con la duración de la pendiente espectral generalizada *D*, y el número *k* de rasgos Δ concatenados.
- El rasgo SDC con las combinaciones de parámetros *(12,2,P,4),(12,3,P,3),(12,3,P,4)* tienen la mejor *MI*, viendo un decremento de la *MI* a medida que el contexto de tiempo se incrementa; esto significa que existe un contexto de tiempo óptimo del rasgo SDC correspondiente con su comportamiento seudo-prosódico, entre 160 y 340 ms. (Ver la Fig. 16)

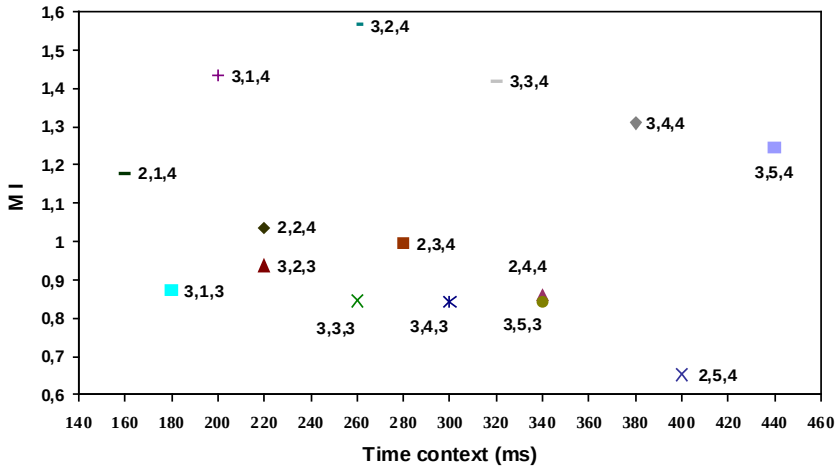


Fig. 16. Información Mutua de las combinaciones de parámetros *(12,2,P,4), (12,3,P,3), (12,3,P,4)* del rasgo SDC en relación con el contexto de tiempo que abarca el rasgo.

12 Experimento de verificación del locutor usando el rasgo MFCC y combinaciones de rasgos SDC con mayor Información Mutua

La evaluación de los resultados previos se llevó a cabo en un experimento de verificación de locutor bajo las mismas condiciones explicadas en los epígrafes 7.1 y 7.2. Los vectores de rasgos cepstrales normalizados MFCC, Δ y SDC se obtuvieron como se explica en el epígrafe 2.2. Se utilizaron como modelos de locutores $\lambda_1, \lambda_2, \ldots . \lambda_{50}$, los previamente obtenidos en el epígrafe 11.1. Se realizó la prueba como se explica en el epígrafe 7.2.2.

Los resultados de EER (%) y DCF (%) del experimento, para cada una de las combinaciones de rasgos, se muestran en la Tabla 16 y en la Tabla 17, respectivamente, recordando que el parámetro *k=1* representa al rasgo MFCC + Δ, la línea base:

Tabla 16. EER (%) de verificación del locutor para cada combinación de *D, P y k*, del rasgo SDC

D	k	1	2	3	4
1	P=1	15.3	18.5	15.8	15.4
	P=2	15.3	16.8	22.1	14.4
	P=3	15.3	16.2	22.3	15.7
	P=4	15.3	18.2	17.6	17.8
	P=5	15.3	18.3	17	18.9
2	P=1	14.4	14.6	14.	12.4
	P=2	14.4	13.4	13.7	11.8
	P=3	14.4	13.8	13.2	12.3
	P=4	14.4	14.5	13.9	13.6
	P=5	14.4	14.2	14.1	14.2
3	P=1	13.1	15	13.2	12.3
	P=2	13.1	13.6	12.4	11.4
	P=3	13.1	14	12.2	10.6
	P=4	13.1	13.6	11.7	11
	P=5	13.1	13.5	12.8	12.6

Tabla 17. DCF (%) de verificación del locutor para cada combinación de *D, P y k*, del rasgo SDC

D	K	1	2	3	4
1	P=1	5.34	6.99	6.75	6.57
	P=2	5.34	6.65	9.14	6.59
	P=3	5.34	6.43	8.9	7.09
	P=4	5.34	6.92	7.26	8.04
	P=5	5.34	7.21	7.76	8.44
2	P=1	5.37	5.98	6.4	5.8
	P=2	5.37	5.56	5.62	5.75
	P=3	5.37	5.43	5.4	5.81
	P=4	5.37	5.77	5.56	5.92
	P=5	5.37	5.62	6.03	6.36
3	P=1	5.38	6	5.96	5.6
	P=2	5.38	5.56	5.55	5.45
	P=3	5.38	5.48	5.4	5.48
	P=4	5.38	5.34	5.37	5.36
	P=5	5.38	5.47	5.78	6.03

Las figuras 17, 18 y 19 reflejan el comportamiento combinado del EER y la DCF para cada combinación de rasgo SDC, manteniendo fijo el parámetro *k*:

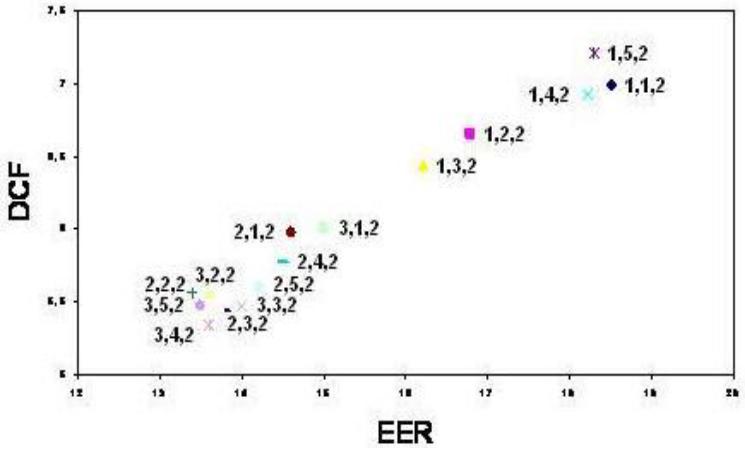


Fig. 17. *EER vs. DCF con SDC(12,D,P, 2)*

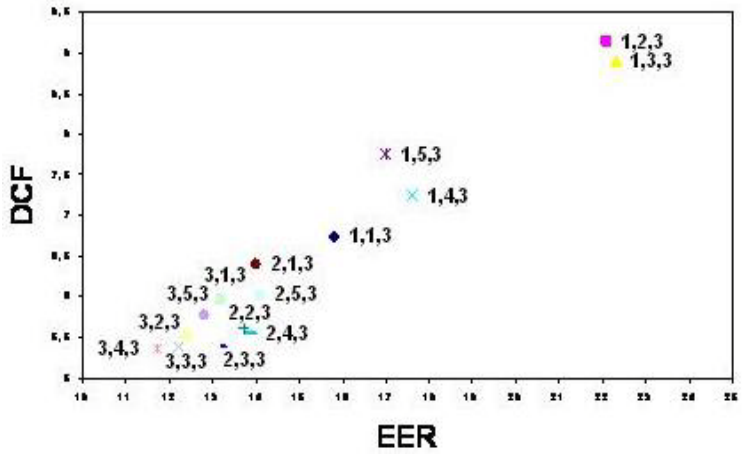


Fig. 18. *EER vs. DCF* con SDC(12,D,P, 3)

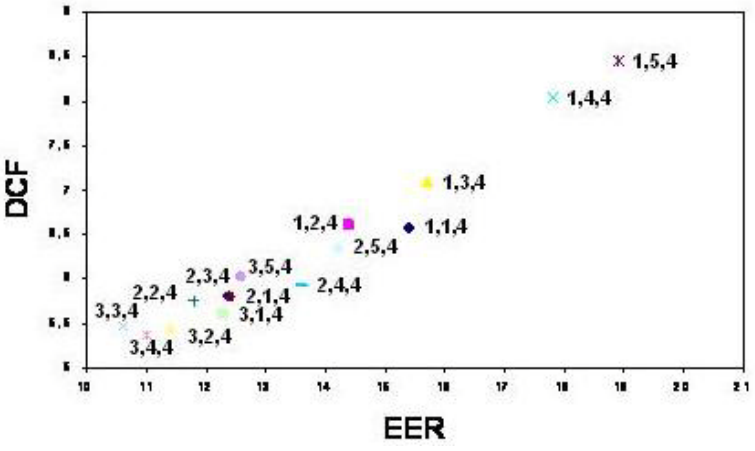


Fig. 19. *EER vs. DCF* con SDC*(12,D,P, 4)*

El comportamiento combinado del EER con la DCF brinda otras conclusiones preliminares:

1. Existe una relación linear entre la DCF y el EER, para cualquier *k.*
2. El EER decrece cuando *k* se incrementa, para cualquier *P* y *D,* en general el EER decrece cuando D se incrementa.
3. La MI y el EER parecen estar correlacionados con la duración de la pendiente espectral D, y el número *k* de rasgos Δ concatenados.
4. El rasgo SDC con k=*2* posee el peor EER, superior a un 13%, confirmándose el resultado obtenido con la Información Mutua. El valor *k=2* no revela el carácter seudo prosódico del rasgo SDC.
5. El rasgo SDC con *D*=1 posee el peor comportamiento DCF-EER para cualquier combinación de *P* y *k,* confirmándose el resultado obtenido con la Información Mutua. Este valor *D=1* no provee información acerca de la dinámica espectral del habla.
6. El rasgo SDC con la combinación *(12, 2, 2, 2)* posee el más bajo EER de entre todos los rasgos SDC con *k=2,* validando su elección por los autores, en todos los experimentos anteriores [23,35,41]
7. Se observa cierta relación inversa entre la Información Mutua y el EER. El rasgo SDC con las combinaciones (12,2,P,3), (12,2,P,4), (12,3,P,3) y (12,3,P,4), muestra el mejor comportamiento DCF-EER, confirmándose el resultado obtenido con la Información Mutua.

Las curvas DET de los experimentos de verificación del locutor con MFCC+Δ y las combinaciones de rasgos SDC (12,2,P,3), (12,2,P,4), (12,3,P,3) y (12,3,P,4), se muestran en las fig. 20, 21, 22 y 23, respectivamente:

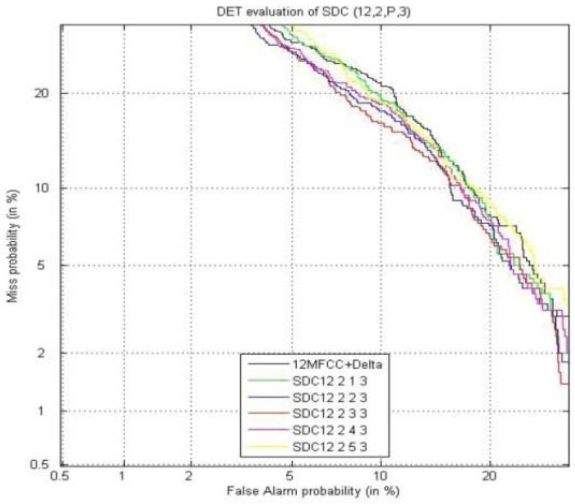


Fig. 20. Verificación del locutor con *MFCC+* Δ y combinación *SDC (12,2,P,3)*

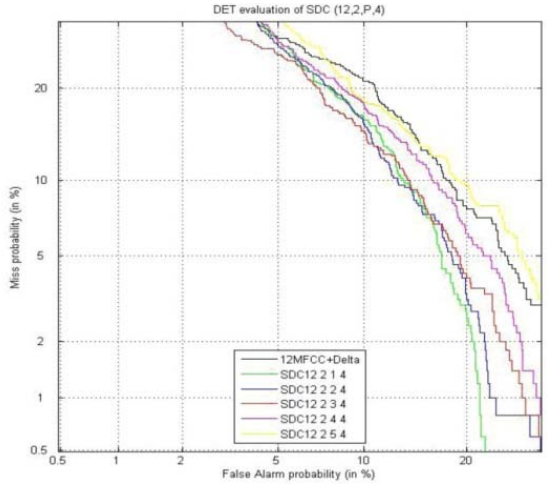


Fig. 21. Verificación del locutor con *MFCC+* Δ y combinación *SDC (12,2,P,4)*

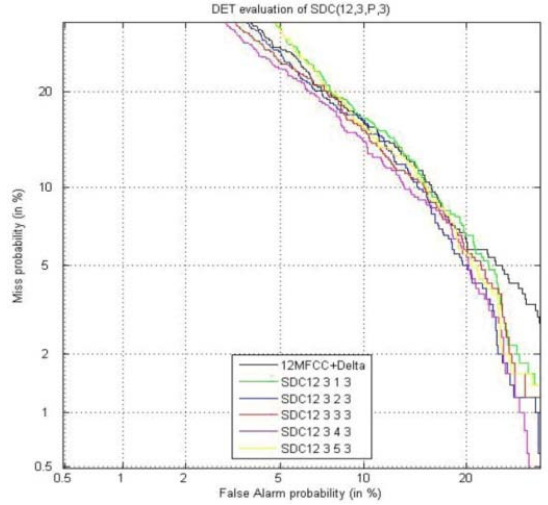


Fig. 22. Verificación del locutor con *MFCC+* Δ y combinación *SDC (12,3,P,3)*

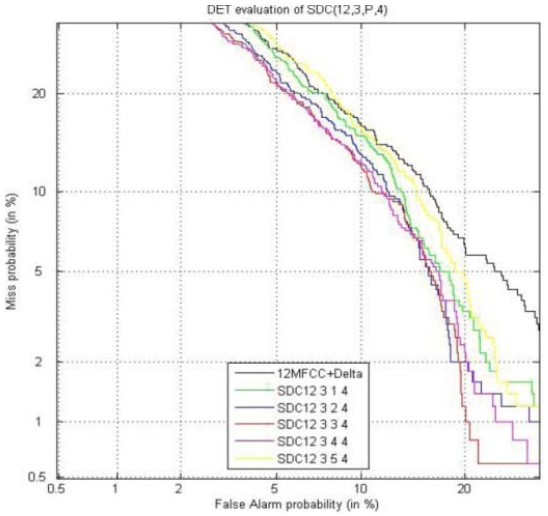


Fig. 23. Verificación del locutor con *MFCC+* Δ y combinación *SDC (12,3,P,4)*

Para dichas combinaciones del rasgo SDC, el valor de Información Mutua y el % de reducción del EER respecto al rasgo MFCC+ Δ se muestra en la Tabla 18:

Tabla 18. *MI* y % de reducción de *EER* respecto a la línea base para las combinaciones del rasgo *SDC*

D	P	k=3			k =4		
		MI	EER	%	MI	EER	%
2	1	0.62	14.	2.7	1.18	12.4	13.8
	2	0.61	13.7	4.8	1.04	11.8	18.0
	3	0.61	13.2	8.3	0.99	12.3	14,6
	4	0.54	13.9	3,4	0.86	13.6	5.5
	5	0.46	14.1	2.1	0.65	14.2	1.4
3	1	0.87	13.2	-0.1	1.43	12.3	6.1
	2	0.94	12.4	5.3	1.57	11.4	12.9
	3	0.84	12.2	6.8	1.42	10.6	19.0
	4	0.84	11.7	10.6	1.31	11	16.0
	5	0.84	12.8	2.3	1.25	12.6	3.8

De las curvas DET y de la tabla 18 podemos concluir:

- En general, el rasgo SDC presenta mejores EER que el rasgo MFCC + Δ para cualquier P.
- El peor resultado obtenido con el rasgo SDC es con P=5, lo que indica que un excesivo desplazamiento temporal entre los rasgos Δ que se concatenan, reduce o desaparece el necesario solapamiento entre las tramas de tiempo del cálculo del Δ.
- El resultado más interesante es el referido al mejor resultado de EER que se obtiene para cada combinación de D, P y k (celdas grises en la Tabla 18). Para una D (pendiente espectral) y una k (número de rasgos Δ que se concatenan) habrá un valor óptimo de P (separación entre rasgos Δ que se concatenan), donde se obtendrá el mejor EER: si D se incrementa o k decrece el mejor resultado será obtenido con una mayor P.
- El rasgo SDC con combinación (12,2,2,4) refleja una reducción relativa del EER del 18 %, y el rasgo SDC con combinación (12,3,3,4) refleja una reducción relativa del EER del 19 %, respecto a la línea base con el rasgo MFCC + Δ.

Los resultados de verificación de locutores confirman la evaluación previa de la Información Mutua entre la identidad de cada locutor y sus diversas combinaciones de rasgos SDC del habla, realizada en el epígrafe 11. Dicha evaluación nos permitió seleccionar la mejor combinación de parámetros del rasgo SDC que le permitiera enfrentar con éxito la variabilidad de canal y del micrófono en el experimento de verificación de locutores. El rasgo SDC con la combinación de parámetros (12,2,P,3), (12,2,P,4), (12,3,P,3) y (12,3,P,4) mostró el mejor comportamiento combinado MI-EER, reflejando en general, una relación inversa entre ambos, a mayor MI, menor EER: si la información mutua que posee el rasgo es mayor, más efectivo será.

Una parte de estos resultados fueron presentados y publicados en las memorias del 10mo Congreso Interspeech 2009, celebrado en Brigthon. [48]

13 Conclusiones

Los resultados mostrados en este reporte, que han sido defendidos y publicados en varios eventos internacionales, confirman que el rasgo SDC posee robustez, tiene carácter seudo prosódico y posee mayor información de la identidad del locutor que el rasgo MFCC+Δ, por lo que debe ser considerado como una alternativa al mismo, sin costo computacional adicional, permitiendo reducir los efectos de variabilidad del canal y del micrófono del teléfono, en verificación de locutores.

Los resultados mostrados permiten explorar nuevas aplicaciones del rasgo SDC del habla en aplicaciones de detección y seguimiento de locutores, teniendo en cuenta su simplicidad computacional.

Referencias bibliográficas

1. A. Acero “Acoustical and Environmental robustness in automatic speech recognition”. Kluwer Academic Publisher, 1993.
2. S. Davis, P. Mermelstein. “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. IEEE Transactions on ASSP. Vol 28 (4), pp. 357–366, 1980.
3. A. Adami, R. Mihaescu, D.A. Reynolds, J.J. Godfrey: “Modeling Prosodic Dynamics for Speaker Recognition”. Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003.
4. D. Reynolds, W. Andrews J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, B. Xiang. “The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition”. Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, pp. 784-787, 2003.
5. S. Furui. “Cepstral analysis for automatic speaker verification”, IEEE Transactions on ASSP, Vol. 29(3), pp 342 - 350, 1981.
6. C. Bernasconi. “On instantaneous and transitional spectral information for text-dependent speaker verification”, Speech Communication, Vol. 9(2), pp 129-139, April 1990.
7. F.K. Soong, A.E. Rosenberg. “On the use of instantaneous and transitional spectral information in speaker recognition”, IEEE Transactions on ASSP, Vol. 36(6), pp 871-879, June 1988.
8. P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, J.R. Deller. “Approaches to language identification using Gaussian Mixture Models and shifted delta cepstral features”. Proc. of Int. Conf. of Spoken Language Processing, pp. 89-92, 2002.
9. P.A. Torres-Carrasquillo, E. Singer, T.P. Gleason, W.M. Campbell, D.A. Reynolds. “Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Recognition”. Proc. of European Conf. on Speech Communication and Technology, pp. 1345-1348, 2003.
10. P.A. Torres-Carrasquillo, T.P. Gleason, D.A. Reynolds. “Dialect identification using Gaussian Mixture Models”. Proc. of Odyssey: The Speaker and Language Recognition Workshop, pp. 297- 300, 2004.
11. P.A. Torres-Carrasquillo, E. Singer, W.M. Campbell, D.A. Reynolds. “Language recognition with support vector machines”. Proc. of Odyssey: The speaker and Language Recognition Workshop, pp. 41- 44, 2004.
12. F. Allen. "Automatic Language Identification", PhD Thesis, University of New South Wales, Sydney, Australia, 2005.
13. J. Lareau. “Application of Shifted Delta Cepstral Features for GMM Language Identification”, MsC Thesis, Rochester Institute of Technology, USA, 2006.
14. T. Wu, D. Van Compernolle, J. Duchateau, Q. Yang, J-P.Martens. “Spectral Change Representation and Feature Selection for Accent Identification”. Proc. of MIDL 2004.
15. L. Besacier, J.F. Bonastre, C. Fredouille. “Localization and Selection of Speaker Specific Information with Statistical Modeling”. Speech Communication. Vol. 31(2-3), 2000.
16. B. Xiang. “Text-Independent speaker verification with dynamic trajectory model”. IEEE Signal Processing Letters. Vol. 10 (5), pp.141-143, 2005.
17. B.R. Wildermoth, K.K.Paliwal. “Reducing Inter-Session Variability With Transitional Spectral Information”. Proc. of Microelectronic Engineering Research Conference, 2001.
18. T. Kinnunen ”Spectral Features for Automatic Text-Independent Speaker Recognition”. Department of Computer Science, University of Joensuu, Finland, 2003.

19. B. Bielefeld, “Language identification using shifted delta cepstrum”, in Proc. Fourteenth Annual Speech Research Symposium, 1994.
20. O. Viikki and K. Laurila. “Cepstral domain segmental feature vector normalization for noise robust speech recognition” . Speech Communication, vol. 25, pp. 133–147, 1998.
21. C. Barras, J. Gauvain. “Feature and score normalization for speaker verification of cellular data”. Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol.2, pp. 49-52, 2003.
22. J. Ortega, J. Gonzalez, V. Marrero. “AHUMADA: A large speech corpus in Spanish for speaker characterization and identification”. Speech communication Vol 31, pp 255-264, 2000.
23. A. Martin, K.G. Doddington, M. Ordowski, M. Przybocki. “The DET curve assessment of detection task performance”. In Proc. of EuroSpeech, Vol. 4, pp. 1895–1898, 1997.
24. Y. Linde, A. Buzo , R. Gray. “An algorithm for vector quantizer design”. IEEE Transactions on Communications, Vol. 28, pp.84-95, 1980.
25. D.A.Reynolds, R.C. Rose. “Robust text-independent speaker identification using Gaussian mixture speaker models”. IEEE Transactions on SAP. Vol 3, pp. 72–83, 1995.
26. J. R. Calvo, R.Fernández, G. Hernández. “Application of Shifted Delta Cepstral Features in Speaker Verification”. Proc. of Interspeech, pp. 734-737, 2007.
27. N. K. Ratha, A. Senior, R.M. Bolle. “Automated biometrics”. 2nd International Conference of Advances in Pattern Recognition. LNCS 2013, pp. 445-474, 2001.
28. J. Ortega-Garcia, J. Bigun, D. Reynolds, J. Gonzalez-Rodriguez. “Authentication gets personal with biometrics”. IEEE Signal Processing Magazine, pp. 50-62, March 2004.
29. L.P. Heck, Y. Konig, M.K. Sonmez, M. Weintraub. “Robustness to telephone handset distortion in speaker recognition by discriminative feature design”. Speech Communication. Vol 31, pp. 181-192, 2000.
30. R. Mammone, X. Zhang, R. Ramachandran. “Robust speaker recognition”. IEEE Signal Processing Magazine, pp. 58-71, September 1996.
31. M.G. Rahim, B.H. Juang. “Signal bias removal by maximum likelihood estimation for robust telephone speech recognition”. IEEE Transactions on SAP, Vol. 4(1), pp. 19-30, 1996.
32. R. Teunen, B. Shahshahani, L.P. Heck. “A model based transformational approach to robust speaker recognition”. Proc. of International Conference of Spoken Language Processing, 2000.
33. K.K. Yiu, M.W. Mak, S.Y. Kung. “Environment Adaptation for Robust Speaker Verification”. Proc. of European Conference on Speech Communication and Technology, pp. 2973-2976, 2003.
34. R. Auckenthaler, M. Carey, H. Lloyd-Thomas. “Score normalization for text-independent speaker verification systems”. Digital Signal Processing, Vol 10 (1–3), pp. 42–54, 2000.
35. J.P. Campbell: “Speaker Recognition: A tutorial”. Proc. of the IEEE, Vol 85 (9), 1997.
36. R.O. Duda, P.E. Hart, D.G. Stork. “Pattern Classification”. John Wiley and Sons, Inc., 2001.
37. D. A. Reynolds, T. F. Quatieri, R. B. Dunn. “Speaker Verification Using Adapted Gaussian Mixture Models”. Digital Signal Processing, vol. 10, pp. 19-41, 2000.
38. J. R.Calvo, R.Fernandez, G. Hernandez. “Channel/Handset mismatch evaluation in a biometric speaker verification using shifted delta cepstral features”. Proc. of of the Iberoamerican Conference of Pattern Recognition, LNCS 4756, pp 96-105, 2007.
39. K. Bartkova, D.Le-Gac, D. Charlet, D. Jouvet : “Prosodic Parameter for Speaker Identification”. Proc. of International Conference of Spoken Language Processing, pp. 1197-1200, 2002.
40. S. Kajarekar, L. Ferrer, A. Venkataraman, K. Sönmez, E. Shriberg, A. Stolcke, H. Bratt, V.R. Gadde. “Speaker Recognition Using Prosodic and Lexical Features”. Proc. of the IEEE Speech Recognition and Understanding Workshop, pp. 19-24, 2003.
41. F.Weber, L. Manganaro, B. Peskin, E. Shriberg. “Using Prosodic and Lexical Information for Speaker Identification, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 141-144, 2002.
42. J. R. Calvo, D. Ribas, R. Fernández, G. Hernández. “Evaluation of linear relation between shifted delta cepstral features and prosodic features in speaker verification”. Proc. of the Iberoamerican Conference of Pattern Recognition. LNCS 5197, pp. 112-119, 2008.
43. T. M. Cover, J. A. Thomas. “Elements of information theory”, Wiley, 2006.
44. R.M. Gray. “Entropy and Information Theory”. Springer-Verlag, 2007.
45. A. P. Dempster, N. M. Laird, D. B. Rubin. “Maximum Likelihood from Incomplete Data via the EM Algorithm”. Journal of the Royal Statistical Society. Series B, Vol. 39(1), pp. 1-38, 1977.
46. T. Eriksson, S. Kim, H-G. Kang, C. Lee. “Theory for speaker recognition over IP”. Proc. of the International Conference of Spoken Language Processing, 2004.

47. T. Eriksson, S. Kim, H-G. Kang, C. Lee. “An information-theoretic perspective on feature selection in speaker recognition”. IEEE Signal Proc. Letters, Vol 12(7), pp. 500-503, 2005.
48. J. R. Calvo, R.Fernández, G. Hernández. “Selection of the Best Set of Shifted Delta Cepstral Features in Speaker Verification Using Mutual Information”. Proc. of Interspeech, pp. 2338-2341, 2009.

Anexo 1

Método para obtención de rasgos cepstrales normalizados del habla [2].

- La señal de habla es pre-enfatizada en frecuencia con un factor de 0.97, y se le aplica un esquema de eliminación de silencios basado en la energía.
- Se aplica una ventana de Hamming a tramas de 20 a 30 mseg. de longitud con un solapamiento entre 30 a 50 %, obteniéndose un espectro de potencia a corto término aplicando la FFT.
- El espectro de potencia se filtra con un banco de 30 filtros Mel-espaciados y el logaritmo de la energía de las salidas de los filtros se transforma con la transformada discreta del coseno, seleccionando N coeficientes cepstrales en frecuencia Mel (comúnmente 12). El rasgo cepstral cero no se usa. Por consiguiente, cada trama de señal se representa por un vector de N coeficientes MFCC.

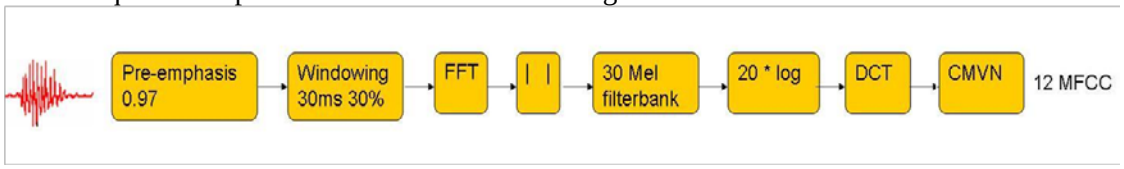
En el reconocimiento del habla y del locutor por canal telefónico, es necesario reducir la influencia de las condiciones medioambientales propias del canal, disminuyendo las diferencias en las condiciones acústicas entre entrenamiento y prueba. En años recientes se ha propuesto un método de normalización de rasgos robustos para reducir el ruido y/o los efectos del canal, la Normalización de la Media y Varianza Cepstral ("CMVN: Cepstral Mean and Variance Normalization") [20,21].

Este método normaliza la distribución espectral gruesa de las expresiones del habla y reduce la variabilidad espectral del habla del locutor a largo término. Asumiendo una distribución gaussiana de los rasgos cepstrales, el método CMVN normaliza cada componente cepstral según la expresión:

$$\hat{c}_i[n] = \frac{c_i[n] - \mu_i}{\sigma_i}$$

donde $c_i[n]$ y $\hat{c}_i[n]$ son los i-ésimos coeficientes del vector de rasgos en la trama n, antes y después de la normalización, respectivamente, y μ_i y σ_i son las estimaciones de media y varianza de la secuencia en tiempo de cada coeficiente c_i.

Un esquema del proceso de obtención de los rasgos MFCC normalizados se muestra:



Este esquema es el utilizado en nuestros experimentos.

RT_040, febrero 2011
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: Di. Alejandro Pérez Abraham
RNPS No. 2142
ISSN 2072-6287
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. No. 21812 e/218 y 222, Rpto. Siboney, Playa;
La Habana. Cuba. C.P. 12200
Impreso en Cuba