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

Métodos de compensación de ruido en reconocimiento de locutores

Ing. Dayana Ribas González,
Dr. C. José R. Calvo de Lara

RT_012 **febrero 2010**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
Ciudad de 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

Métodos de compensación de ruido en reconocimiento de locutores

Ing. Dayana Ribas González,
Dr. C. José R. Calvo de Lara

RT_012 **febrero 2010**

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

Índice

Métodos de compensación de ruido en reconocimiento de locutores

Ing. Dayana Ribas González, Dr. C. José R. Calvo de Lara

Centro de Aplicaciones de Tecnología de Avanzada, 7a #21812 e/218 y 222, Siboney, Playa, Ciudad de La Habana, Cuba
dribas@cenatav.co.cu

RT_ 012 CENATAV
Fecha del camera ready: 31 de julio de 2009

Resumen: En condiciones poco controladas, el ruido es un problema al que se tienen que enfrentar los Sistemas de Reconocimiento de Locutores para obtener resultados robustos. Este trabajo presenta un análisis taxonómico de los métodos que se utilizan en esta tarea, ofreciendo una explicación en síntesis de los métodos de compensación de ruido y profundizando en el método de Rasgos Perdidos (*Missing Features*, MF) por su potencialidad en la solución de este problema. Se presentan también los resultados de experimentos de verificación de locutores realizados utilizando algunas técnicas del método MF.

Palabras clave: reconocimiento de locutores, ruido, métodos de compensación, rasgos perdidos

Abstract: Noise is a problem that Speaker Recognition Systems have to face to obtain robust performances in slightly controlled terms. This job presents a taxonomic analysis of the methods used in this task, offering a synthetical explanation of the noise compensation methods and deepen in Missing Features Method (MF) because of its potentiality in the solution of this problem. The results of speaker verification experiments executed using some MF method techniques are introduced, too.

Keywords: Speaker Recognition, Noise, Compensation Methods, Missing Features

1. Introducción

Cualquier sistema que involucre la transmisión, adquisición o generación de voz es vulnerable a ser afectado por influencias que deterioran la calidad de la señal de voz. Ejemplos de estos son: el ruido de fondo a la hora de captar la señal de voz, los efectos del eco y las no-linealidades introducidas por los dispositivos electroacústicos que intervienen en el proceso. La siguiente figura ejemplifica las zonas propensas a ser afectadas por diferentes tipos de ruido en el proceso de captación de la voz.

Algunas de estas afectaciones se pueden controlar ajustando los parámetros del sistema, como la frecuencia de muestreo o el número de bits que representan una palabra. Otras se atacan con diferentes técnicas de Procesamiento Digital de Señales (DSP, por sus siglas en inglés)[1]. Los Sistemas Automáticos de Reconocimiento de Locutores (SARL) no están exentos de estos problemas. La precisión de sus resultados es atacada por dos factores fundamentales: los efectos en consecuencia de la desigualdad entre la sesión de prueba y la de entrenamiento y las afectaciones a causa del ruido [2]. Cuando la relación

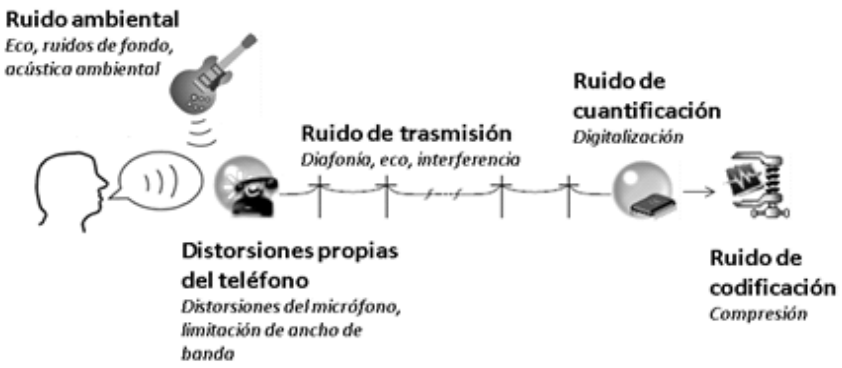

Fig. 1: Tipos de ruidos que afectan la señal de voz

señal-ruido (SNR, por sus siglas en inglés)[1] es menor que 10 dB, la señal de voz es muy sensible a corrupción por ruido. En ese caso, ni siquiera asegurar el mismo nivel de ruido en entrenamiento y prueba resuelve el problema[3].

En este trabajo se presenta un estudio taxonómico de los principales métodos que se encargan de mitigar el ruido en sistemas digitales de procesamiento de voz, orientado a las aplicaciones de reconocimiento de locutores. Se profundiza en los métodos de compensación de ruido por su relevancia en las aplicaciones con señales altamente corruptas.

2. Ruido

2.1. Tipos

2.1.1. Ruido ambiental

Cuando un transductor está captando la voz de un locutor, todos los sonidos que estén en el entorno se pueden clasificar como ruido ambiental. Por ejemplo: otras personas hablando cerca del lugar donde se capta la señal, un carro pasando o un ventilador.

Un ejemplo de esto es la reverberación, fenómeno derivado de la reflexión del sonido que consiste en una ligera prolongación del sonido una vez que se ha extinguido el original, debido a las ondas reflejadas. Estas ondas reflejadas sufrirán un retardo no superior a 100 msg, que es el valor de la persistencia acústica, tiempo que corresponde a una distancia recorrida de 34 metros a la velocidad del sonido (el camino de ida y vuelta a una pared situada a 17 metros de distancia). Cuando el retardo es mayor, entonces se le llama eco.

2.1.2. Distorsiones propias del equipo

El amplificador introduce distorsiones no-lineales que ocurren cuando éste se sobrecarga, produciendo acortamiento de la señal en casos extremos. También puede provocarse ruido

[1] La relación señal a ruido es la razón entre la potencia de la señal y la potencia del ruido que usualmente se expresa en dB.

por influencias térmicas, que puede entrar en el rango de frecuencias audibles, principalmente cuando la amplificación es de alta ganancia.

Los filtros paso bajo que se ocupan de evitar el solapamiento entre las tramas de la señal pueden provocar distorsión al eliminar componentes de alta frecuencia que estén dentro del espectro de la señal.

2.1.3. Ruido de transmisión

La realimentación de eco acústico ocurre cuando la comunicación es *full duplex*, la voz del locutor A es captada por el micrófono del locutor B y le llega de nuevo al locutor A y así sucesivamente. Este fenómeno sucede frecuentemente en teléfonos de manos-libres y produce graves problemas de inteligibilidad.

El *jitter* es una distorsión irreversible provocada por la desincronización del reloj para el muestreo de la señal, sin embargo el *jitter* sólo afecta cuando la razón de muestreo es alta, lo que no sucede usualmente en aplicaciones para voz y audio.

2.1.4. Ruido de cuantificación

En la conversión analógico-digital se introduce distorsión al aproximar los valores de las muestras de la señal.

2.1.5. Ruido de compresión

La codificación y compresión de la señal está sujeta al compromiso entre una implementación exhaustiva y la distorsión de la señal. La idea es potenciar el proceso en la información que el oído humano no es capaz de captar, pero incluso así se escapan algunas distorsiones sobre las zonas audibles.

2.2. Características del ruido y su manifestación en la señal de voz

El efecto del ruido en la señal de voz está muy relacionado con la naturaleza de éste. El ruido ambiental se puede modelar como una perturbación aditiva y no correlacionada con la señal de voz, mientras los efectos del canal de comunicación se combinan con la señal de forma convolucional. El ruido provoca la degradación de la calidad de la señal, así como la disminución de la inteligibilidad. Si es una perturbación no-lineal, provoca cambios en el espectro. La señal ruidosa puede tener una distribución no gaussiana, incluso cuando las fuentes de ruido y de señal son gaussianas. Esto puede afectar la etapa de creación de modelos en el reconocimiento de locutores, pues muchos de los métodos que aquí se usan asumen que la señal se describe como una distribución gaussiana. Por otro lado, si en el proceso de adquisición de la señal de voz estaban otras personas hablando en la escena, el reconocimiento del locutor deseado se complica más de lo usual. Cuando la media de la señal de ruido no cambia en el tiempo, se dice que éste es estacionario, en caso contrario será no estacionario. El ruido estacionario es más sencillo de enfrentar con algoritmos

computacionales que el ruido no estacionario. En este último caso cuesta trabajo encontrar un patrón que lo defina y de esta forma poderlo eliminar.

3. Métodos para enfrentar el ruido en reconocimiento de locutores

3.1. Taxonomía

Los métodos y técnicas para enfrentar el ruido son tan antiguos como el mismísimo procesamiento de señales. Para lidiar con señales ruidosas, los sistemas de reconocimiento de locutores emplean una gran diversidad de métodos (fig 2) que atacan el problema del ruido de diferentes formas. En cada uno de estos grupos se especifica en qué zona del reconocimiento actúa cada método.

Las soluciones para múltiples canales se emplean en los casos específicos de que las señales hayan sido grabadas con varios transductores. Se dirigen a elevar la SNR aprovechando la direccionalidad de los transductores y combinándolo con técnicas de DSP. Estos métodos son más costosos que el resto de los que se mencionan. Frecuentemente se utilizan en aplicaciones combinadas con la localización del hablante.

Las soluciones para canales simples agrupan a los métodos clásicos de DSP para el tratamiento de ruido. Actúan directamente sobre la señal, intentando eliminar el ruido y obtener una señal limpia para pasar al reconocimiento. Ejemplos de estos son los filtros en el dominio del tiempo, los filtros espectrales, los algoritmos recursivos, los filtros predictivos, las técnicas de estimación de ruido y señal, etc.

Los métodos de adaptación actúan sobre los modelos, haciendo iteraciones sucesivas hasta obtener el modelo que mejor represente al locutor, por ejemplo MLLR y MAP.

Entre estos se puede incluir la solución conocida por entrenamiento multicondición o multiestilo, que consiste en entrenar al sistema en diferentes condiciones de ruido para que a la hora de la prueba el sistema este preparado para recibir señales grabadas en diferentes condiciones.

Los métodos de normalización actúan en diferentes secciones del proceso de reconocimiento. En la extracción de rasgos se han desarrollado parametrizaciones lo más inmunes posible al ruido. Por ejemplo los rasgos PLP presentan la mayor robustez ante el ruido alcanzada con rasgos cepstrales, incluso aplicando un proceso de filtrado en el tiempo (RASTA-PLP) se obtienen mejores resultados aún. También es común la eliminación de la media y la varianza cepstral en los rasgos. Por su parte, la normalización de la puntuación se encarga de reducir la diferencia entre las condiciones en entrenamiento y prueba, a través de la búsqueda de un umbral global independiente del locutor, para el proceso de toma de decisión.

Generalmente los sistemas de reconocimiento de locutores emplean métodos clásicos de DSP, técnicas de adaptación y normalización de forma simultánea a lo largo del proceso. Sin embargo, todas estas soluciones tienen un límite de efectividad para enfrentarse a señales con un alto nivel de ruido. En la práctica, las señales que tengan una SNR menor de 10 dB aproximadamente, requieren de un proceso de compensación más específico para obtener buenos resultados en el reconocimiento de locutores. Cuando se trata de aplicaciones con

¿Cómo enfrentar el ruido en Reconocimiento de Locutores?

Métodos de adaptación
a) Entrenamiento multiestilo[4]
De modelos:[1]:
a) MLLR: Maximum Likelihood Linear Regression
b) MAP: Maximum A Posteriori Adaptation
c) EMAP: Extended Maximum A Posteriori Adaptation

Métodos de normalización
De rasgos:
a) CMN: Cepstral Mean Normalization [1]
b) CMVN: Cepstral Mean and Variance Normalization [1]
c) PLP: Perceptual Linear Prediction [5]
d) RASTA [6]
De modelos:
a) FA: Factor Analysis[7]
b) NAP: Nuisance Attribute Projection [8]
De puntuación:[9]:
a) T-norm: Test Normalization
b) Z-norm: Zero Normalization

Soluciones en canales simples
En la señal:
a) Filtrado en el dominio del tiempo
b) Filtrado espectral
c) Estimación de voz y ruido
d) Filtros Predictivos
e) Algoritmos recursivos
f) Redes neuronales
g) Descomposición de subespacios

Soluciones en múltiples canales
En la señal:
a) Arreglos de micrófonos[10]

Algoritmos de compensación
En la señal:
a) Spectral Substraction[1]
b) Filtro de Wiener [11]
En los rasgos:
a) VTS: Vector Taylor Series Compensation []
En los modelos:
a) PMC: Parallel Model Combination[12]
b) JA: Adaptación Ambiental Jacobiana [13]
En todo el proceso de reconocimiento:
a) Missing Fatures [14]

Fig. 2: Métodos para enfrentar el ruido en reconocimiento de locutores.

señales verdaderamente corruptas, por ejemplo en el campo forense, es necesario aplicar técnicas más contundentes y orientadas a lidiar con ruido.

Los algoritmos de compensación son técnicas especificas para el tratamiento de señales muy afectadas por ruido. Actúan en diferentes zonas del proceso de reconocimiento: en la señal, en los rasgos y en los modelos. En general, las compensaciones de señal y de rasgos son más sencillas y eficientes de implementar, pero la compensación de modelos da mejores resultados de robustez. Un caso muy común en las aplicaciones de reconocimiento de locutores, es que no se cuente con la certeza de la cuantía de corrupción que tiene la señal de voz. En este caso recientemente se han obtenido alentadores resultados en el plano experimental utilizando la teoría de Rasgos Perdidos (Missing Features, MF) [2,15]. En los siguientes epígrafes se profundiza en los algoritmos de compensación y especialmente en el método de Rasgos Perdidos.

3.2. Substracción espectral

Es uno de los primeros métodos creados para realzar una señal frente al ruido [16]. Existen dos tipos de substracción espectral: la lineal, también conocida por substracción espectral a secas, y la no-lineal [1].

El método de substracción espectral lineal (LSS, por sus siglas en inglés) consiste en cancelar el ruido aditivo no correlacionado a partir de la señal de voz corrupta; para esto se estima la cuantía de ruido que porta la señal y luego se substrae de ésta para obtener un estimado de la señal limpia original. La estimación del espectro de potencia de ruido se realiza con diferentes técnicas, que en muchos casos se apoyan en las zonas de silencio, es decir, se asume que al inicio de la señal sólo aparece ruido y con este dato se inicializa la estimación del espectro de potencia de ruido. Finalmente, la compensación de la señal se realiza restando el espectro de potencia de ruido del espectro de potencia de la señal corrupta, hasta obtener un estimado del espectro de potencia de la señal limpia original.

A pesar de que este método es ampliamente utilizado, presenta varias limitaciones. Por ejemplo un problema muy común es que el estimado de la energía del espectro de ruido sea mayor a la energía de la señal corrupta correspondiente, resultando en valores negativos de energía en el estimado de la señal limpia original. En este caso, el estimado del ruido no es representativo del espectro de ruido real que está afectando a la señal, ocasionando que la substracción espectral no sea efectiva. Cuando esto sucede, el espectro estimado de la señal limpia original, en el mejor de los casos, presenta picos y ceros indeseados, muy molestos desde el punto de vista perceptual, fenómeno al que se le conoce como ruido musical. En el peor de los casos, la cuantía de ruido en vez de mitigarse se amplifica, distribuyéndose de otra forma en el espectro, fenómeno conocido por ruido residual y puede ser hasta más dañino que el ruido original que tenía la señal corrupta [17].

La substracción espectral no-lineal (NSS, por sus siglas en inglés) se dirige a robustecer el estimado del espectro de ruido, reduciendo la incidencia de varios problemas que presenta la substracción espectral lineal. Por ejemplo los valores negativos en zonas del espectro del estimado de la señal limpia original, se pueden eliminar forzando un mínimo positivo en la formulación de la NSS. Además, al aplicar NSS generalmente el estimado obtenido no presenta el ruido musical, muy frecuente en los resultados de la LSS.

El estimado del ruido se obtiene por la misma vía explicada para la LSS, pero en este caso la señal corrupta actúa como umbral a partir de la siguiente regla 1:

$$|N(t,s)| = \begin{matrix} |Y(t,s)| & si \quad |Y(t,s)| < |N(t,s)| \\ |N(t-1,s)| & otro \quad caso \end{matrix} \tag{1}$$

La característica esencial de la NSS es que separa el proceso de substracción espectral en la obtención de dos componentes tanto del estimado de ruido como de la señal ruidosa: un valor instantáneo en trama y subbanda y un valor a largo término[2]. La componente instantánea es el valor del espectro en trama y subbanda. La componente a largo término se calcula, haciendo un promedio pesado del espectro en la trama anterior y el espectro instantáneo en la trama actual. Se construye un filtro con ambas componentes, para formular la ecuación final de substracción espectral (ec. 2,3) que se encarga de compensar a la señal por trama.

$$|\hat{S}(t,s)| = H(t,s)|Y(t,s)| \tag{2}$$

$$\hat{S}(t,s) = |\hat{S}(t,s)|e^{iY(t,s)} \tag{3}$$

3.3. Filtros de Wiener

El filtro de Wiener es un filtro lineal que se utiliza en el tratamiento de señales de voz ruidosas en general, por lo que puede aparecer asociado a cualquier sistema de procesamiento de voz que se vea afectado por ruido. Específicamente en el reconocimiento de locutores, se utiliza para mejorar la SNR [11]. La respuesta al impulso de este filtro se diseña con el objetivo de minimizar el error cuadrático esperado entre la señal de voz limpia y la corrupta. Se puede formular tanto en el dominio del tiempo como en el de la frecuencia, aunque es más utilizado en esta última variante [1]. A continuación se presentan diferentes planteamientos del filtro de Wiener.

El filtro de Wiener en el dominio del tiempo se obtiene a partir de minimizar el Error Cuadrático Medio (MSE, por sus siglas en inglés) entre una señal y su estimado. El filtro se define como sigue:

$$h_0 = h_1 - R_y^{-1} r_{vv} \tag{4}$$

Como se observa en la ecuación 4 está formado por dos componentes h_1 y $R_y^{-1} r_{vv}$. Cada una tiene una función diferente desde el punto de vista físico: h_1 crea una réplica de la señal ruidosa (y) y $R_y^{-1} r_{vv}$ genera un estimado del ruido. La reducción del ruido se logra en dos pasos: primero se crea un estimado óptimo del ruido y luego se le resta a la señal ruidosa y(n), muestra a muestra. Este tipo de filtro logra reducir el ruido pero a la vez provoca distorsión en la señal de voz, motivo por el cual apareció el Filtro Subóptimo que logra controlar el compromiso entre la reducción de ruido y la distorsión de la voz.

El filtro Subóptimo consiste en una adecuación empírica del filtro de Wiener, con el objetivo de controlar el compromiso entre la reducción de ruido y la distorsión de la voz.

[2] Se refiere a más de 20ms de la señal, en muchos casos a más de una trama.

Logra mejorar la SNR a pesar de ser menos efectivo que el Filtro de Wiener en cuanto a la reducción de ruido.

También están los métodos de subespacio, que son un grupo de técnicas que consisten en obtener un estimado de la señal limpia (x) aplicando una transformación lineal al vector de muestras ruidosas (y). Como en el filtro de Wiener, estas técnicas se formulan a partir del algoritmo MMSE con rigor matemático, por lo que logra estimaciones óptimas. Sin embargo, a diferencia del filtro de Wiener, la transformación lineal de este caso se deriva de un problema de optimización restringida.

El filtro de Wiener también se puede plantear en el dominio de la frecuencia. Una vía para lograr esto es transformar el filtro de Wiener en el dominio del tiempo. En este caso, ambas representaciones del filtro (la del tiempo y la de la frecuencia) tienen exactamente el mismo comportamiento. Sin embargo, el filtro en la frecuencia se formula estimando directamente el espectro de habla limpia (x) a partir del espectro de habla ruidosa (y). El filtro resultante difiere del anterior en dos aspectos: el del tiempo es un filtro causal, mientras que el de la frecuencia no lo es; el del tiempo aplica una técnica de filtrado en toda la banda, mientras que el de la frecuencia aplica una técnica de análisis en subbanda, donde cada filtro de subbanda es independiente a los filtros correspondientes en otras bandas de frecuencia.

El filtro de Wiener en el dominio de la frecuencia incrementa la SNR a menos que todas las subbandas tengan igual valor de SNR. Ambas señales, el ruido y la voz, tienen la misma PSD bajo la condición de resolución dada. En este caso, el filtro de Wiener no puede distinguir al ruido de la voz y por lo tanto no puede mejorar la SNR. En el resto de los casos, el filtro de Wiener de subbanda puede mejorar la SNR.

3.4. Combinación Paralela de Modelos

La técnica de Combinación Paralela de Modelos (PMC, por sus siglas en inglés) surgió para incrementar la robustez ante el ruido en sistemas de reconocimiento de voz. Los primeros trabajos que la describieron se reportaron en la década del 90 [12,18]. Luego se empezó a utilizar en reconocimiento de locutores, en el 2001 se reportó el primer trabajo que implementó PMC para compensar el ruido aditivo en verificación de locutores dependiente del texto, modelando con HMM[3] [19]. A partir de aquí siguieron apareciendo trabajos que utilizaban a PMC para el tratamiento de señales afectadas por ruido en reconocimiento de locutores [20].

La Combinación Paralela de Modelos se aplica independientemente al método de clasificación (GMM[4], HMM, etc) que se utilice, aunque en la mayoría de los casos en que el reconocimiento es dependiente del texto el PMC se acompaña con HMM y cuando es independiente del texto se emplea GMM. PMC considera que la señal limpia y el ruido son distribuciones gaussianas, pero aunque estas no lo sean PMC asume que la combinación de estas dos, es decir la señal ruidosa resultante, sigue una distribución gaussiana. Típica-

[3] Modelos Ocultos de Markov, HMM (Hidden Markov Models) método de modelado estadístico en el tiempo que se utiliza en sistemas de reconocimiento de voz y locutores.
[4] Modelo de Mezclas Gaussianas, GMM (Gaussian Mixture Models.)

mente se necesitan muestras de ruido aislado para estimar adecuadamente los parámetros del PMC [21].

En esencia, PMC trata de transformar los vectores de media y las matrices de covarianza de la distribución acústica del modelo para hacerlos similares a la distribución ideal cepstral de la voz ruidosa. Se crea un estado combinado que representa a la señal ruidosa dentro del dominio cepstral convolucionando el estado HMM simple de la señal limpia con el estado HMM simple del ruido. Seguidamente se presenta un ejemplo.

Dados los estados gaussianos simples en el dominio cepstral denotados por σ_s y σ_n, que representan a la señal limpia y al ruido respectivamente. El vector de media y la matriz de varianza de σ_s y σ_n, se denota por: $\{\mu_s, \xi_s\}$, $\{\mu_n, \xi_n\}$. PMC crea un estado combinado que representa a la señal de voz afectada por el ruido en el dominio cepstral a través de una convolución. El proceso básico para conformar el método se muestra en la figura 3 y sus pasos son los siguientes:

- Se realiza la transformada inversa de los conjuntos de media y varianza $\{\mu_s, \xi_s\}$, $\{\mu_n, \xi_n\}$ para llevarlos al dominio espectral.
- Se convolucionan los datos que describen a la señal limpia con los que describen al ruido.
- El resultado de la convolución se devuelve al dominio cepstral.

Para llevar a cabo este proceso, PMC asume que se conocen de antemano el comportamiento del ruido y del canal; la señal de voz y el ruido son independientes y aditivas. Durante el proceso de combinación, la suma de dos distribuciones log-normales da una distribución log-normal [11,19].

El método se divide en dos clases que se conocen por PMC iterativo y no iterativo. La primera asume que la alineación del estado no se altera, permitiendo que varíe un estado es posible mejorar la distribución ruidosa. La segunda asume que la señal limpia no se altera por la adición de ruido. Esto implica que la señal ruidosa tiene una distribución aproximadamente gaussiana, lo cual es falso.

A modo de ventajas se puede señalar que PMC opera en el dominio cepstral, por lo que goza de las ventajas de la decorrelación de parámetros. Por otro lado, los parámetros dinámicos y estáticos que normalmente se incluyen en una representación cepstral para reconocimiento de habla se pueden compensar usando PMC. Además PMC maneja la idea de que la señal de voz ruidosa está compuesta por dos señales o modelos estocásticos simples, el del habla y el del ruido. Claro que estas técnicas asumen que exactamente dos fuentes están presentes en el modelo estocástico y que este ha sido entrenado para fuentes de ruido y de voz. Es decir que conoce el comportamiento de la señal ruidosa. Esto no siempre sucede en las situaciones que requieran procesamiento de señales de voz afectadas por ruido, por lo que no es posible utilizar PMC en determinadas situaciones objetivas.

3.5. Adaptación Ambiental Jacobiana

Es una técnica de compensación de modelos acústicos que se usa para adaptar los vectores de media del HMM, partiendo de determinadas condiciones de ruido a otras [13], con el objetivo de atenuar la desigualdad entre el entrenamiento y la prueba. Se emplean matrices Jacobianas para relacionar cambios en el ruido ambiental con cambios en modelo acústico

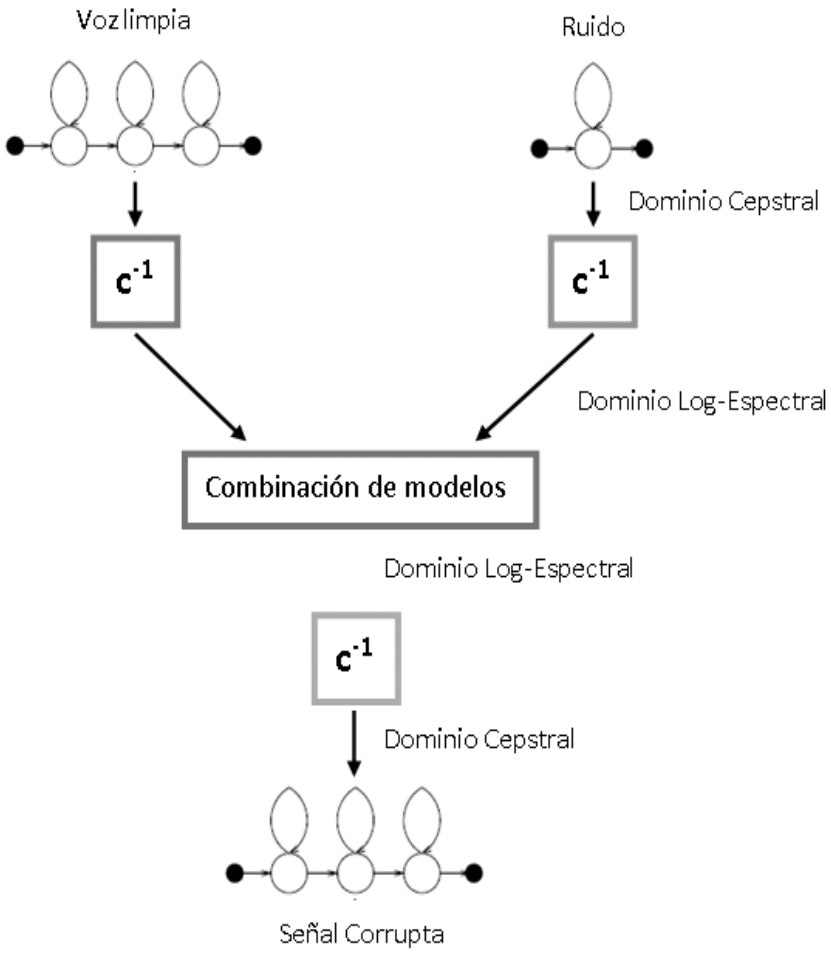

Fig. 3: Proceso PMC en HMM

de la señal ruidosa. La adaptación se basa en la diferencia entre el ruido presente en la fase de entrenamiento y la de prueba, sin embargo es necesario estimar una referencia de ruido para ambas fases, la más precisa sería la estimación del ruido y la mejor la adaptación de los modelos [22].

La transformación Jacobiana se conforma según la ecuación 5.

$$\hat{C}_{s+n} = C_{s+n} + \frac{\partial C_{s+n}}{\partial C_n}(\hat{C}_n - C_n) \tag{5}$$

Donde $\frac{\partial C_{s+n}}{\partial C_n}$ es la matriz Jacobiana; $\hat{C}_{s+n}$ es el vector de media del rasgo de habla ruidoso adaptado; C_{s+n} es el vector de media del rasgo de habla ruidoso original; C_n es el vector de rasgos ruidoso de referencia, es decir el ruido presente en la etapa de prueba y $\hat{C}_n$ es el vector de rasgos ruidoso del *target*, es decir el ruido presente en la etapa de reconocimiento.

Una forma más efectiva de utilizar la JA para reconocimiento del locutor es empleando parámetros FF (Frecuency filtering) [23], en este caso la matriz queda según a ecuación 6

$$\frac{\partial C_{s+n}}{\partial C_n} = Hdiag(\frac{\alpha N}{S+\alpha N})H^{-1} \quad (6)$$

Donde N representa el vector de energías (FBE, por sus siglas en inglés) de la referencia de entrenamiento del banco de filtros; S el vector FBE del modelo de habla ruidosa; H la transformación de la matriz FF; H^{-1} es la inversa de H; α la constante que se incluye para aliviar el problema de JA para grandes desmacheos entre el entrenamiento y la prueba [24] y diag()la matriz diagonal formada por los elementos del vector que está entre paréntesis. El cociente se calcula elemento por elemento.

En general en JA solo se aplica una referencia de ruido para calcular todas las matrices jacobianas y adaptar todos los modelos. En [25] se propone una modificación del método que emplea una referencia de ruido para adaptar cada modelo.

Esta técnica se usó para reducir la razón de error en verificación de locutores, ya que en sistemas de reconocimiento de locutores las frases dichas por unos locutores no se usaban para entrenar los modelos de otros. En la etapa de entrenamiento, una referencia de ruido se estima para cada locutor a partir de las señales de voz que se usaron para entrenar su modelo. Entonces las matrices Jacobianas (una para cada vector de media de cada mezcla de cada modelo) se calculan usando la ecuación 6, donde el vector N es el ruido específico de referencia FBE de cada modelo. En la etapa de prueba los modelos se adaptan usando la ecuación 5 donde el vector Cn es el vector cepstral de ruido específico de referencia del modelo a adaptar [22].

Experimentalmente JA ha demostrado brindar buenos resultados especialmente cuando sólo se tiene una pequeña muestra (de menos de 1 segundo) del ruido presente en la prueba. Es ventajosa en términos del costo computacional [13].

3.6. Método de Rasgos Perdidos

Previo a la década del 90 se trataba el ruido en señales de voz, ya sea para reconocimiento de palabra como de locutores, utilizando técnicas de substracción espectral, PMC y otras. Para esta fecha en la Universidad de Sheffield se estaba investigando en el modelado computacional del Análisis de la Escena de Audición (ASA, por sus siglas en inglés)[5]. En 1994 se comenzaron investigaciones [26] acerca de cómo el ASA pudiera utilizarse en la tarea de reconocer el habla en señales ruidosas. La situación que se presenta es la siguiente: dada una señal de voz afectada por un patrón de ruido no conocido por el sistema, reconocer las palabras o el locutor. Esta situación es muy común y conveniente para enfrentarla con enfoque de la teoría de rasgos perdidos.

A partir de aquí se comenzó a trabajar en cómo aplicar el método de rasgos perdidos (*Missing Feature*s, MF) en reconocimiento de voz bajo condiciones ruidosas, presentándose numerosos trabajos sobre el tratamiento y desarrollo de las técnicas que conforman el

[5] ASA, en inglés Auditory Scene Analysis. Consiste en separar las fuentes de sonido de una escena de audición y prestar especial atención a alguna de ellas. Los humanos tienen gran capacidad para conformar un análisis de este tipo.

método MF [26,27,28,29]. A partir de 1998 aparecen los primeros reportes [17,30,31], de cómo aplicar este enfoque al reconocimiento de locutores con señales ruidosas. En años posteriores, especialistas de varias universidades, bajo la tutoría del Laboratorio Lincoln del Instituto Tecnológico de Massachusetts (MIT, por sus siglas en inglés) MIT, presentaron trabajos sobre este tema [2,15,32].

Los rasgos perdidos no son más que datos ruidosos al límite de considerarse prácticamente sólo ruido [33]. En el reconocimiento de locutores se consideran rasgos perdidos las componentes espectro-temporales de la señal de voz que tienen un alto nivel de afectación por el ruido.

El método MF modela el efecto del ruido como una corrupción en cada una de las componentes que representan en tiempo y frecuencia a la señal de voz. Dichas componentes no son más que los rasgos que caracterizan a la voz y que se obtienen utilizando métodos de extracción de rasgos [14]. En muchos casos se utilizan los métodos acústicos por las ventajas que estos ofrecen [34], siendo el más frecuentemente empleado el Método de Extracción de Coeficientes Cepstrales en escala Mel, MFCC [6] [35,36,37].

MF clasifica a las componentes de la señal en confiables y no confiables. Las confiables son las que no están alarmantemente afectadas por el ruido, se conocen por componentes *R*, del término en inglés *reliable"*. Mientras que las no confiables son aquellas que están altamente afectadas por el ruido, al punto de en ocasiones considerarse sólo ruido y se denominan componentes *U*, del término en inglés *ünreliable"*. Las técnicas que se utilizan para determinar esta clasificación se apoyan en diferentes características de la señal de voz, por ejemplo: la SNR, los rasgos acústicos, los atributos perceptuales, los parámetros estadísticos, etc [14]. Estas consisten en determinar un umbral, basado en los recursos antes mencionados, que les permita decidir si una componente es *U* o *R*. A este proceso de clasificación se le conoce como procedimiento de enmascaramiento espectrográfico, y al resultado del mismo: máscara espectrográfica. Partiendo de este dato hay dos formas de realizar el reconocimiento, o se efectúa con esta nueva medida incompleta del espectro o se reconstruyen las zonas corruptas, a este paso se le conoce por compensación de MF.

MF agrupa un conjunto de técnicas que actúan sobre las distintas fases del sistema de reconocimiento de locutores, provocando que este mejore su robustez frente al ruido. Es por ésto que se puede definir como un método para enfrentar el ruido. Las técnicas que conforman al método se estructuran según el esquema que aparece a continuación 4.

El método consiste en dos procesos fundamentales, la identificación de las componentes *U* y la compensación de componentes U, luego se pasa al reconocimiento. Estos se ponen en práctica a través de diversas técnicas que se combinan en dependencia de las características de la situación.

3.6.1. Identificación de componentes no confiables

El paso más complicado del método de MF es estimar la máscara espectrográfica que identifica a las componentes espectrales U [14]. La estimación del nivel de corrupción se hace para cada trama *(t)* y subbanda *(s)*. Se puede hacer de varias formas por ejemplo: estimando la SNR de cada componente espectral, extrayendo rasgos que identifiquen la

[6] Mel Frequency Cepstral Coeficients, (MFCC por sus siglas en inglés).

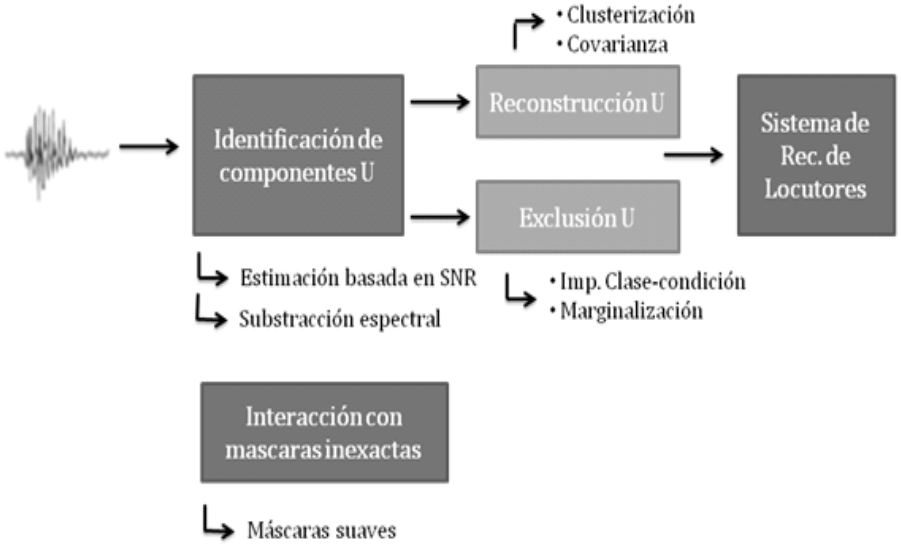

Fig. 4: Estructura de las técnicas que conforman el método MF.

señal sobre el ruido (utilizando un clasificador para decidir si es U o R) o empleando criterios de percepción.

A continuación se presentan métodos de estimación de máscaras basados en el criterio SNR, que son los que más se han utilizado en las aplicaciones de MF al reconocimiento de locutores. La SNR es la razón entre el espectro de potencia de la señal y el espectro de potencia del ruido que la afecta. Por lo tanto para calcular la SNR en los casos que no se conozca el ruido que afecta a la señal es necesario estimarlo. Existen muchos métodos para estimar el espectro ruidoso de una señal. Por ejemplo, una forma simple y muy usada es partir de las regiones de la señal que no contienen voz, también llamadas zonas de silencio, se promedia el espectro de potencia de estas tramas y se asume como espectro ruidoso, otra forma es utilizando una recursión de primer orden [38]. Cualquiera de estos u otro sirve como punto de partida para crear la máscara.

Máscaras Oráculo

Las máscaras que se obtienen a partir de los datos reales de SNR de la señal contaminada son conocidas por el término en inglés *Oracle Mask*. Estas se utilizan como medida de comparación de la efectividad que alcanza un determinado método de cálculo de máscara. Las máscaras Oráculo se calculan contando independientemente con la señal limpia y la señal de ruido que la contamina. El proceder es sencillamente calcular la SNR de cada componente a partir de la potencia de señal y potencia de ruido correspondientes, si $SNR > 1$ la componente se clasifica como R, si por el contrario $SNR < 1$ se clasifica como U.

Criterio de energía negativa

Este método fue introducido en el trabajo [17]. Consiste en obtener las componentes R a través de una estimación de la señal limpia que se logra llevando a cabo una substracción espectral. Se considera que la señal ruidosa *(Y)* está compuesta por una señal de voz limpia *(X)* y un ruido *(N)* según la siguiente ecuación:

$$Y(n) = X(n) + N(n) \tag{7}$$

De esta forma el espectro de potencia de la señal amplificada queda planteado según la siguiente regla de reducción de ruido:

$$|\hat{X}_m(w)|^2 = \begin{array}{ll} |Y_m(w)|^2 - |\bar{N}_m(w)|^2 & si \quad |Y_m(w)|^2 > |\bar{N}_m(w)|^2 \\ 0 & otro \quad caso \end{array} \tag{8}$$

Donde $|Y_m(w)|^2$ es el espectro de potencia de la trama de habla ruidosa y $|N_m(w)|^2$ es la potencia promedio de ruido estimada con la ayuda de un detector voz/pausa.

Una vez que la substracción sea calculada en el dominio espectral con la ecuación 7, se retorna al dominio del tiempo, obtiéndose la señal amplificada a través de la siguiente transformación:

$$\hat{x}(n) = IFFT[\hat{X}(w)e^{jargY(w)}] \tag{9}$$

Independientemente del algoritmo que se use para la substracción, se introduce un error de ruido residual como consecuencia de la diferencia que existe entre la señal original de voz limpia y la estimación de ésta generada en el proceso de SS anteriormente explicado. Este ruido es completamente diferente al ruido original *(N(n))* y en algunos casos puede causar más distorsión que éste.

Se puede decir que la substracción espectral hace una limpieza gruesa del contenido de ruido presente en el espectro de *Y(w)*. Las componentes que aún quedan se representan en *r(n)*, aunque en algunos casos se generan nuevos tonos en frecuencias aleatorias. Por lo que se puede concluir que la técnica clásica de SS mejora la SNR pero al precio de distorsionar la señal en algunos casos.

Criterio de SNR

Este métodos identifica las bandas espectrales como *U*, si la SNR estimada de alguna componente espectral se sitúa por debajo de los 0 dB. Los estimados de los espectros de la señal limpia y del ruido se obtienen a partir de la substracción espectral. Entonces, una componente espectral *Y(t,s)* se asumirá como *U*, si el espectro de potencia estimado de la señal limpia es menor que el de ruido:

$$\hat{X}(t,s) < \hat{N}(t,s) \tag{10}$$

Alternativamente Y(t,s) se considera *U* si:

$$\hat{X}(t,s) < 0{,}5Y(t,s) \tag{11}$$

En la práctica la mejor estimación de la máscara se obtiene cuando se combinan el criterio de energía negativa y el criterio de SNR [14]. En general se espera que esta técnica sea efectiva cuando el ruido es estacionario o pseudo estacionario. Para ruidos no-estacionarios o inestables la estimación del espectro de ruido se complica y esta técnica puede resultar en máscaras espectrográficas altamente inexactas.

Criterio MMSE

En este caso se obtiene una estimación de la señal limpia a partir del método de Estimación del Error cuadrático Medio (MMSE, por sus siglas en inglés).Este método fue definido por Drygajlo y El-Maliki [30]. La ecuación que define el método es la siguiente:

$$|\hat{X}_m(w) = G(w)Y(w) \tag{12}$$

Donde:
Y(w) magnitud del habla ruidosa
G(w) función de ganancia

$$G(w) = \frac{\sqrt{\pi}}{2}\sqrt{\frac{1}{SNR_{post}}\frac{SNR_{prio}}{1+SNR_{prio}}}F(SNR_{post}\frac{SNR_{prio}}{1+SNR_{prio}}) \tag{13}$$

$$F(x) = exp(\frac{-x}{2})[(1+x)I_0(\frac{x}{2}) + xI_1(\frac{x}{2})] \tag{14}$$

Donde I_0, I_1 son las funciones de bessel modificadas de orden 0 y 1; SNR_{post} es la SNR posterior que expresa la razón entre el espectro de potencia del habla ruidosa $|Y(w)|^2$ y el estimado de ruido durante las pausas $|N(w)|^2$; SNR_{prior} es la SNR previa es un parámetro dominante, la atenuación del ruido depende de la estimación de este parámetro, que se obtiene en la n-esima trama de cada subbanda a través de una estimación de máxima similitud (EM, por sus siglas en inglés).

A continuación se define como se estima la SNRprior empleando el método EM.

$$SNR_{prior} = \begin{matrix} \overline{SNR}_{post}(n) - 1 & si \quad \overline{SNR}_{post}(n) - 1 \geq 0 \\ 0 & otro \quad caso \end{matrix} \tag{15}$$

Donde:
$\overline{SNR}_{post}(n) = \alpha\overline{SNR}_{post}(n-1) + (1-\alpha)\frac{SNR_{post}(n)}{\beta}$
$0 < \alpha < 1$
$\beta \geq 1$ factor de sobreestimación del ruido
La ecuación 14 produce valores nulos en las bandas de frecuencia donde el ruido estimado domina al ruido $\overline{SNR}_{post}(n) - 1 < 0$); por consecuente la función de ganancia en el MMSE produce valores nulos como estimado del habla limpia en esas bandas de frecuencias. Las componentes correspondientes a estos valores nulos se clasifican como las perdidas.

Interacción con máscaras inexactas. Máscaras suaves

En la práctica se hace imposible determinar con verdadera exactitud cuáles son las componentes U de un espectro dado. Los errores en las máscaras espectrográficas causan que se degrade el proceso de reconocimiento utilizando el método MF. La implementación de máscaras suaves es una solución que ha arrojado buenos resultados en trabajos recientes [15]. Como la mayoría de estas técnicas, primero se utilizaron en reconocimiento de palabra [29].

Para estimar máscaras suaves, a cada componente de la señal corrupta se le calcula la probabilidad de que sea R o U. Esto se puede hacer representando el estimado de SNR de la señal corrupta con una función sigmoidal.

$$Y \approx \frac{1}{1 + e^{-\alpha(SNR(t,s)-\beta)}} \tag{16}$$

Donde:

α pendiente de la función S

β intercepto de la función S con el eje y

El rango válido de variación de α es $[0, \infty)$. Para altos valores de α la función S tiende a convertirse en un escalón y por tanto se pierde el concepto de máscara suave. En este caso se está considerando de forma implícita que la estimación de error del ruido tiene una varianza (σ) pequeña. Por otro lado, a medida que el valor de α tiende a 0 va creciendo el desconocimiento del ruido y por tanto no se podrá crear una máscara confiable. Los valores apropiados para los parámetros β y α se obtienen de forma empírica después de una serie de ajustes. Aunque en [14] se reportan valores de alfa y beta para 3 y 0 respectivamente. La SNR se estima a partir de las diferencias de nivel estimadas entre la voz y el ruido, usando los procedimientos que se describen en epígrafes previos.

3.6.2. Compensación de las componentes no confiables

Una vez obtenida la cuantía de corrupción a través de la estimación de la máscara espectrográfica, es necesario decidir que hacer con las componentes U para pasar al proceso de reconocimiento. Una opción es reconstruir las componentes afectadas por el ruido y reconocer con este nuevo espectro estimado. La otra opción es excluir estas componentes dañadas y reconocer sólo con la información que no está cuantiosamente deteriorada por el ruido. A continuación se profundiza en métodos para ambas opciones previamente utilizados en reconocimiento de locutores.

Marginalización

La marginalización se comenzó a utilizar en reconocimiento de voz[39].El principio detrás de esta técnica es una clasificación óptima basada en ignorar las componentes U y clasificar utilizando solamente las componentes *(R)* correspondientes a la zona confiable del espectrograma. La marginalización se puede utilizar sobre cualquier método de clasificación (GMM, HMM).

Este tipo de métodos usualmente utilizan los vectores espectrales como parámetros para el reconocimiento. Esto se debe a que utilizan las mismas componentes U que salen de la máscara y en la conformación de ésta generalmente se usan rasgos espectrales. Esta contiene información localizada sobre la confiabilidad de cada componentes espectral, lo que es incompatible con los parámetros ortogonalizados, por ejemplo los MFCC [40].

La marginalización también se conoce como método de integración. Existen dos tipos: la marginalización completa y la acotada. La técnica de marginalización asume que las componentes de los parámetros no están acotadas y ocupan un espacio ilimitado $(-\infty, \infty)$. Por otro lado la técnica de marginalización o integración acotada toma en cuenta un poco de la información sobre el espacio de rasgos perdidos en el proceso de reconocimiento [30]. En este caso los índices de integración se escogen teniendo en cuenta los siguientes criterios: si el ruido en el dominio espectral es aditivo, las componentes espectrográficas enmascaradas por ruido tendrán una cota superior igual al estimado de la energía de

ruido; si las componentes de los parámetros están definidas en el dominio logarítmico, la cota inferior de energías es igual a $-\infty$.

Experimentos han demostrado que la marginalización acotada mejora los resultados de reconocimiento para bajos valores de $SNR < 10dB$, mientras que para altos SNR funciona peor que la completa [31]. Estos resultados demuestran que añadir información sobre las componentes perdidas puede mejorar el resultado de reconocimiento siempre y cuando el número de componentes U sea alto.

Por ejemplo si se usa un sistema basado en HMM las probabilidades de los estados se remplazan por el término:

$$\hat{P}(X|s) = P(X_r, -\infty \leq X_u \leq Y_u|s) \tag{17}$$

Se asume que la probabilidad está dada por la ecuación 21 y que todas las gaussianas en la mezcla de la probabilidad en esa ecuación tienen matrices de covarianza diagonal. Ahora, teniendo en cuenta que R es el índice de componentes R y U el planteamiento es el siguiente:

$$\hat{P}(X|s) = \sum_v c_{s,v} P(X_r, -\infty \leq X_u \leq Y_u; \mu_{s,v}, \theta_{s,v}) \tag{18}$$

Si Xu se asumen como desconocidas totalmente, dado que Yu no provee ninguna información que restrinja a Xu, entonces los límites de la integral estarán entre $-\infty, \infty$ y los términos de la integral pueden evaluarse en 1, es decir marginalizando las componentes espectrales a partir de la probabilidad de salida del estado (marginalización completa). Por otro lado, si fuera posible fijar cotas entre los valores posibles de Xu, la integral se pudiera fijar entre límites, por ejemplo $(-\infty, Yu)$ (marginalización acotada).

Imputación clase-condición

El método de imputación consiste en reemplazar las componentes corruptas por valores estimados, en muchos casos a partir de las muestras de entrenamiento[31]. Por ejemplo la técnica de imputación de media consiste en sustituir las componentes U por la media del modelo. Otra técnica que se utiliza también es estimar la media condicional a través de una regresión lineal, y con esta medida sustituir las componentes U. También se puede aplicar la técnica de modelado del background de voz integrado. Esta técnica se desarrolló en principio para el reconocimiento de voz en señales ruidosas [28] y luego se extendió al reconocimiento de locutores [41].

Al igual que en el método PMC la clave detrás de esta técnica consiste en crear modelos estadísticos del ruido de fondo y de la voz, y combinarlos para obtener un modelo ruidoso de la señal de voz. De esta forma se pueden reducir las consecuencias de la desigualdad entre las condiciones de entrenamiento y prueba. Esta técnica también se puede utilizar para hacer un estimado condicional de la señal limpia esperada, dadas las muestras de ruido y los modelos de ruido y señal limpia.

Trabajos previos [4] muestran que al realizar experimentos de verificación de locutores, utilizando estas técnicas de imputación de parámetros no se mejoran los resultados de EER con respecto a la marginalización.

Reconstrucción basada en agrupamiento

El método de Reconstrucción basada Agrupamiento (CBR, por sus siglas en inglés) hasta el momento solo ha sido utilizado en reconocimiento de palabra. Este método estima las componentes dañadas del vector espectral *(U)* a partir de las componentes R, utilizando un modelo estadístico que los relaciona. Este método considera que cada vector espectral del espectrograma es un proceso aleatorio independiente de los otros y con la misma distribución de probabilidad que éstos. La compensación de MF usando CBR ha arrojado buenos resultados en reconocimiento de voz, en reconocimiento de locutores no se han reportado resultados previos al experimento que se explica en el siguiente capítulo. CBR modela la distribución de vectores de la señal limpia como clusters de mezclas gaussianas. La media, covarianza y probabilidad a priori de cada cluster se estima a partir de un grupo de voces de entrenamiento usando estimación de máxima similitud, específicamente el algoritmo de maximización de la expectancia [42].

Si $Y = Yr + Yu$ es un vector espectral con componentes ruidosas y $X = Xr + Xu$ es su correspondiente vector reconstruido de señal limpia, el primer paso para compensar las componentes U es determinar la probabilidad del vector ruidoso que pertenece a cada cluster, dado por la siguiente ecuación:

$$P(k|Y) = \frac{w_k P(Y|k)}{\sum_{j-1}^{k} w_j P(Y|j)} \tag{19}$$

Donde:

w_k probabilidad a priori del cluster

Para calcular $P(Y|k)$ se debe tener en cuenta que Y está compuesto por una componente R y una U, y que para ruido aditivo $Xr = Yr$ y $Xu < Yu$. Sin embargo, se puede evaluar la distribución gaussiana en las componentes R e integrar fuera de las U.

$$P(Y|k) = P(X_r, X_u \leq Y_u|k) = \int_{-\infty}^{Y_u} P(X_r, X_u|k) dX_u \tag{20}$$

Si se asumen matrices de covarianza diagonales, la ecuación 18 se plantea de otra forma [42] A partir de cada cluster se puede obtener una estimación de las componentes U:

$$\hat{X}_u^k = argmax_{X_u}\{P(X_u|k, X_u \leq Y_u, X_r = Y_r)\} \tag{21}$$

Finalmente las componentes U se obtienen calculando la probabilidad posterior de los elementos de cada cluster al combinar las estimaciones de las componentes U.

$$\hat{X}_u = \sum_{k=1}^{K} P(k|Y)\hat{X}_u^k \tag{22}$$

La figura 5 muestra un ejemplo del algoritmo. Usando las componentes R de X_1 el procedimiento determina que los vectores de rasgos pertenecen al cluster 1 y sustituye las componentes U de Y_2 con una estimación limpia de X_2. Una vez recuperados todos los vectores de rasgos Mel-espectrales se pueda pasar a parametrizar de la forma tradicional

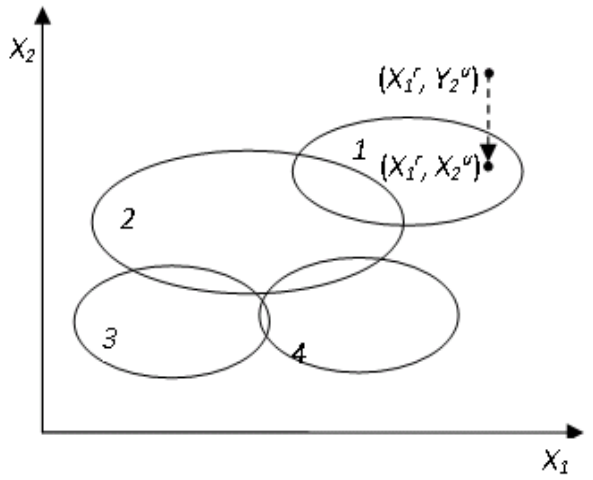

Fig. 5: Proceso CBR.

(MFCC, etc) para comenzar el proceso de reconocimiento. Aplicar este método para la compensación de MF permite utilizar cualquier tipo de parametrización y post procesamiento de ésta. No se requiere modificar el reconocedor para aplicar MF a una verificación o identificación de locutores, ofreciendo la cobertura de utilizar cualquier sistema que se tenga a disposición.

Reconstrucción basada en covarianza

Este método de compensación de componentes U sólo se ha utilizado en experimentos para reconocimiento de palabra. Esta forma de reconstrucción por covarianza considera las correlaciones estadísticas entre todas las componentes del espectrograma. Por este motivo, también se le conoce como reconstrucción basada en correlación.

Los vectores espectrales del espectrograma de la señal limpia se asumen como las salidas de un proceso estacionario en sentido amplio (WSS, por sus siglas en inglés) aleatorio y gaussiano. Todos los espectrogramas de las señales limpias se consideran observaciones independientes del mismo proceso estacionario. Esta asunción de WSS implica que las medias de los vectores espectrales y las covarianzas entre las componentes del espectrograma son independientes de su posición en el espectrograma, es decir las variables estadísticas del espectrograma no dependen del tiempo [14]. Similarmente la covarianza entre las componentes de dos vectores espectrales sólo depende de la distancia (τ) entre los vectores a lo largo del tiempo y no de su valor en tiempo.

La información a priori de los parámetros estadísticos (valor esperado de los vectores y covarianza entre sus componentes) se puede obtener a partir del entrenamiento realizado con señales limpias. Como el proceso se asume gaussiano la distribución conjunta de las componentes espectrales en una secuencia de vectores se asume gaussiana también y por tanto la distribución probabilística de un subgrupo de componentes lo será también, lo que implica que los valores estimados de la media y la covarianza caracterizan el proceso sin necesidad de estimar otros parámetros estadísticos (Papoulis 1991).

Dadas estas asunciones, la tarea de reconstrucción se va a centrar en obtener un estimado de las componentes U en el espectrograma original de la señal ruidosa usando MAP

y estimando las componentes ruidosas en los vectores individuales basados en las componentes no ruidosas que se encuentran en su vecindad. Para esto se construye un vecindario que reuna a los vectores de componentes espectrales cercanos a la zona U. Esto se hace a partir de las componentes R del espectrograma que tengan una covarianza relativa mayor que un valor de umbral, covarianza normalizada de por lo menos $0,5$ y un vecindario con al menos una componente U. La figura 6 ilustra la construcción del vector de vecindad a partir de la representación de un pequeño espectrograma de 4 vectores espectrales con 4 componentes cada uno. Las componentes grises son U y se desea estimar las componentes U del segundo vector espectral, las componentes bordeadas en negro son las que se escogen para formar la vecindad.

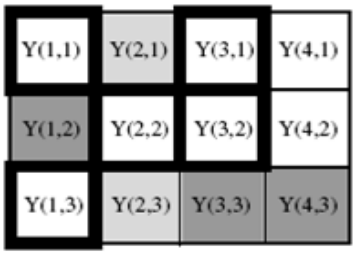

Fig. 6: Esquema para la construcción del vector de vecindad.

Los parámetros estadísticos para calcular el estimado de las componentes U dentro de cada vecindad se obtienen apoyándose en las muestras de entrenamiento y luego se aplica un estimado Máximo a Posteriori [43].

4. Experimentos y Resultados

En los siguientes epígrafes se presentan los resultados experimentales de la aplicación del método MF al reconocimiento de locutores. Se desarrolló un experimento de verificación de locutores en condiciones ruidosas aplicando técnicas del método MF, con el objetivo de evaluar la eficacia en el reconocimiento usando este método. Como detector de componentes no confiables se implementó una máscara espectrográfica basada en sustracción espectral [17]. Para compensar dichas componentes se utilizó una imputación de ceros reportada en [30]. Finalmente se realizó el reconocimiento de locutores.

4.1. Materiales

- Corpus: Sesión 1 microfónica de la base Ahumada [44] y muestra de ruido blanco de la base Noisex 92.
- Clasificador: Alize [45]

4.2. Diseño del experimento

Para determinar el nivel de corrupción de las señales se implementó una máscara espectrográfica a partir de substracción espectral empleando el Criterio de Energía Negativa [17] se utilizó el algoritmo clásico de SS [16] y el estimador de ruido de Martin [46]. La compensación de componentes U se realizó anulando sus valores, simulando una marginalización. Se simuló una verificación de locutores con señales ruidosas de aproximadamente 15 dB, aplicando el método MF. Como línea base para comparar se realizó una prueba de verificación con señales afectadas por ruido sin aplicar el método MF. Las señales limpias se ensuciaron digitalmente con ruido blanco.

La evaluación del comportamiento del método MF se hizo a través de una verificación de locutores utilizando Alize con las siguientes especificaciones. Como parámetros se tomaron 30 rasgos espectrales en escala Mel. Se utilizó un UBM de 512 gaussianas, obtenido de 50 señales de una sección microfónica de la base Ahumada. El entrenamiento se realizó con señales limpias de la base Ahumada que se adaptaron con MAP.

4.3. Resultados

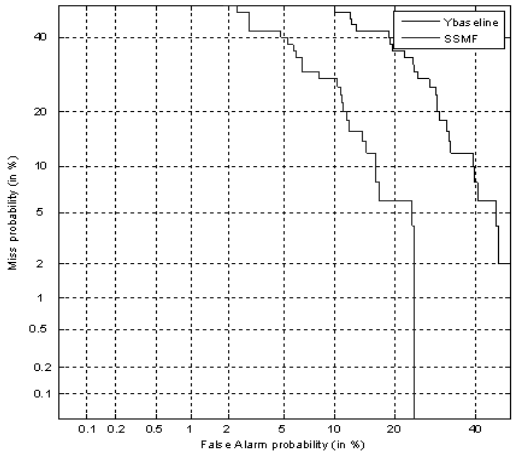

Fig. 7: Evaluación DET en Alize sin/con ruido blanco, 50 locutores de Ahumada

Tabla 1. Resumen de las curvas DET

	EER	DCF
Prueba con máscara	14.1632	6.5294
Prueba sin máscara (línea base)	27.7132	8.3698

Como se muestra en la curva, el comportamiento de la verificación de locutores mejoró notablemente al aplicar el método MF.

5. Conclusiones

Reconocer locutores en señales corruptas por ruido afecta la eficacia de los resultados. Cuando la voz tiene un nivel de SNR menor que 10dB ni siquiera asegurar el mismo nivel de ruido en entrenamiento y prueba asegura resultados eficaces de reconocimiento. La mayoría de los métodos para manejar el ruido fueron diseñdos para el procesamiento de voz, muchos se crearon para aplicaciones de reconocimiento de palabras. Actualmente los métodos que se utilizan en reconocimiento de locutores no fueron diseñados teniendo en cuenta las caracteristicas identificativas del locutor.

El problema del ruido que afecta a la señal de voz se ataca de diferentes formas según las características de la aplicación. La mayoría de los sistemas de reconocimiento de locutores, utilizan técnicas de adaptación y normalización, así como uno que otro método clásico de DSP para mitigar el ruido. Los métodos de compensación se utilizan en aplicaciones en las que el ruido provoca una afectación severa de la SNR de la señal de voz. Aplicando los métodos existentes los resultados de reconocimeitno de locutores no son eficaces para señales con bajos niveles de SNR afectadas por ruido desconocido, sobretodo si estos son ruidos no estacionario, mezclas de ruidos o ruido correlacionado con la voz. Las situaciones en que aparezcan estos tipos de ruidos son muy comunes en los escenarios donde se adquieren las voces en aplicaciones reales.

El caso de que el ruido provoque un escenario de datos perdidos es muy común en aplicaciones de reconocimiento de locutores. Por lo tanto, es muy útil desarrollar métodos que sean capaces de lidiar con estos casos. El método de rasgos perdidos para el procesamiento de voz se ha desarrollado mucho en el marco de las investigaciones para el reconocimiento de palabra, desarrollándose múltiples técnicas en las diferentes etapas del proceso. En el reconocimiento de locutores algunos grupos han utilizado el método de rasgos perdidos, apareciendo resultados publicables mayoritariamente a partir del 2005 [31,15,2,47,48].

Dado que es un campo que se está explorando todavia en las aplicaciones de reconocimiento de locutores quedan varios problemas no resueltos que señalan caminos de investigación. Por ejemplo:

- Quedan varias técnicas, ya sea para la identificación de componente U, como para la compensación de éstas, que solo se han probado en aplicaciones de reconocimiento de voz, que se pueden probar en reconocimiento de locutores y evaluar sus resultados.
- Según Raj y Stern [14]: "La creación de máscaras espectrográficas es la tarea más difícil de afrontar en el método MF". A pesar de que ya existen algunas soluciones en este campo, se pueden probar otras técnicas que sean capaces de determinar el grado de corrupción del espectro de forma más eficiente que las que están usando actualmente.
  - Ya se han desarrollado métodos de creación de máscaras a partir de clasificadores bayesianos [42]. Probar con otros clasificadores buscando el óptimo para la aplicación.
  - Según las características del ruido que afecte a la señal, buscar los métodos de creación de máscaras más adecuados para el caso.
  - Aplicación de otras parametrizaciones de más alto nivel que representen otras características de la señal y contribuyan a enriquecer el proceso de clasificación de rasgos de máscara.

Ya se han dado algunos pasos desarrollando experimentos en el marco de introducir el método de MF en el reconocimiento de locutores. Los resultados han sido alentadores, pero sobre todo un buen punto de partida para continuar experimentando en este campo. El trabajo futuro está dirigido a desarrollar alguna de estas líneas con el objetivo de crear una alternativa robusta para aplicaciones de reconocimiento de locutores.

Referencias bibliográficas

1. G. M. Davis, *Noise Reduction in Speech Applications*. New York: CRC Press LLC, 2002.
2. J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Transactions on Speech and Audio Processing*, vol. 15, pp. 1711–1723, 2007.
3. A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*. PhD thesis, 1990.
4. R. Teunen, B. Shahshahani, and L. P. Heck, "A model-based transformational approach to robust speaker recognition," 2000.
5. H. Hermansky, "Perceptual linear prediction (plp) analysis for speech," *Journal of the Acoustic Society of America*, 1990.
6. H. Hermansky, *RASTA processing of speech*. IEEE Trans. on Speech and Audio Processing, 1994.
7. P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, *Factor Analysis Simplified*. Canada: Centre de recherche informatique de Montreal (CRIM), 2005.
8. A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for svm speaker recognition," 2005.
9. F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, SylvainMeignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds, *A Tutorial on Text-Independent Speaker Verification*, vol. 4. Hindawi Publishing Corporation, 2004.
10. J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin Heidelberg: Springer-Verlag, 2008.
11. J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Berlin, Germany: Springer-Verlag, 2008.
12. M. Gales and S. Young, "Hmm recognition in noise using parallel model combination," 1993.
13. S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, *Jacobian Approach to Fast Acoustic Model Adaptation*. Proceedings of the ICASSP´97, 1997.
14. B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, 2005.
15. M. T. Padilla, T. F. Quatieri, and D. A. Reynolds, "Missing feature theory with soft spectral subtraction for speaker verification," 2006.
16. M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *IEEE*, pp. 208–211, 1979.
17. A. Drygajlo and M. El-Maliki, *Speaker Verification in Noisy Enviroments with Combined Spectral Subtraction and Missing Feature Theory*. Signal Processing Laboratory, Swiss Federal Institute of Technology at Lausanne, 1998.
18. M. Gales and S. Young, *Robust Continuos Speech Recognition using Parallel Model Combination*. Cambridge: Cambridge University, 1996.
19. L. P. Wong and M. Russell, *Text-Independent Speaker Verification Under Noisy Conditions Using Parallel Model Combination*. Edgbaston, Birmingham B15 2TT, United Kingdom: School of Electronic and Electrical Engineering The University of Birmingham, 2001.
20. Z. Tufekci and S. Gurbuz, *Noise Robust Speaker Verification Using Mel-Frequency Discrete Wavelet Coefficients and Parallel Model Compensation*. Urla-Izmir , Turkey: Department of Electrical and Electronics Engineering, Izmir Institute of Technology, ICASSP 05, 2005.
21. P. J. Moreno, *Speech Recognition in Noisy Environments*. PhD thesis, 1996.
22. J. Anguita and J. Hernando, *On the Use of Jacobian Adaptation in Real Speaker Verification Applications*. Barcelona, Spain: TALP Research Center, Universitat Politècnica de Catalunya, Department of Signal Theory and Communications, 2006.

23. C.Ñadeu, D. Macho, and J. Hernando, *Time and Frequency Filtering of Filter-Bank Energies for Robust HMM Speech Recognition.* Speech Communication, 2001.
24. C. Cerisara, L. Rigazio, R. Boman, and J.-C. Junqua, "a-jacobian environmental adaptation," *Speech Communication*, vol. 42, 2004.
25. J. Anguita, J. Hernando, and A. Abad, *Improved Jacobian Adaptation for Robust Speaker Verification.* IEICE Transactions on Information and Systems, 2005.
26. M. Cooke, P. Green, and M. Crawford, *Handling Missing Data in Speech Recognition.* Yokahama, Japan: Presented at the International Conference on Spoken Language Processing (ICSLP), 1994.
27. P. Green, M. Cooke, and M. Crawford, *Auditory Scene Analysis and Hidden Marcov Model recognition of Speech in Noise.* Sheffield, UK: Speech and Hearing research Group, Department of Computer Science, University of Sheffield, 1995.
28. M. Cooke, A. Morris, and P. Green, *Missing Data Techniques for Robust Speech Recognition.* Sheffield, UK: Computer Science, University of Sheffield, 1997.
29. J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," 2000.
30. A. Drygajlo and M. El-Maliki, "Speaker verification in missing features detection and handling for robust speaker verification," 1999.
31. M. El-Maliki and A. Drygajlo, *Integration and Imputation Methods for Unreliable Feature Compensation in GMM Based Speaker Verification.* Crete, Greece: A Speaker Odyssey, Tha Speaker Recognition Workshop, 2001.
32. J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in unknown noisy conditions," tech. rep., MIT Lincoln Laboratory, 2005.
33. S. Ahmad and V. Tresp, *Some solutions to the missing feature problem in vision.* Munchen, Germany: Siemens, 1993.
34. D. R. González, *Los rasgos dinámicos espectrales del habla y su relación con la prosodia en el reconocimiento de locutores.* 2008.
35. J. Campbell, *Speaker recognition: A tutorial.*, vol. 85. Procedings of the IEEE, 1997.
36. R. Cole, J. Mariani, H. Uszkoreit, G. Batista, A. Zaenen, A. Zampolli, and V. Zue, "Survey of the state of the art in human language technology," tech. rep., Cambridge University Press and Giardini, 1997.
37. T. Kinnunen, *Spectral Features for Automatic Text-Independent Speaker Recognition.* PhD thesis, 2003.
38. H. Hirsch and C. Ehrlicher, *Noise Estimation Techniques for Robust Speech Recognition.* ICASSP-95, 1995.
39. M. Cookeand, P. Greenand, L. Josifovski, and A. Vizinho, "Robust asr with unreliable data and minimal assumptions," 1999.
40. R. Togneri, M. Kühne, and S.Ñordholm, *Technologies and Applications*, ch. Time-Frequency Masking: Linking Blind Source Separation and Robust Speech Recognition. 2009.
41. R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, *Integrated Models of Signal and Background with Application to Speaker Identification in Noise*, vol. 2. IEEE Transactions on Speech and Audio Processing, 1994.
42. M. Seltzer, B. Raj, and R. M. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, 2004.
43. M. Seltzer, B. Raj, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, pp. 275–296, 2004.
44. J. Ortega, J. Gonzalez, and V. Marrero., "Ahumada: A large speech corpus in spanish for speaker characterization and identification.," *Speech Communication*, vol. 31, pp. 255–264, 2000.
45. J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, and J. Mason, "Alize/spkdet: a state-of-the-art open source software for speaker recognition," 2008.
46. R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, 2001.
47. D. Pullella, M. Kuhne, and R. Togneri, "Robust speaker identification using combined feature selection and missing data recognition," 2008.
48. M. Kühne, D. Pullella, R. Togneri, and S.Ñordholm, "Towards the use of full covariance models for missing data speaker recognition," 2008.

RT_012, febrero 2010
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: DCG Matilde Galindo Sánchez
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;
Ciudad de La Habana. Cuba. C.P. 12200
Impreso en Cuba