Técnicas de robustez frente al ruido para sistemas de reconocimiento de habla en teléfonos móviles y PDAs

Ascensión Gallardo Antolín[1], Javier Macías Guarasa[2], Rubén San Segundo Hernández[2] Javier Ferreiros López[2] y José Manuel Pardo Muñoz[2]
[1]Departamento de Teoría de la Señal y Comunicaciones. Escuela Politécnica Superior
Universidad Carlos III de Madrid
Avda. de la Universidad, 30, 28911 Leganés, Madrid
[2]Grupo de Tecnología del Habla. Departamento de Ingeniería Electrónica. ETSI Telecomunicación
Universidad Politécnica de Madrid
Ciudad Universitaria, s/n. 28040 Madrid
gallardo@tsc.uc3m.es, {macias, lapiz, jfl, pardo}@die.upm.es

Resumen: En los últimos años, las tecnologías de reconocimiento del habla han alcanzado un alto grado de madurez, de modo que son un componente básico en el diseño de las interfaces conversacionales de las que están dotadas muchas aplicaciones, incluyendo las desarrolladas sobre nuevos dispositivos como teléfonos móviles y agendas personales (*Personal Digital Assistant*, PDA). En este contexto, los sistemas de reconocimiento automático del habla (RAH) deben poseer dos características: en primer lugar, robustez frente a un rango amplio de condiciones acústicas adversas (por ejemplo, ruido de fondo) y en segundo lugar, unos requerimientos bajos en memoria y carga computacional. En este artículo se abordan ambas cuestiones mediante el estudio de técnicas basadas en transformaciones de parámetros con bajos requerimientos de memoria y proceso, cuyo objetivo es el de dotar de robustez frente al ruido a los reconocedores de voz.
Palabras clave: Reconocimiento automático del habla, robustez frente al ruido, técnicas de transformación de parámetros

Abstract: In recent years, speech recognition technology has reached a level of maturity to play an important role as a key component of conversational interfaces in many applications, including new services developed on portable devices with limited resources like mobile phones and Personal Digital Assistants (PDA). In this context, Automatic Speech Recognition (ASR) systems needs to meet two conditions: firstly, robustness against a great range of adverse acoustic conditions (for example, additive noise) and secondly, low memory and processing requirements. This paper deals with these questions by means of the study of low-resource techniques based on parameter transformations aimed at providing ASR systems with a great robustness.
Keywords: Automatic speech recognition, robustness, parameter transformation techniques

1. Introducción

A medida que las tecnologías del habla se han ido implicando cada vez más como parte integral de aplicaciones prácticas en escenarios reales (como acceso a bases de datos por línea telefónica, marcado automático de números telefónicos, control de instrumental en coches, máquinas de dictado, etc.), se ha hecho patente la necesidad de desarrollar sistemas automáticos de reconocimiento de habla (RAH) robustos, es decir que mantengan sus prestaciones dentro de un amplio margen de condiciones de entorno, incluso en el caso en que dichas condiciones varíen de forma rápida. Asimismo, en los últimos años, han aparecido nuevos escenarios de aplicación de los sistemas de RAH, que pueden funcionar sobre dispositivos nuevos, tales como teléfonos móviles o agendas personales (*Personal Digital Assistant*, PDA). En estos casos, no sólo es fundamental la robustez frente al ruido, si no que es necesario tener en cuenta que estos dispositivos presentan limitaciones en cuanto a memoria y capacidad de cómputo que determina que los reconocedores han de ser los más sencillos

posibles.

Este artículo se enmarca en este contexto. En concreto, hemos estudiado un conjunto de técnicas que, por una parte, mejoren las prestaciones de los reconocedores de voz en escenarios ruidosos, y que, por otra parte, demanden un bajo coste computacional y de memoria.

El contenido de este artículo es el siguiente. En la sección 2 se describen las técnicas de robustez basadas en transformaciones, que son las que utilizaremos en el resto de esta comunicación. En la sección 3 se trata del cálculo teórico de la función de transformación que relaciona los parámetros de habla limpia y habla ruidosa. La aplicación práctica de dicha función a un sistema de RAH se desarrolla en la sección 4. En la sección 5 se describen los experimentos y resultados obtenidos. Finalmente, en la sección 6 se mencionan las conclusiones más importantes de este trabajo.

2. *Técnicas de robustez frente al ruido basadas en transformaciones*

Las prestaciones de un sistema de RAH se degradan drásticamente cuando las condiciones acústicas en las que se grabaron los datos de voz utilizados para entrenarlo, difieren sustancialmente de las condiciones del entorno real en el que ha de funcionar. En los últimos años, se ha investigado y desarrollado un gran número de técnicas que tratan de paliar este problema (Junqua, 2000).

De entre ellas, nos centraremos en las técnicas de robustez basadas en transformaciones de parámetros. Estas técnicas consisten en la transformación de los parámetros de la señal de voz ruidosa en parámetros de voz sin contaminar, más similares acústicamente a los utilizados en la fase de entrenamiento del sistema. Para determinar la función de transformación (también denominada función de entorno), es necesaria la definición de un modelo matemático que relacione analíticamente la voz limpia, la contaminada y el ruido.

En este trabajo, hemos optado por técnicas pertenecientes a este tipo, porque una de sus principales características es que, en general, requieren de una menor carga computacional y de memoria en comparación con otros métodos. Esto las hace especialmente adecuadas para su incorporación a reconocedores de voz sobre teléfonos móviles o agendas personales, en los que existen fuertes restricciones computacionales. Además, presentan otras ventajas como la rápida adaptación a condiciones cambiantes del ruido de fondo y que no requieren de datos de voz en el entorno ruidoso para su funcionamiento.

Sin embargo, también presentan el inconveniente de que la expresión matemática de la función de entorno suele ser muy compleja, por lo que, para su simplificación, se suelen realizar una serie de aproximaciones que no siempre resultan válidas.

En este trabajo abordaremos este último punto. En particular, trataremos de analizar las limitaciones de las funciones de transformación utilizadas habitualmente y propondremos dos funciones de entorno nuevas que modelan de forma más eficaz la relación entre la voz sin contaminar y la ruidosa. Finalmente, compararemos su funcionamiento con una de de las técnicas más conocidas y utilizadas de robustez basada en transformaciones de parámetros: la substracción espectral generalizada.

3. *Efecto del ruido aditivo sobre la señal de voz*

En esta sección, derivaremos la forma genérica de la función de entorno que aplicaremos, con distintas aproximaciones a la transformación de parámetros para reconocimiento de habla con ruido. La expresión analítica del efecto del ruido de fondo sobre la señal de voz se puede expresar como una distorsión aditiva en el dominio del tiempo (Huang, Acero, y Hon, 2001), tal y como se indica en la siguiente expresión, en la que $y[n]$, $s[n]$ y $n[n]$ representan, respectivamente, las señales muestreadas de voz contaminada, de voz limpia (sin contaminar con ruido aditivo) y de ruido de fondo.

$$y[n] = s[n] + n[n] \tag{1}$$

La misma relación es también aditiva en el dominio de la frecuencia:

$$S_y(t,k) = S_s(t,k) + S_n(t,k),\ 0 \leq k < N \tag{2}$$

en la que $S_y(t,k)$, $S_s(t,k)$ y $S_n(t,k)$ representan la transformada de Fourier enventanada de N puntos en el punto de frecuencia *k-ésimo* de la trama *t-ésima* de la voz contaminada, limpia y el ruido, respectivamente. A

partir de este momento, y por claridad, prescindiremos del subíndice t, entendiendo que las expresiones siguientes se aplican a cada trama.

A partir de la expresión anterior, la densidad espectral de la señal de voz ruidosa puede calcularse con la siguiente ecuación:

$$\begin{aligned} |S_y(k)|^2 &= |S_s(k)+S_n(k)|^2 = \\ &= |S_s(k)|^2+|S_n(k)|^2+ \\ &+ S_s(k)S_n^*(k)+S_s^*(k)S_n(k) = \\ &= |S_s(k)|^2+|S_n(k)|^2+ \\ &+ 2|S_s(k)||S_n(k)|cos(\phi(k)) \end{aligned} \quad (3)$$

En la ecuación (3), $\phi(k)$ representa la diferencia de fase en el punto de frecuencia *k-ésimo* existente entre el espectro de la voz limpia y el ruido. Dicha expresión nos indica que la densidad espectral de la voz ruidosa es igual a la suma de las densidades espectrales de la señal limpia y del ruido más los términos $S_s(k)S_n^*(k)$ y $S_s^*(k)S_n(k)$ que están relacionados con las correlaciones cruzadas de la voz limpia y el ruido.

Para calcular las energías en banda, se aplica sobre la densidad espectral un banco de N_b filtros triangulares con pesos $H_m(k)$, cuya definición puede encontrarse en (Huang, Acero, y Hon, 2001). Por tanto, el cálculo de la energía en la banda *m-ésima* de la voz ruidosa, $Y(m)$, puede ser realizado a partir de la siguiente ecuación:

$$\begin{aligned} Y(m) &= \sum_{k=0}^{N-1} H_m(k)|S_y(k)|^2 = \\ &= \sum_{k=0}^{N-1} H_m(k)|S_s(k)|^2 + \sum_{k=0}^{N-1} H_m(k)|S_n(k)|^2 + \\ &+ 2\sum_{k=0}^{N-1} H_m(k)|S_s(k)||S_n(k)|cos(\phi(k)) \end{aligned} \quad (4)$$

Para conseguir una relación matemáticamente tratable es necesario suponer que el espectro de la señal es lo suficientemente suave dentro de la banda m, de modo que pueda ser aproximado a una constante en el interior de dicha banda igual al valor de espectro en el punto de su frecuencia central k_m. Teniendo en cuenta esta consideración, el último término de la ecuación (4) puede expresarse como:

$$\begin{aligned} &2\sum_{k=0}^{N-1} H_m(k)|S_s(k)||S_n(k)|cos(\phi(k)) \approx \\ &2|S_s(k_m)||S_n(k_m)|cos(\phi(k_m))\sum_{k=0}^{N-1} H_m(k) \end{aligned} \quad (5)$$

Aplicando la misma aproximación sobre la fórmula convencional del cálculo de las energías en banda, llegamos a la siguiente expresión:

$$\begin{aligned} S(m) &= \sum_{k=0}^{N-1} H_m(k)|S_s(k)|^2 \approx \\ &\approx |S_s(k_m)|^2 \sum_{k=0}^{N-1} H_m(k) \end{aligned} \quad (6)$$

en la que $S(m)$ representa la energía en la banda *m-ésima* de la señal $s[n]$. La misma expresión es válida para las energías en banda del ruido, $N(m)$. Esto quiere decir que la energía en cada banda es aproximadamente proporcional al módulo al cuadrado del espectro en la frecuencia central de la banda.

Combinando las ecuaciones (4), (5) y (6) obtenemos la expresión analítica que describe la distorsión que el ruido aditivo produce sobre la voz limpia en el dominio de las energías en banda:

$$\begin{aligned} Y(m) &= S(m)+N(m)+ \\ &+ 2\sqrt{S(m)}\sqrt{N(m)}cos(\phi(k_m)) \end{aligned} \quad (7)$$

Esta expresión indica que las energías en banda de la voz ruidosa, $Y(m)$, pueden calcularse sumando la contribución de las energías en banda de la voz limpia, $S(m)$, la del ruido, $N(m)$, y un término relacionado con la correlación cruzada entre el habla limpia y el ruido.

4. Transformación de parámetros

Como hemos mencionado previamente, la transformación se realiza en los parámetros de entrada al reconocedor. Por tanto, el problema consiste en estimar las energías en banda de la voz limpia a partir de las de la voz contaminada y una estimación del ruido de fondo de la siguiente forma:

$$\hat{S}(m) = T\{Y(m), N(m)\}, 0 \leq m < N_b - 1 \quad (8)$$

en la $T\{Y(m), N(m)\}$ es la función de entorno considerada, que en nuestro caso tendrá relación con la deducida en el apartado anterior.

4.1. Substracción espectral generalizada (SS_GEN)

La primera alternativa consiste en suponer que la señal de voz limpia y el ruido están incorrelladas y que ambas son de media cero. De esta forma, y partiendo de la ecuación (7) el término relacionado con la correlación cruzada puede ser eliminado. De este modo, la función de transformación es:

$$\hat{S}(m) = E\{S(m)\} = Y(m) - E\{N(m)\} \quad (9)$$

que corresponde con la ecuación básica de partida de las técnicas de substracción espectral clásicas (Boll, 1979).

La aplicación práctica más utilizada de la fórmula (9) es la denominada substracción espectral generalizada (SS_GEN) (Berouti, Schwartz, y J.Makhoul, 1979), (Kermovant, 1999) y que está definida por:

$$\overline{S}(m) = \begin{cases} T_{Y\overline{N}}(m), & \text{si } T_{Y\overline{N}}(m) \geq \beta\overline{N}(m) \\ \beta\overline{N}(m), & \text{en caso contrario} \end{cases} \quad (10)$$

donde $T_{Y\overline{N}}(m)$ es la función de transformación que se calcula con la siguiente expresión:

$$T_{Y\overline{N}}(m) = Y(m) - \alpha\overline{N}(m) \quad (11)$$

en la que $\overline{N}(m)$ es la estimación del ruido de fondo, que suele calcularse como el promedio de las energías en banda sobre las primeras tramas de silencio previas al inicio de la palabra. α es el factor de sobreestimación que intenta compensar la estimación deficiente del ruido y depende de la relación señal a ruido de la trama. β es el factor de umbral mínimo (*flooring*) cuyo objetivo es evitar que se obtengan estimaciones de $S(m)$ con valores muy pequeños o negativos, lo que daría lugar a la presencia de importantes distorsiones claramente audibles (ruido musical). Este último factor ha de ser determinado empíricamente dentro del intervalo $0 < \beta << 1$.

4.2. Substracción espectral con término de correlación fijo (SS_CORRF)

La suposición de que la voz limpia y el ruido están incorrelados y ambos son de media cero, no es adecuada al realizar el análisis espectral a corto plazo (como ocurre en el módulo de parametrización de un sistema de RAH) (Hung, Shen, y Lee, 2001). Además, resulta particularmente incorrecta para relaciones señal a ruido (SNR) bajas.

Con el siguiente ejemplo, podremos comprobar que la estimación de las energías de voz limpia es especialmente sensible a la incorporación del término de correlación cruzada cuando la relación señal a ruido es baja.

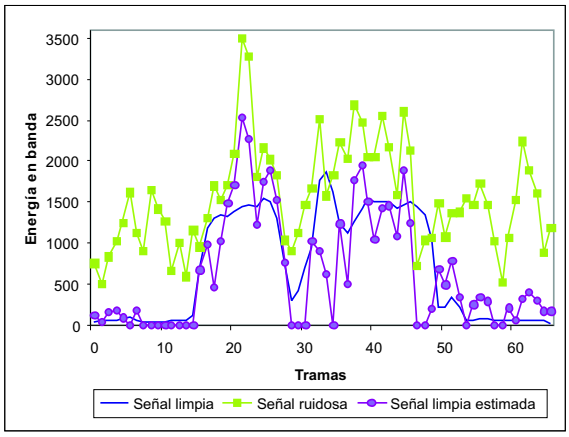

Figura 1: Evolución temporal de la energía en la banda 2 para la señal limpia, la señal ruidosa y la señal limpia estimada a partir de SS_GEN

La figura 1 representa la evolución temporal de la energía en la segunda banda para la voz limpia, la voz contaminada (artificialmente) con ruido de ventilador a una SNR baja y la estimación de la voz limpia realizada con la ecuación (10) con $\beta = 0,0$. Como podemos observar en dicha figura, existe una gran diferencia entre la señal estimada y la limpia, lo que indica que el término de correlación no es insignificante para estas condiciones.

A la vista de estos resultados, cabe esperar que la inclusión de una estima del término de correlación cruzada sea efectiva en condiciones de relación señal a ruido bajas, como veremos a continuación. Después de ciertas manipulaciones matemáticas sobre la ecuación (7) podemos deducir la siguiente

expresión que indica la forma de calcular una estimación de las energías en banda de la voz limpia a partir de las de la voz ruidosa y el ruido:

$$\hat{S}(m) = E\{Y(m)\} + E\{N(m)\} - \\ - 2E\{\sqrt{Y(m)}\sqrt{N(m)}cos(\theta(k_m))\} \quad (12)$$

en la $\hat{S}(m)$, $Y(m)$ y $N(m)$ son las energías en la banda *m-ésima* de la estimación de la voz limpia, la voz ruidosa y el ruido respectivamente y $\theta(k_m)$ es el desfase existente entre el espectro de la voz ruidosa y el ruido en el punto central de la banda, k_m.

En la ecuación anterior, puesto que la voz ruidosa es conocida, su media coincide con el valor verdadero. Por otra parte, la esperanza del ruido se suele calcular sobre las tramas de silencio inicial previas a la aparición de la señal de voz. La principal dificultad en la aplicación de la ecuación (12) es la estimación de la esperanza de su último término.

A continuación, proponemos un método sencillo para su cálculo basado en la siguiente desigualdad (Papoulis, 1980) (cap. 7, pág. 244). Dadas dos variables aleatorias X_1 y X_2, se verifica que:

$$[E\{X_1X_2\}]^2 \leq E\{X_1^2\}E\{X_2^2\} \Rightarrow \\ \Rightarrow E\{X_1X_2\} \leq \sqrt{E\{X_1^2\}}\sqrt{E\{X_2^2\}} \quad (13)$$

Además, hemos de hacer la suposición adicional de que la variable aleatoria $\sqrt{Y(m)}\sqrt{N(m)}$ es aproximadamente independiente de $cos(\theta(k_m))$. Aplicando esta aproximación junto con la desigualdad de la ecuación (13) en el último término de la ecuación (12) y dado que $E\{cos(\theta(k_m))\} \leq 1$ obtenemos que:

$$E\{\sqrt{Y(m)}\sqrt{N(m)}cos(\theta(k_m))\} = \\ E\{\sqrt{Y(m)}\sqrt{N(m)}\}E\{cos(\theta(k_m))\} \\ \leq \sqrt{E\{Y(m)\}\{N(m)\}} \quad (14)$$

De este modo, hemos calculado el límite superior del término de correlación cruzada o dicho de otro modo, hemos determinado que el término de correlación cruzada es proporcional a la raíz cuadrada del producto de las esperanzas de la energía ruidosa y la energía del ruido. Finalmente, llevando este resultado a la ecuación (12) obtenemos la nueva función de entorno dada por:

$$\hat{S}(m) = E\{S(m)\} = E\{Y(m)\} + E\{N(m)\} - \\ - 2\gamma\sqrt{E\{Y(m)\}E\{N(m)\}} \quad (15)$$

en la que γ es un factor constante (que puede ser tanto negativo como positivo) que engloba las correcciones a las aproximaciones antes formuladas y su valor se calcula de forma empírica.

Del mismo modo que en el caso de la substracción espectral generalizada, resulta conveniente introducir el factor de sobreestimación, α, para compensar las deficiencias en la estimación del ruido. Sin embargo, en este caso este valor es una constante, ya que esperamos que la estimación del término de correlación cruzada sea el que regule el proceso de substracción de ruido en función de la SNR local de la trama. También se ha incluído en la fórmula final el umbral mínimo, β, para evitar la aparición de ruido musical. De este modo, la transformación que aplicaremos sobre las energías en banda ruidosas es equivalente a la de la ecuación (10), con la diferencia de que la función de desajuste T es ahora:

$$T_{Y\overline{N}}(m) = \\ Y(m) + \alpha\overline{N}(m) - 2\gamma\sqrt{Y(m)\overline{N}(m)} \quad (16)$$

Hemos denominado a esta transformación como substracción espectral con término de correlación fijo (SS_CORRF), puesto que el parámetro γ es un factor constante para todas las tramas de vectores de parámetros. En este caso, las constantes a determinar empíricamente son tres: α, β y γ.

4.3. Substracción espectral con término de correlación variable (SS_CORRV)

El proceso de derivación de la transformación que hemos denominado substracción espectral con término de correlación variable (SS_CORRV) es idéntico al desarrollado en el caso anterior (SS_CORRF). La única diferencia consiste en que el factor γ no se considera

un valor constante para todos los vectores de parámetros y que ha de ser ajustado manualmente, sino que es variable trama a trama y depende del grado de similitud entre los parámetros de voz contaminada y el ruido.

De hecho, γ está relacionado con el coeficiente de correlación entre las energías en banda de la señal ruidosa y las del ruido (L.Hu, Bhatnagar, y Loizou, 2001). Por ello, para cada trama, se calcula una estimación de dicho coeficiente de correlación utilizando la siguiente expresión:

$$\gamma = \frac{\frac{1}{N_b}\sum_{m=0}^{N_b-1}\sqrt{Y(m)\overline{N}(m) - \mu_Y\mu_{\overline{N}}}}{\sigma_Y\sigma_{\overline{N}}} \quad (17)$$

en la que μ_Y, σ_Y, $\mu_{\overline{N}}$ y $\sigma_{\overline{N}}$ son la media y varianza de las energías en cada trama para la voz contaminada y el ruido, respectivamente.

La función de desajuste para SS_CORRV es la misma que para SS_CORRF (ecuación 16) con la diferencia de que γ es ahora variable. En este caso, existen únicamente dos parámetros a ajustar empíricamente: α (que también es constante) y β.

5. Entorno experimental y resultados

5.1. Bases de datos y tareas

Para la experimentación, hemos utilizado un subconjunto de la base de datos SpeechDat compuesto por palabras aisladas(Moreno, 1997). SpeechDat es una base de datos de voz grabada sobre línea telefónica por 1000 locutores distintos a una frecuencia de muestreo de 8 KHz. El subconjunto de entrenamiento consta de 5080 ficheros de voz y el subconjunto de test de 2203 ficheros. Los locutores del subconjunto de entrenamiento no están incluidos en el test. El diccionario de la tarea está compuesto por 1043 palabras (comandos, dígitos, apellidos, nombres de ciudades, etc.).

Esta base de datos ha sido contaminada artificialmente con ruidos pertenecientes a la base de datos estándar NOISEX (Varga et al., 1992). En concreto, hemos utilizado el ruido *volvo* grabado en el interior de la cabina de un coche circulando a 140 Km/h. y el ruido *leopard* grabado en el interior de la cabina de un automóvil militar circulando a 40 Km/h.

5.2. Sistema de referencia

El sistema de reconocimiento que hemos utilizado está basado en modelos ocultos de Markov continuos. Cada modelo, que consta de tres estados y 3 mezclas por estado, representa una unidad acústica similar a un alófono. El alfabeto consta de 45 unidades más dos adicionales correspondientes a los modelos de silencio inicial y final que ayudan a reducir los errores producidos por los fallos en el detector de principio y fin. Los modelos acústicos han sido entrenados utilizando el subconjunto de entrenamiento original, es decir, sin contaminar con ruido. Hemos utilizado un número limitado de mezclas por estado, puesto que nuestro objetivo es aplicar las técnicas de robustez frente al ruido en un sistema con grandes limitaciones en carga computacional y memoria (teléfono móvil con reconocedor local o PDA).

Con respecto a la parametrización, hemos utilizado 10 parámetros mel-cepstrales convencionales (MFCC), calculados cada 12,5 ms con ventanas de análisis de Hamming de 25 ms sin preénfasis y un banco formado por 17 filtros triangulares en escala mel, junto con la log-energía total en la trama y sus correspondientes derivadas.

Las marcas de inicio y final de palabra de los ficheros de voz fueron extraídas de forma automática utilizando un detector de principio y fin basado en umbrales de energía. El hecho de utilizar un detector automático nos aproxima a un escenario real en el que, justamente, los fallos de clasificación voz-no voz son una fuente importante de errores para el reconocedor de habla.

La tasa de reconocimiento de referencia del sistema, es decir, cuando la fase de entrenamiento y test se realiza con ficheros de voz sin contaminar es de 82,94 % con una banda de confianza de [81,37 % - 84,51 %].

5.3. Resultados

En esta subsección mostraremos los resultados obtenidos con las tres técnicas descritas en la sección 4.

En la figura 2 se representan las tasas de reconocimiento obtenidas para voz contaminada con el ruido *volvo* a las siguientes relaciones señal a ruido: 0 dB, 3 dB y 12 dB. La figura 3 representa los resultados para el caso de ruido *leopard* con relaciones señal a ruido de 6 dB, 9 dB y 18 dB. En todas las gráficas anteriores están representados los resultados

obtenidos para los siguientes casos:

- 'SC' (*sin compensación*): Corresponde con el caso en el que no se aplica transformación de parámetros.
- 'SS_GEN': Resultados obtenidos con substracción espectral generalizada (ver sección 4.1).
- 'SS_CORRF': Resultados obtenidos con substracción espectral con término de correlación fija (ver sección 4.2). Como existe una relación obvia entre α (que ajusta el nivel de ruido) y γ (que ajuste la correlación entre la señal ruidosa y el ruido), relacionamos ambos parámetros mediante: $\gamma = \alpha^{0,5}$. Después de una experimentación preliminar, se observó que para un intervalo de valores de α comprendido entre 0,2 y 0,9 no existía una variación significativa en la tasa de reconocimiento del sistema. En las gráficas adjuntas, los resultados corresponden con un valor intermedio de α ($\alpha = 0{,}5$ y por tanto $\gamma = 0{,}707$).
- 'SS_CORRV': Resultados obtenidos con substracción espectral con término de correlación variable (ver sección 4.3). El valor α fue determinado de forma empírica. Las tasas obtenidas resultaron ser estables para un intervalo de valores de α entre 0,001 y 0,6. En los resultados que se presentan se utilizó $\alpha = 0{,}2$.

Además, en dichas gráficas están representadas mediante líneas verticales las bandas de confianza para cada uno de los resultados (el solapamiento entre bandas indica que las diferencias entre los resultados correspondientes no son estadísticamente significativas).

Para todos las SNRs, el valor de β fue empíricamente fijado a 0,1. Además, la estimación del ruido se realizó promediando las energías en banda correspondientes a las 20 tramas previas al inicio de la palabra.

Con la transformación de parámetros SS_GEN se obtienen mejoras para relaciones señal a ruido bajas aunque éstas son estadísticamente significativas únicamente para el ruido *leopard* con SNR de 6 y 9 dB. Esto significa que SS_GEN mejora las prestaciones del sistema únicamente en los casos en los que hay un fuerte desajuste entre las condiciones acústicas de entrenamiento y test.

Pero para SNRs altas (ruido *volvo* a partir de 12 dB y ruido *leopard* a partir de 18 dB) y debido a las distorsiones que introduce en la señal, SS_GEN no sólo no produce mejoras sino que degrada el funcionamiento del sistema respecto al caso en el que no se aplica ninguna técnica de robustez. Esto hace que su utilización sea cuestionable cuando no se conoce de antemano el grado de contaminación de la señal de voz entrante o cuando presenta una SNR elevada.

Para SNRs bajas, SS_CORRF y SS_CORRV siempre mejoran significativamente la tasa con respecto al caso de no realizar ninguna transformación sobre los parámetros. Para SNRs altas (*volvo* con 12 dB y *leopard* con 18 dB) las tasas de reconocimiento son similares con respecto a no usar transformaciones. Lo interesante es que aunque no mejoran tampoco degradan el funcionamiento del sistema (como ocurre con SS_GEN para el ruido *volvo* y 12 dB). Esto proporciona una ventaja adicional de SS_CORRF y SS_CORRV que radica justamente en que son técnicas estables para distintas condiciones de SNR, favoreciendo, por tanto, su utilización en entornos más realistas en los que el nivel de ruido sea variable.

Además, en general, la inclusión del término de correlación cruzada (SS_CORRF y SS_CORRV) incrementa la tasa de reconocimiento del sistema para todas las relaciones señal a ruido consideradas respecto al caso de SS_GEN. Sin embargo, la transformación SS_CORRV es la única que consigue diferencias significativas respecto al caso de SS_GEN en casi todas las condiciones de SNR (excepto en el caso de ruido *leopard* y 18 dB).

Por último, respecto a la carga computacional introducida por SS_CORRF y SS_CORRV es un solamente un poco más elevada que con respecto a SS_GEN.

6. Conclusiones

En este artículo hemos estudiado diversas técnicas encaminadas a dotar de un mayor grado de robustez frente al ruido a los sistemas de reconocimiento automático del habla que funcionen en dispositivos con restricciones de memoria y capacidad de cómputo como teléfonos móviles o agendas personales. Las técnicas que más se ajustan a este tipo de escenarios son las basadas en transformaciones de parámetros puesto que

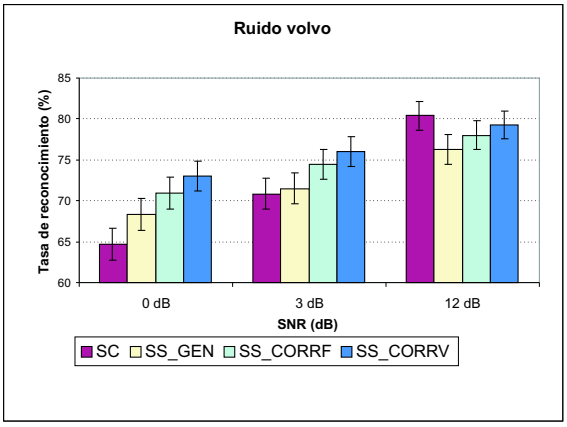

Figura 2: Resultados obtenidos con ruido *volvo* para los siguientes casos: 'SC' (sin compensación), 'SS_GEN', 'SS_CORRF' y 'SS_CORRV'

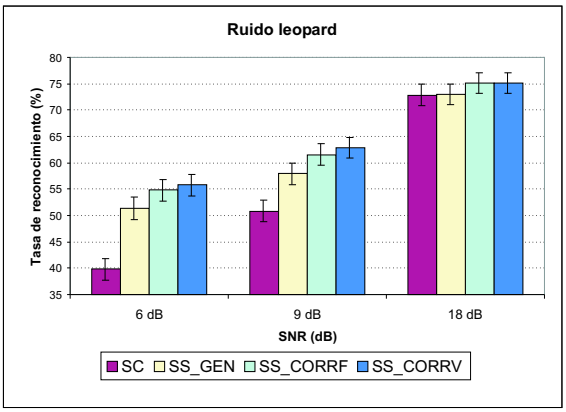

Figura 3: Resultados obtenidos con ruido *leopard* para los siguientes casos: 'SC' (sin compensación), 'SS_GEN', 'SS_CORRF' y 'SS_CORRV'

son las que menor carga computacional y de memoria demandan.

En este contexto, hemos propuesto dos técnicas nuevas (SS_CORRF y SS_CORRV) cuya novedad reside en la utilización de funciones de entorno más realistas obtenidas a partir del estudio de la influencia del ruido aditivo sobre la señal de voz. Hemos comprobado experimentalmente que ofrecen mejores resultados (especialmente SS_CORRV) que el método clásico de substracción espectral generalizada (SS_GEN) para dos tipos distintos de ruido y en diversas condiciones señal a ruido. Además, no degradan las prestaciones del sistema de reconocimiento en ausencia de ruido, efecto que se produce con el método convencional. Este resultado sugiere que son técnicas más estables con respecto a variaciones en el grado de contaminación de la señal, hecho frecuente en aplicaciones reales.

Bibliografía

Berouti, M., R. Schwartz, y J.Makhoul. 1979. Enhancement of Speech Corrupted by Acoustic Noise. En *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing: ICASSP 1979*, páginas 208–211.

Boll, S. F. 1979. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 27:113–120.

Huang, X., A. Acero, y H.-W. Hon. 2001. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development.* Prentice-Hall, Upper Saddle River, New Jersery 07458.

Hung, J.-W., J.-L. Shen, y L.-S. Lee. 2001. New Approaches for Domain Transformation and Parameter Combination for Improved Accuracy in Parallel Model Combination (PMC) Techniques. *IEEE Trans. on Speech and Audio Processing*, vol. 9(nº 8):843–855.

Junqua, J.-C. 2000. *Robust Speech Recognition in Embedded Systems and PC Applications.* Kluwer Academic Publishers.

Kermovant, C. 1999. *A Comparison of Noise Reduction Techniques for Robust Speech Recognition.* IDIAP Research Report 99-10. Disponible en http://www.idiap.ch.

L.Hu, M. Bhatnagar, y P. C. Loizou. 2001. A Cross-Correlation Technique for Enhancing Speech Corrupted with Correlated Noise. En *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing: ICASSP 2001.*

Moreno, A. 1997. *Documentación de SpeechDat.* Universitat Politécnica Catalunya.

Papoulis, A. 1980. *Probabilidad, Variables Aleatorias y Procesos Estocásticos.* Universitaria de Barcelona, Barcelona.

Varga, A. P., J. M. Steenneken, M. Tomlinson, y D. Jones. 1992. *The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition.* Technical Report. DRA Speech Res. Unit. Malvern, Worcestershire, U. K.