Nuevos parámetros acústicos para la clasificación de voz cercana, voz lejana y voz procedente de varios locutores

Oscar Varela Serrano [1], Rubén San-Segundo Hernández [2], Luis Alfonso Hernández Gómez[3]
[1] Telefónica I+D. Madrid, Spain
[2] Grupo de Tecnología del Habla, UPM
[3] Grupo de Aplicaciones del Procesado de Señal, UPM
Poster

ABSTRACT — This paper describes new acoustic features for improving VAD (Voice Activity Detection) when dealing with speech mixed with far-field and multi-speaker speech. Background voices are one of the major causes for the degradation of speech recognition performance in spoken dialog systems (specially over mobile phones). Also, in any audio indexing application, it can be necessary to separate the voice of a target speaker from others background speakers. This paper studies three new features to discriminate between near-field, far-field and background multi-speakers speech: 1) the percentage of frame-by-frame change for the best HMM mixture in a HMMs-based VAD; 2) the Mahalanobis distance between MFCCs from consecutive speech frames, and 3) the maximum auto-correlation value for each speech frame. Experimental results on the Av16.3 speech database for the best feature, obtain classification errors below 19% for near-field vs. far-field speech, and 3.5% for one-speaker vs. multi-speaker.

I. INTRODUCCIÓN

Este trabajo se centra en el problema que tienen muchas aplicaciones basadas en reconocimiento automático del habla cuando otros locutores, estáticos o en movimiento, distintos del locutor principal hablan a cierta distancia de él. Este problema de las voces de fondo (habla lejana) es especialmente importante en aplicaciones telefónicas basadas en reconocimiento de voz, sobre todo en aplicaciones de telefonía móvil: el locutor principal puede encontrarse en un entorno abierto o en una sala de reuniones, en ambos casos con la posible existencia de voces de fondo. El detector de actividad juega un papel muy importante en este sentido ya que es un sistema capaz de discriminar entre la ausencia (ruido, silencio) o presencia de voz para que un reconocedor automático de habla use esta información de forma adecuada. En la actualidad, los modelos de voz de los detectores de actividad tradicionales no pueden evitar reconocer estas voces de fondo como parte del diálogo hombre-máquina, situación que da lugar a un error de reconocimiento que hace que el sistema de diálogo falle.

En algunos trabajos previos, se han usado similares parámetros acústicos para técnicas de dereverberación. En [1] por ejemplo, los autores usan la idea de reverberación para reconstruir la voz degradada, debido a los rebotes del sonido producidos en una habitación, con la medida de dos micrófonos: se realizan operaciones cepstrales cuando el espectro de las observaciones no desaparece. Otra técnica de dereverberación usa el “pitch” como primera característica [2]. Este método estima el “pitch” y la estructura armónica de la señal de voz y obtiene un operador de dereverberación. Más tarde, el mencionado operador, basado en una operación de filtrado inverso, amplifica la señal. Por otro lado [3] propone un nuevo método de dereverberación con enventanado en la función de auto-correlación de ciertas tramas inteligentemente elegidas. Bees en [4] muestra una técnica que reduce la reverberación en salas y consiste en una deconvolución cepstral compleja y en el comportamiento que tiene la respuesta de un impulso en una sala. Se usan filtros cuadráticos inversos para recuperar la voz resultando una reducción importarte de la reverberación. Yegnanarayana propone en [5] un método para obtener el retardo entre dos señales de voz recogidas por dos micrófonos estimado usando información espectral a corto plazo (amplitud, fase o ambas a la vez) ya que la reverberación y el ruido degradan la señal de la voz y las características espectrales se ven afectadas. Finalmente, Cournapeau nos presenta en [6] un VAD (Voice Activity Detector) que funciona en tiempo real. Este VAD se basa en estadísticas de orden superior, que con la ayuda de la auto-correlación del LPC residual, puede discriminar entre habla lejana y habla cercana. Actualmente existe interés en sistemas que incorporan un VAD en tiempo real. En este sentido se propone el uso de nuevas técnicas para la utilización de estos detectores de actividad, por ejemplo, Ramírez en [7] propone un VAD para reconocimiento de voz embebida en ruido basado en la medida de divergencia de Kullback-Leibler.

En este trabajo se presenta el análisis de distintas características o parámetros acústicos para clasificar habla de campo cercano, habla de campo lejano y habla simultánea de distintos locutores: el porcentaje de cambios, entre tramas consecutivas, para la mejor gaussina en un VAD basado en HMMs, la distancia de Mahalanobis entre los MFCCs de tramas de voz consecutivas, y el máximo de auto-correlación obtenido del cálculo del “pitch” en cada trama.

II. BASE DE DATOS

La base de datos usada en este trabajo es la Av16.3. Está formada por datos, tanto de audio como de video, grabados en una sala como la que se puede ver en la Fig.1 y en la Fig. 2. Los ficheros de audio se han grabado con 16 micrófonos perfectamente sincronizados y convenientemente calibrados. Para cada grabación, hay 16 ficheros WAV grabados con dos arrays circulares de 8 micrófonos cada uno (Fig. 1) y muestreados a 16 Khz y otros ficheros WAV grabados desde

micrófonos de la solapa del locutor también muestreados a 16 Khz. Es importante mencionar que en algunas de las mencionadas grabaciones existe voz de habla simultánea, esto es, distintos locutores hablan a la vez. Los ficheros de voz se nombran en función de las características de los locutores. Ver referencia [8] para más información. Para concluir este apartado es importante decir que todos los ficheros de audio de la base de datos se han inframuestreado a 8 Khz (para simular el canal de telefonía móvil) y se han agrupado aleatoriamente en 3 categorías: ficheros de entrenamiento (80%), de validación (10%) y de testeo (10%). El análisis de las nuevas características se ha realizado a partir de lo obtenido en el proceso de entrenamiento.

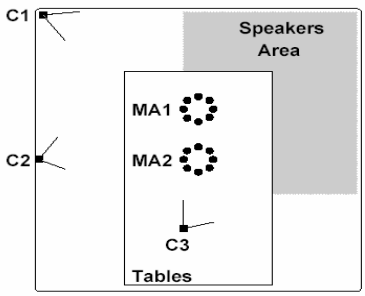

Fig.1. MA1 y MA2 son arrays circulares de 8 micrófonos. Observar la "Speakers Area". Esta figura se ha obtenido de [8].

Fig.2. Sala de grabación obtenida de [8].

III. Porcentaje de cambios para la mejor gaussiana en un detector basado en HMM

Este apartado presenta un estudio sobre el poder de discriminación de la máxima probabilidad obtenida a partir de un detector de actividad basado en HMMs, en tramas de voz consecutivas. Estas tramas poseen una duración de 24 milisegundos. El sistema de detección de actividad usa dos HMMs (voz y ruido) de un estado cada uno y con 200 gaussianas. El número tan elevado de gaussianas pretende introducir una mayor variabilidad de gaussianas a la hora de calcular la mejor. El detector de actividad usa un vector de MFCCs (generado mediante el análisis de 12 filtros Mel) formado por 8 coeficientes cepstrales, energía normalizada y la delta de energía. Los HMMs se entrenaron con el algoritmo de Baum-Welch.

Al comenzar los experimentos de cálculo de la gaussiana con la que se obtenía la mayor probabilidad por trama, se tuvieron en cuenta los pesos de las mencionadas gaussianas, con el problema de que se obtenía una variabilidad bastante pobre, esto es, el número de candidatos de la gaussiana ganadora era muy pequeño, en torno a 5. Por lo tanto, para aumentar esta variabilidad, se realizaron los mismos cálculos pero sin tener en cuenta estos pesos. Las figuras 3 y 4 muestran la distribución de este porcentaje de cambios considerando agrupaciones de N=100 y N=1000 tramas respectivamente. Como se puede observar en las figuras 3 y 4, el porcentaje de cambios de la mejor gaussiana es mayor si la voz procede de varios locutores que hablan al mismo tiempo. Esta característica puede discriminar muy bien entre la voz de un locutor principal y las voces procedentes de distintos locutores que hablan simultáneamente. Para este caso, el error de clasificación es menor del 26% y del 10% considerando 100 y 1000 tramas respectivamente (al incrementar el número de tramas consideradas de 100 hasta 1000 se tiene una mejor estimación del porcentaje de cambios). Por otro lado, el poder de discriminación entre la voz del locutor principal y la voz de un locutor lejano (con un porcentaje de cambios mayor) no es tan buena.

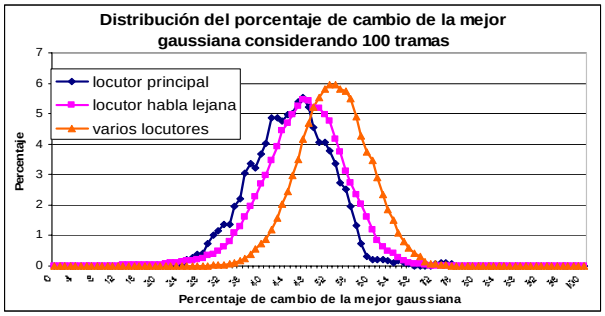

Fig. 3. Distribución del porcentaje de cambios considerando N=100 tramas.

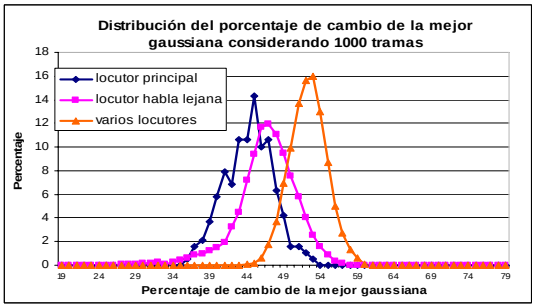

Fig. 4. Distribución del porcentaje de cambios considerando N=1000 tramas.

IV. Distancia de Mahalanobis entre coeficientes MFCC

Esta característica se basa en el cálculo de la distancia de Mahalanobis entre vectores de componentes MFCC, a partir de un banco de 12 filtros Mel con un previo filtrado de pre-énfasis para suavizar la señal, y calculados en tramas de voz consecutivas. Cada uno de estos vectores está formado por los primeros 8 MFCC, energía normalizada y la delta de energía. La distancia de Mahalanobis, se usa para determinar la similitud entre variables multidimensionales aleatorias. Las

distribuciones de la distancia de Mahalanobis entre tramas consecutivas para un locutor principal, un locutor de voz lejana y varios locutores se presentan en la Fig.5. Como se puede ver, la voz del locutor principal presenta la menor distancia, mientras que la mayor es para el caso de voz procedente de varios locutores. Una vez visto esto, se considera el análisis para grupos de N tramas (N=50 y N=500 tramas) calculando la distancia mínima en las N tramas. Las figuras 6 y 7 muestran las distribuciones de la mínima distancia para los tres casos: locutor principal, locutor de habla lejana y varios locutores.

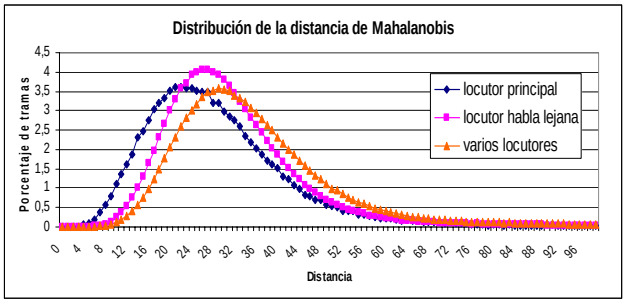

Fig. 5. Distribución de la distancia de Mahalanobis para un locutor principal, un locutor de habla lejana y varios locutores.

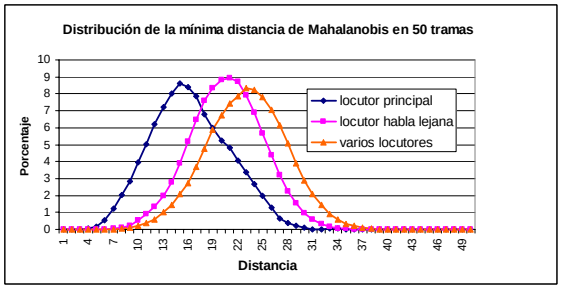

Fig. 6. Distribución de la mínima distancia de Mahalanobis para N=50 tramas.

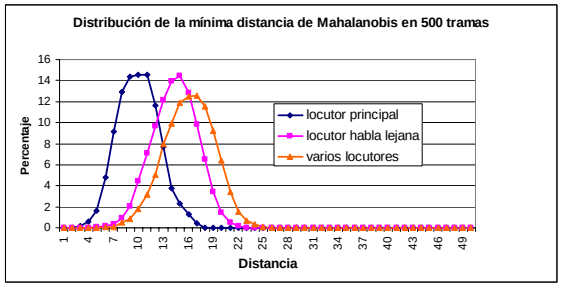

Fig. 7. Distribución de la mínima distancia de Mahalanobis para N=500 tramas.

Como se puede observar en las figuras 6 y 7, la mínima distancia en N tramas es mayor para el habla que procede de varios locutores. Esta característica puede también discriminar muy bien entre la voz de un locutor principal y las voces procedentes de varios locutores, incluso mejor que la característica anterior. Para este caso, el error es menor que el 24% y el 14% para agrupaciones de 50 y 500 tramas respectivamente. Además, cuando aumenta en número de agrupaciones de tramas consideradas desde 50 hasta 500 para el cálculo de esta mínima distancia, los resultados son mejores y el poder de discriminación aumenta. Por otro lado, el poder de discriminación entre la voz de un locutor principal y la voz de un locutor de lejano es mejor: los errores son menores que del 35% y el 27% para agrupaciones de 50 y 500 tramas respectivamente.

V. MÁXIMO DE AUTO-CORRELACIÓN CUANDO SE CALCULA EL PITCH

En este caso, el estudio se centra en el comportamiento de los valores de auto-correlación cuando se realiza el cálculo del pitch o frecuencia fundamental en cada trama. Se trata de tramas de 32 milisegundos, solapadas 16 y de 256 muestras (frecuencia de muestreo de 8 Khz.). Teniendo en cuenta sólo las tramas de voz, se calcula el máximo de auto-correlación en regiones de pitch. En la fig. 9 se presentan las distribuciones del máximo de auto-correlación:

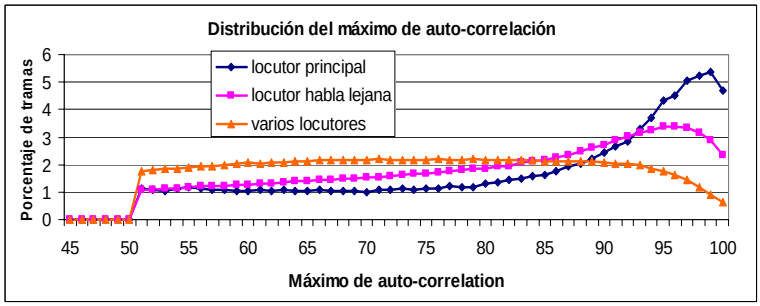

Fig. 8. Distribución del máximo de auto-correlación.

La fig. 8 presenta comportamientos muy diferentes en cuanto al máximo valor de auto-correlación obtenido para los tres casos diferentes, especialmente para valores de auto-correlación mayores de 0,9 (90% en la fig. 9). Hay muchas más tramas para el caso de la voz de un locutor principal y muy pocas para el caso de voces precedentes de varios locutores o un locutor de habla lejana. Tras considerar este efecto, se calcula para los tres casos el porcentaje de tramas (en N tramas) con un valor máximo de auto-correlación superior a 0,9. Las figuras 9 y 10 presentan las distribuciones del porcentaje de los valores del máximo de auto-correlación mayores del 0,9 para la voz de un locutor principal, la de un locutor de habla lejana y la de varios locutores para agrupaciones de 50 y 500 tramas. Como se puede apreciar, el porcentaje de agrupaciones de N tramas

es menor para el caso de habla de varios locutores. Este parámetros acústico es la que mejor discrimina entre la voz procedente del locutor principal y la que procede de varios locutores (el error es menor del 15% y del 3,5% para agrupaciones de 50 y 500 tramas respectivamente). Conforme se aumenta el número de tramas considerado, mejor es el resultado y el poder de discriminación aumenta. El poder de discriminación entre la voz de un locutor principal y la de un locutor de campo lejano es bastante mejor que el obtenido con las dos características anteriores. Para este caso se obtiene un error menor del 33,5% y 19% considerando agrupaciones de 50 y 500 tramas respectivamente.

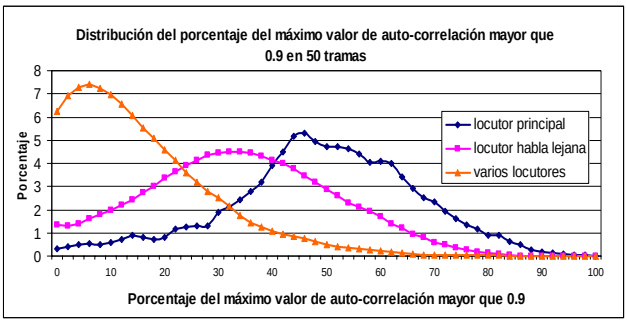

Fig. 9. Distribución con un valor máximo de auto-correlación mayor del 90% (N=50 tramas).

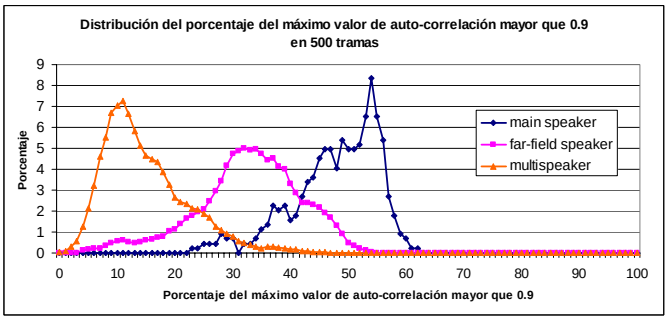

Fig. 10. Distribución con un valor máximo de auto-correlación mayor del 90% (N=500 tramas).

VI. Conclusión

Este trabajo presenta nuevos parámetros o características acústicas de la voz adecuadas para mejorar el comportamiento de un detector de actividad (VAD: Voice Activity Detection) cuando la voz del locutor principal se mezcla con voces de un locutor de campo lejano o varios locutores debido al entorno en que este se encuentra. Este estudio se ha realizado con la base de datos Av16.3 en la que sus ficheros de audio han sido muestreados a 8 Khz para simular el canal telefónico.

El primer parámetro presentado ha sido el porcentaje de cambios de la gaussiana que obtiene la mejor probabilidad obtenida a partir de un detector de voz basado en HMMs. Los resultados muestran que esta característica rechaza mejor el caso de habla procedente de varios locutores: el error entre la voz de un locutor principal y el habla procedente de distintos locutores es menor del 26% y del 10% considerando agrupaciones de 100 y 1000 tramas respectivamente.

La segunda característica es la distancia de Mahalanobis entre coeficientes MFCC de tramas de voz consecutivas. En este caso los resultados son mejores que para la anterior: el error entre la voz de un locutor principal y el habla procedente de distintos locutores es menor del 24%y del 14% considerando agrupaciones de 50 y 500 tramas respectivamente. Por otro lado, si se compara la voz de un locutor principal con una voz de campo lejano, el error de clasificación es menor del 35%, y del 27% para agrupaciones de 50 y 500 tramas.

Finalmente, el tercer parámetro es el máximo de auto-correlación en cada trama de voz. Esta característica es la que mejor resultados ha ofrecido. Puede discriminar casi de forma completa entre la voz de un locutor principal y el habla procedente de varios locutores que hablan simultáneamente: el error es menor del 15%, y del 3.5% considerando agrupaciones de 50, y 500 tramas. Si se compara la voz de un locutor principal con la de un locutor de campo lejano se obtienen errores menores del 33,5%, y del 19% para agrupaciones de 50 y 500 tramas.

Las tres características coinciden en que cuando se aumenta el número de tramas, el solape entre gaussianas es menor. Además, los resultados son bastante buenos en el caso de la segunda y la tercera característica para aplicaciones de tiempo real. Su poder de discriminación es bueno si se consideran agrupaciones de 50 tramas para sus cálculos.

Referencias

[1] Petropulu, A. P., and Subramaniam, S.,"Cepstrum based deconvolution for speech dereverberation", *IEEE Trans. Speech and Audio Proc.*,pp. 9-12,1994.

[2] Nakatani, T. and Miyoshi, M.,“Blind dereverberation of single channel speech signal based on harmonic structure”, pp. 92-95, ICASSP 2003.

[3] Ohta, K. and Yanagida, M., “Single channel blind dereverberation based on auto-correlation functions of frame-wise time sequences of frecuency components”, *Iwaenc 2006* – Paris – September 12-14, 2006.

[4] Bees, D., Kabal, P., and Blostein, M., “Application of complex cepstrum to acoustic dereverberation”, *Proc. Biennial Symp. Commun. (Kingston, ON)*, pp. 324-327, June 1990.

[5] Yegnanarayana, B., Mahadeva Prasana, S. R., Duraiswami, R. and Zontkin, D., “Processing of Reverberant Speech for Time-Delay Estimation”, *IEEE Trans. Speech and Audio Proc.*, pp. 1110-1118, vol. 13, nº 6, November 2005.

[6] Cournapeau, D. And Kawahara, T., “Evaluation of Real-Time Activity Detection based on High Order Statistics”, pp. 2945-2948, *Interspeech 2007*.

[7] Ramírez, J., Segura, J., Benítez, C. and Rubio, A., “A New Kullback-Leibler VAD for Speech Recognition in Noise”, *IEEE Signal Proc.*, vol. 11, nº 2, pp. 266-269, 2004.

[8] AV16.3: An Audio-Visual Corpus for Speaker Localization and Tracking.