TECNIACUSTICA
AVILA · 1999

MEJORA DE VOZ CON ARRAYS DE MICRÓFONOS MEDIANTE DESCOMPOSICIÓN EN COMPONENTES DE FASE MÍNIMA Y PASO TODO

REFERENCIA PACS: 43.72.-p

Joaquín González Rodríguez, José Luis Sánchez Bote y Javier Ortega García
Área de Tratamiento de Voz y Señales (ATVS)
Dpto. Ingeniería Audiovisual y Comunicaciones (DIAC)
E.U.I.T. Telecomunicación - Universidad Politécnica de Madrid
Ctra. Valencia, km. 7 - Campus Sur
28031 Madrid
Tel: 913 367 796
Fax: 913 367 784
email: jgonzalz@diac.upm.es

ABSTRACT

In this contribution, the speech signals obtained from a microphone array in a reverberant room are separately processed through its minimum phase and all-pass components, obtaining a great reduction in the reverberation present in the recorded speech signals. In order to also cancel out the effects of diffusse and coherent noise, a Wiener-based postprocessing scheme is proposed. Several experiments and results, performed with actual multichannel reverberant and noisy data, will be shown, confirming the noise and reverberation cancellation abilities of the system.

RESUMEN

En esta contribución, presentaremos un esquema basado en el procesado por separado de las componentes de fase mínima y paso todo de las señales de entrada al array, obteniendo de este modo una reducción significativa de la reverberación presente en la señal grabada, junto a un esquema de postfiltrado Wiener que nos permite cancelar los efectos tanto del ruido difuso, debido fundamentalmente a la reverberación, y del ruido coherente debido a las fuentes de ruido presentes en el recinto. Para demostrar la efectividad de este algoritmo, presentaremos resultados reales sobre señales multicanal grabadas en presencia de ruido y reverberación con los algoritmos descritos anteriormente.

INTRODUCCIÓN

La adquisición de señal de voz en presencia de ruido y reverberación es un campo de gran interés, donde el objetivo es lograr establecer comunicaciones "manos-libres" entre el hablante y el sistema receptor. Para aprovechar las características espaciales del campo acústico generado por el hablante, se han venido usando desde hace tiempo los arrays (o cadenas) de micrófonos [1][2][3] para realizar un filtrado espacial de la señal, de forma que apuntamos el haz directivo de recepción a la dirección estimada del locutor, mejorando por tanto en recepción la relación entre señal directa y señal reverberante.

Sin embargo, esta mejora sólo se hace apreciable para grandes cantidades de micrófonos, y además se encuentra limitada por la presencia de reverberación en el recinto. Con el objetivo de reducir el ruido remanente debido al campo difuso que se introduce en el sistema, se han propuesto diversas estructuras basadas en la

realización de un filtrado de Wiener posterior a la etapa de conformación de haz [4]. Nuevamente, esta estructura reduce fuertemente el ruido difuso presente a la entrada del array, pero no es capaz de cancelar fuentes de ruido que lleguen al array con alto grado de coherencia espacial (fuentes próximas y/ó muy sonoras), lo que puede resolverse mediante estimación de dichas componentes en ausencia de la señal de voz mediante un detector automático de actividad hablada, y su inclusión en el postfiltro modificado de Wiener [5], como ya presentaron los autores en [6].

La novedad que presentamos en este trabajo es la posibilidad de juntar en un único sistema la capacidad de reducción tanto de ruido coherente como difuso recién expuesta, con la extraordinaria capacidad de reducir la reverberación presente en la señal grabada en un recinto mediante el procesado por separado de las componentes de fase mínima y paso todo de las señales de entrada al array [7][8], las cuales se ven afectadas de forma muy diferente por la reverberación existente en el recinto.

COMPONENTES FASE-MÍNIMA Y PASO-TODO DE LA RESPUESTA IMPULSIVA DE UN RECINTO

Solemos representar H(w), transformada de Fourier (FFT) de la respuesta al impulso del recinto h(n) entre el transmisor y el receptor en un recinto reverberante, en la forma módulo-fase, es decir:

$$H(w) = |H(w)| \cdot \exp[jf(w)]$$

Nosotros, sin embargo, representaremos H(w) factorizada en sus componentes fase-mínima y paso-todo:

$$H(w) = H_{min}(w) \cdot H_{All}(w)$$

donde $H_{min}(w)$ y $H_{All}(w)$ son respectivamente, las componentes fase-mínima y paso-todo de *H(w).*

Se dice que una señal es de fase-mínima cuando su *Transformada-Z* (*TZ*) no contiene ningún polo o cero fuera del circulo unidad en el dominio Z. Las señales fase-mínima son especialmente interesantes ya que su señal inversa es causal y estable. Por desgracia, las respuestas al impulso de los recintos no son, por lo general, de fase-mínima ya que poseen ceros fuera del circulo unidad [10], quedando estos ceros representados en la componente paso-todo.

La componente fase-mínima, $H_{min}(w)$, puede expresarse de la forma siguiente:

$$H_{min}(w) = |H(w)| \cdot \exp[jf_{min}(w)]$$

Vemos cómo $H_{min}(w)$ depende sólo de la magnitud de H(w) y no de su fase. La información sobre la fase de H(w) está totalmente contenida en la componente paso-todo, $H_{All}(w)$, que puede obtenerse sencillamente dividiendo H(w) por $H_{min}(w)$:

$$H_{All}(w) = \exp\{j[f(w) - f_{min}(w)]\}$$

donde podemos comprobar cómo la componente paso-todo sólo tiene término de fase, es decir, tiene módulo unidad.

Si observamos en la *figura 1* el aspecto que tienen las componentes fase mínima y paso todo correspondientes a una función de transferencia entre dos puntos de un recinto, podemos extraer dos importantes conclusiones:
La componente fase-mínima se ve mucho menos afectada por la reverberación que la componente paso–todo e incluso que la respuesta impulsiva original. Así, toda la información correspondiente a la señal original se conserva en la componente fase-mínima de la señal recibida.
La componente paso-todo conserva intacta la información relativa al retardo de la señal directa y por tanto de la localización de la fuente emisora.

Figura 1.- Representación temporal de la respuesta impulsiva entre dos puntos de un recinto, y sus componentes fase-mínima y paso-todo.

IMPLEMENTACIÓN DEL ALGORITMO

El sistema completo de procesado en array que hemos utilizado está compuesto por las siguientes tres etapas situadas en cascada:
Conformación robusta de haz: para lograr el apuntamiento del haz de banda ancha del array, debemos ser capaces de estimar correctamente los retardos entre canales adyacentes en presencia de reverberación. En este trabajo, esta estimación robusta se consigue haciendo uso de la fase del espectro cruzado de potencia, como se describe en [9][11].
Descomposición y procesado en componentes fase-mínima y paso-todo: partiendo de la apreciación de que el cepstrum complejo de la señal de voz se encuentra concentrado en las bajas *quefrencies*, y que la información de los ecos se encuentra bastante separada sobre este eje, se realiza la descomposición de las señales de entrada al array en sus componentes de fase mínima y paso todo, utilizando un procesado diferente para cada una de dichas componentes [8]:

Procesado fase-mínima: la componente de fase mínima, directamente relacionada con el cepstrum real, se ve afectada en pequeña medida por la reverberación. De este modo, en cada una de las componentes de fase mínima del array, se encontrará la misma información de señal y una información característica de la función de transferencia entre la fuente y cada uno de los micrófonos, que cambia canal a canal. Así, el procesado propuesto consiste en realizar un promediado espacial de las mismas, de forma que la información de señal se ve realzada. Además, este procesado es equivalente a un promediado geométrico en el plano Z, por lo que se conserva la propiedad de secuencia de fase mínima (ceros y polos dentro de la circunferencia unidad). Este promediado es seguido de un proceso de *liftering* paso bajo que selecciona únicamente la parte de señal, concentrada en la parte de bajas *quefrencies*, eliminando componentes debidas a la reverberación.
Procesado paso-todo: en las componentes paso todo se conserva totalmente la información de fase de cada uno de los canales, y por tanto la información de posición de la fuente, de forma que lo que se realiza es un proceso de filtrado espacial mediante suma en frecuencia de cada una de las componentes. La señal resultante no es paso todo, por lo que nos quedamos únicamente con la componente paso todo de la señal resultante para combinarla con la componente fase mínima recién calculada.
Post-filtrado modificado de Wiener: con el objetivo de reducir el ruido residual que permanece en la señal de voz resultante del proceso anterior, se diseña un filtro óptimo a la salida del 'de-reverberador', basado en la teoría de filtrado óptimo de Wiener. Así, si suponemos que las componentes de ruido, debido a la separación entre micrófonos, se encuentran incorreladas entre sí, podemos utilizar un filtro como el que se muestra en [4]. Sin embargo, y como se demuestra en [6], existen componente correladas de ruido, debidas tanto a la separación espacial de los micrófonos [12], como a fuentes de ruido próximas ó muy sonoras, y que al ser consideradas componentes coherentes, atravesarán el filtro sin problema al suponer éste que toda componente coherente entre micrófonos es debida a señal deseada. Para evitar este problema, en [5] se hace uso de un detector de actividad hablada, para estimar en ausencia de voz las componentes coherentes de ruido y poder así incluirlas en el filtro según:

$$H_{LBF}(f,k) = \frac{g_{x_1x_2}(f,k) - \hat{g}_{n_1n_2}(f)}{g_{x_1x_2}(f,k)}$$

Con el objetivo de obtener una mejora aún mayor, los autores proponen en [6] el uso de un filtro basado en coherencia:

$$C_{x_ix_j}(f,k) = \frac{g_{x_ix_j}(f,k)}{\sqrt{g_{x_ix_j}(f,k) \cdot g_{x_ix_j}(f,k)}}$$

en las subbandas de baja coherencia espacial, de forma que usamos finalmente:

$$if \quad C_m > T \Rightarrow H_m(f,k) = H_{LBF}(f,k)$$
$$if \quad C_m < T \Rightarrow H_m(f,k) = C_m(f,k)^a$$

BASE DE DATOS DE EVALUACIÓN

Para este trabajo, hemos usado una base de datos multicanal real grabada por T.M. Sullivan y R.M. Stern en la Universidad de Carnegie Mellon (Pittsburg, PA., USA). Esta base de datos está formada por grabaciones simultáneas tanto de voz limpia, grabada mediante un micrófono de alta calidad montado sobre la cabeza del hablante, como de la voz recogida por un array de micrófonos a una cierta distancia del locutor, con lo que obtenemos una referencia exacta de los efectos introducidos por la propagación de la señal de voz por el recinto. Esta base de datos, muestreada a 16 kHz y con 16 bits por muestra, contiene diferentes subcorpora:
arrA: 10 locutores varones hablando a una distancia de 1 metro desde el centro de un array de 8 elementos

espaciados linealmente 7 centímetros, grabado en un laboratorio ruidoso debido fundamentalmente a múltiples ordenadores y equipos encendidos.
Los 6 subcorpora restantes fueron grabados por el mismo locutor con un array de 15 elementos espaciados de forma que se disponga de tres subarrays intercalados de 7 elementos cada uno, con espaciamientos lineales respectivos de N, 2N y 4N centímetros, como podemos ver en la figura siguiente:

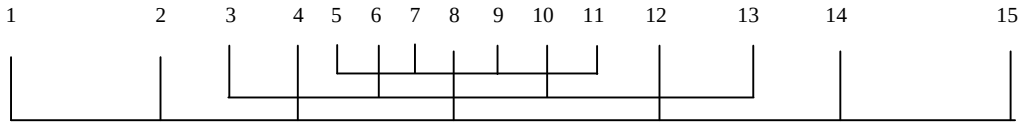

Figura 2.- Disposición del array de micrófonos con espaciamientos de N, 2N y 4N para cada subarray de 7 micrófonos.

arr3A: mismo laboratorio ruidoso anterior, con espaciamiento mínimo (N) de 3 cm., y con el locutor sentado a 1 metro (d=1 m.) del centro del array.
arr4A: laboratorio ruidoso, N=4 cm., d=1 m.
arrC1A: grabado en una sala de reuniones, más grande que el laboratorio ruidoso pero mucho más tranquilo, N=4 cm. y d=1m.
arrC3A: misma sala anterior, N=4 cm., pero ahora d=3 m.
arrCR1A: misma sala anterior, N=4 cm., d=1 m., pero incluyendo junto a la voz del locutor una señal de voz interferente procedente de una emisora de radio AM, llegando al array aproximadamente a 45º de su eje.
arrCR3A: igual que el anterior, pero con d=3 m.
Para este trabajo, hemos usado únicamente los subcorpus *arr4* (laboratorio ruidoso) y *arrC1A/arrC3A* (sala de reuniones, d=1m./3m.). Así, para el espaciado mínimo elegido de N=4 cm., cada uno de los subarrays de 4N, 2N y N cubrirán las subbandas de 0-1 kHz, 1-2 kHz y 2-8 kHz respectivamente.

EXPERIMENTOS Y RESULTADOS

Para la evaluación del algoritmo propuesto, se han realizado pruebas con todos los ficheros de los subcorpus *arr4A*, *arrC1A* y *arrC3A* de la base de datos CMU. Por limitaciones de espacio, no podemos exponer aquí tanto los resultados objetivos (distancia LAR, mejora en SNR) como subjetivos (tests perceptuales) que se encuentran publicados en [9], y mostramos únicamente un ejemplo de los espectrogramas de la señal de entrada a uno de los micrófonos del array, con gran cantidad de ruido y reverberación, y la señal a la salida tras procesar las señales de entrada al array con la estructura propuesta:

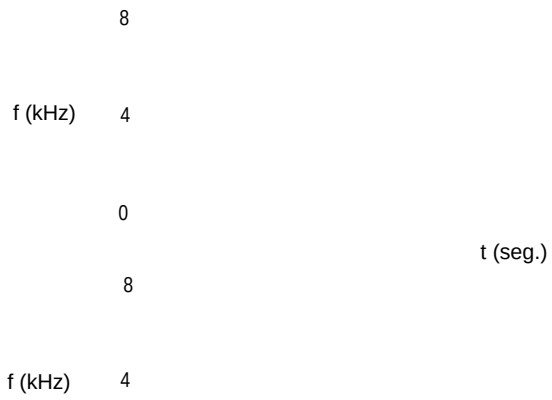

Figura 3.- Espectrogramas de entrada (voz con ruido y reverberación) y de salida.

CONCLUSIONES

En este trabajo hemos presentado una estructura de procesado en array capaz de reducir de forma apreciable tanto la reverberación, mediante el procesado separado de las componentes fase-mínima y paso-todo de las señales de entrada, como el ruido residual, ya sea éste de naturaleza difusa (debido a la reverberación) o coherente (debido a fuentes de ruido). Los resultados, tanto objetivos como subjetivos, obtenidos sobre ficheros multicanal reales grabados en recintos con dirferentes características acústicas, son excelentes, especialmente en la compleja tarea de reducción de la reverberación.

Debemos destacar que en estos momentos nos encontramos realizando la implementación en tiempo real de estos algoritmos en un ordenador personal tipo PC mediante una tarjeta de adquisición multicanal de alta velocidad de transferencia y una tarjeta de procesado de señal (DSP) de última generación.

AGRADECIMIENTOS

Los autores quieren expresar su especial agradecimiento a Dña. Mª Teresa Chamorro Calvo, por el excelente trabajo desarrollado durante la realización de su proyecto fín de carrera en el diseño y evaluación de sistemas de procesado en array de micrófonos.

REFERENCIAS BIBLIOGRÁFICAS

[1] J.L. Flanagan et al., “Computer Steered Microphone Arrays for Sound Transduction in Large Rooms”, J. Acoust. Soc. Am., vol. 78 (5), Noviembre 1985.

[2] J.L. Flanagan, A.C. Surendran and E.E. Jan, “Spatially Selective Sound Capture for Speech and Audio Processing”, Speech Communication, vol. 13, pp. 207-222, 1993.

[3] J. González Rodríguez and J. Ortega García, “Robust Speaker Recognition through Acoustic Array Processing and Spectral Normalization”, Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’97), vol. 2, pp. 1103-1106, Munich, Alemania, Abril 1997.

[4] R. Zelinski, “A Microphone Array with Adaptive Postfiltering for Noise Reduction in Reverberant Rooms”, Proc. IEEE Intl. Conf. Acoust. Speech and Signal Proc. (ICASSP’88), pp. 2578-2581, 1988.

[5] R. Le Bouquin et al., “Enhancement of Speech Degraded by Coherent and Incoherent Noise Using a Cross-Spectral Estimator”, IEEE Trans. on Speech and Audio Processing, vol. 5, No. 5, pp. 484-487, Septiembre 1997.

[6] J. González Rodríguez et al., “Coherence-based Subband Decompositon for Robust Speech and Speaker Recognition in Noisy and Reverberant Rooms”, Proc. of the International Conference on Spoken Language Processing (ICSLP’98), pp. 385-388, Sydney, 1998.

[7] M. Tohyama, *The Nature and Technology of Acoustic Space*, apdo. 6.3, Academic Press, 1995.

[8] Q.G. Liu, B. Champagne and P. Kabal, “A Microphone Array Processing Technique for Speech Enhancement in a Reverberant Space”, Speech Communication, vol. 18, pp. 317-334, 1996.

[9] Mª T. Chamorro Calvo, *Implementación y evaluación de sistemas de procesado en array para señal de voz*, Proyecto Fín de Carrera, SSR-ETSIT-UPM, Madrid, Junio 1999.

[10] S.T. Neely and J.B. Allen, “Invertibility of a room impulse response”, Journal of the Acoustical Society of America, vol. 66 (1), pp. 165-169, Julio 1979.

[11] M. Omologo and P. Svaizer, “Use of the Cross-Power Spectrum Phase in Acoustic Event Localization”, IEEE Trans. on Speech and Audio Processing, vol. 5, No. 3, pp. 288-292, Septiembre 1997.

[12] F. Jacobsen and T.G. Nielsen, “Spatial Correlation and Coherence in a Reverberant Sound Field”, Journal of Sound and Vibration, vol. 118 (1), pp. 175-180, 1987.