Anales Acad. Nac. de Cs. Ex., Fís. y Nat., tomo 68 (2016): 7-8.

SEPARACIÓN CIEGA DE FUENTES SONORAS: REVISIÓN HISTÓRICA Y DESARROLLOS RECIENTES

Leandro Ezequiel Di Persia

Instituto de Investigación en Señales, Sistemas e Inteligencia computacional sinc(*i*), Universidad Nacional del Litoral (UNL) – Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
Ciudad Universitaria – Ruta Nacional 168 – 4° piso, CP3000, Santa Fe, Argentina
E-mail: ldipersia@sinc.unl.edu.ar

Resumen

La separación ciega de fuentes sonoras apunta a obtener las fuentes sonoras que generaron un determinado campo sonoro, a partir de mediciones registradas del mismo con uno o más micrófonos. El escenario típico es el llamado *cocktail party*, en el cual múltiples hablantes están dialogando, y las personas son capaces de focalizarse en una conversación, ignorando las demás fuentes sonoras que los rodean. Esto que las personas podemos hacer en forma transparente, se ha presentado como un problema muy difícil de resolver cuando se trata de dispositivos electrónicos que deben captar el sonido ambiente, segregarlo y extraer información de las fuentes individuales. El problema es más complicado cuantas más fuentes sonoras haya, cuanto mayor sea el tiempo de reverberación del recinto, y cuantos menos micrófonos en relación al número de fuentes activas se disponga. En este artículo se realizará una revisión del tema, repasando en particular los métodos de separación ciega de fuentes en el dominio frecuencial, realizando un repaso histórico del desarrollo del área y de los métodos más recientes.

Palabras clave: separación ciega de fuentes sonoras; reverberación; dominio tiempo-frecuencia; análisis de componentes independientes.

Abstract

Blind sound source separation: historical revision and recent developments. The blind sound source separation task aims at obtaining the sound sources that produced a specific sound field, from measurements of the same obtained using one or more microphones. The typical setup is that of a cocktail party, in which multiple people are talking to each other and humans have the capability of focusing in a particular conversation, ignoring the other sound sources surrounding them. This task, that human beings can perform in a transparent way, is a very hard to solve problem when dealing with electronic devices that must capture the sound, separate the sources and extract information from individual sources. The problem grow in complexity as the number of sound sources is increased, as the reverberation time of the environment increases, and also when the number of available microphones is small compared to the number of active sound sources. In this article, a review on this subject will be presented, with particular emphasis on frequency domain blind source separation algorithms. A historical review of the development of the research on this area will be presented, together with the last advances and newer algorithms.

Keywords: blind sound source separation; reverberation; time-frequency domain; independent component analysis.

Introducción

El problema de separación ciega de fuentes sonoras puede formularse como la extracción individual de las fuentes sonoras que generaron un determinado campo sonoro en un determinado ambiente, a partir de mediciones realizadas sobre el efecto conjunto de todas las fuentes [1,2]. En un ambiente abierto, donde el sonido sólo pudiera llegar a los micrófonos por el camino directo, a él o los micrófonos utilizados para la adquisición del campo sonoro llegarían copias retardadas y atenuadas de las fuentes sonoras. Supongamos el escenario más completo, de múltiples fuentes y múltiples micrófonos. Sea $s_j(t)$ la j-ésima fuente sonora, α_{ij} la atenuación sufrida por la misma y τ_{ij} el retardo introducido por el tiempo de viaje desde la fuente j-ésima hasta el micrófono i-ésimo. Entonces la señal captada por el mismo $x_i(t)$ puede expresarse como

$$x_i(t) = \sum_{j=1}^{M} \alpha_{ij} s_j(t - \tau_{ij}), \quad (1)$$

donde hemos supuesto que había *M* fuentes activas y *N* micrófonos midiendo el campo sonoro.

Cuando este mismo escenario se repite en un ambiente cerrado como se ve en la Fig. 1, a cada micrófono llegará no sólo el sonido viajando por propagación directa, sino todos los ecos producidos por las reflexiones de todos los órdenes que se producen en todas las superficies sólidas del ambiente. Este proceso puede modelarse como un filtrado lineal, donde el ambiente actúa como un filtro que modifica la fuente sonora. Sea $h_{ij}(t)$ la respuesta al impulso medida desde la fuente sonora j-ésima hasta el micrófono i-ésimo.

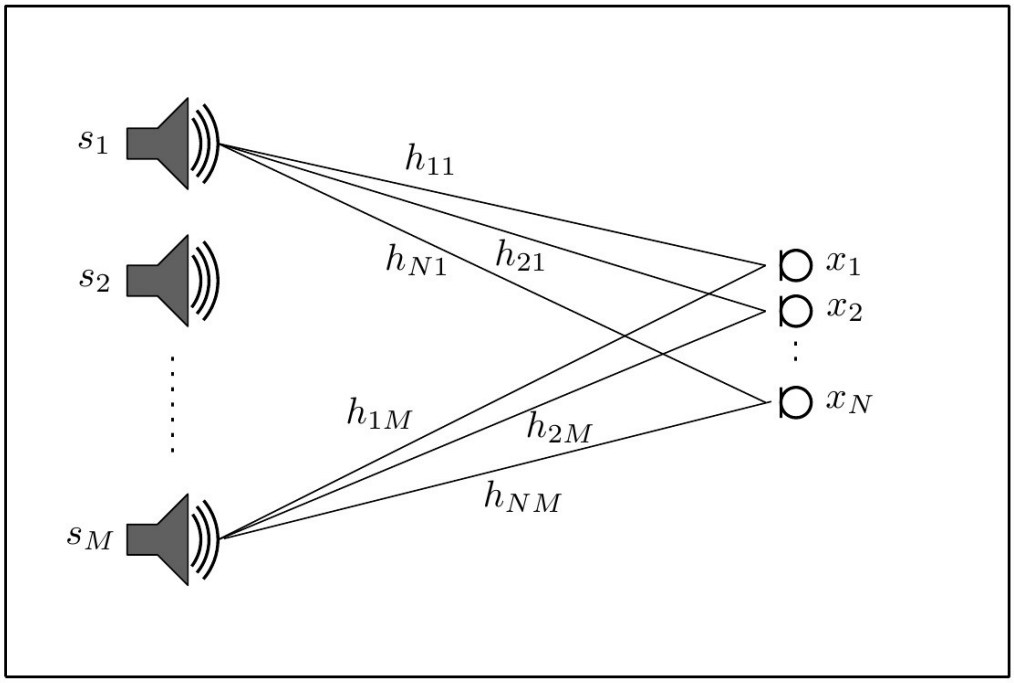


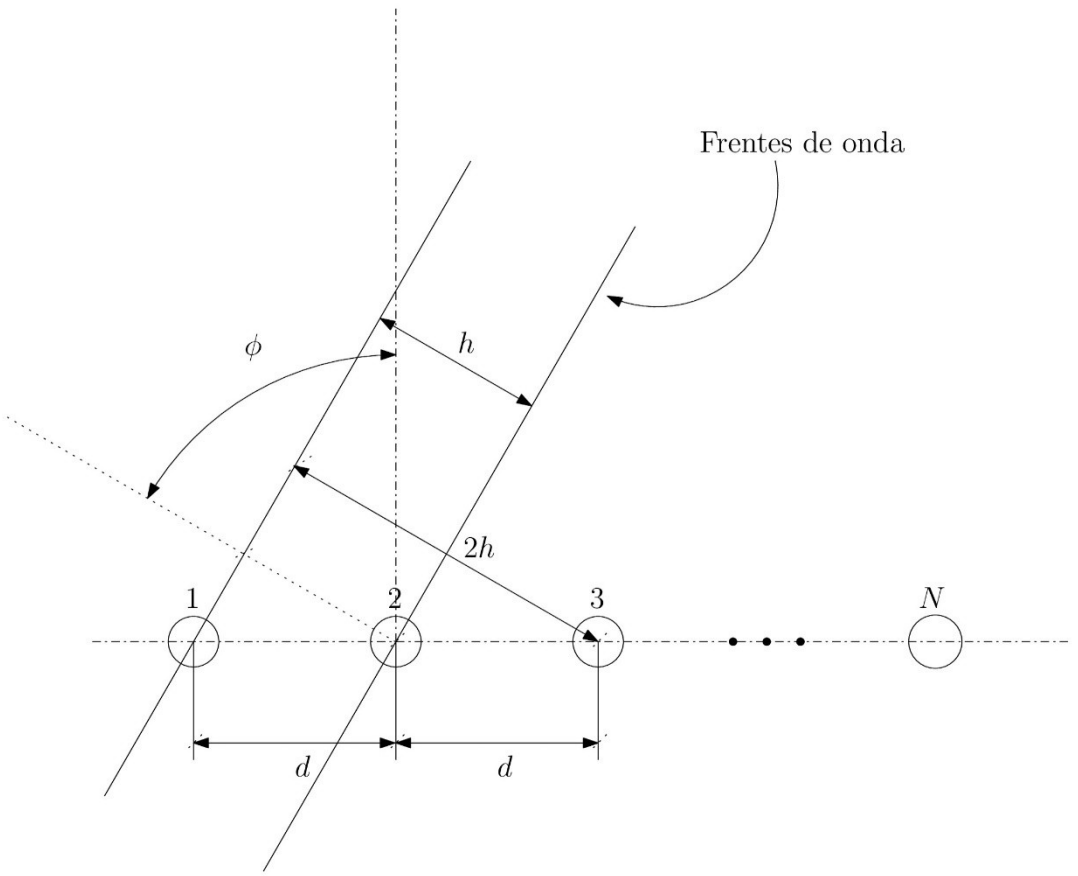
Fig. 1 - *Un caso de mezclas sonoras con M fuentes y N micrófonos en un ambiente cerrado*

En este contexto, la señal adquirida por el i-ésimo micrófono puede expresarse como

$$x_i(t) = \sum_{j=1}^{M} h_{ij} * s_j(t) , \qquad (2)$$

donde el símbolo * representa la operación de convolución. Esto es lo que se conoce como modelo de mezcla convolutiva. En este contexto, el objetivo de la separación ciega de fuentes puede plantearse como la obtención de M señales $y_j(t)$ tales que sean lo más parecidas posibles a las fuentes, $y_j(y) \approx s_j(t)$ [2].

A lo largo de los años se han propuesto diferentes alternativas para obtener aproximaciones a las fuentes sonoras. Los diferentes enfoques difieren en las condiciones de adquisición (por ejemplo, cantidad de micrófonos y disposición espacial de los mismos), en hipótesis sobre las fuentes sonoras (independencia estadística, decorrelación temporal), dominio en el que se realiza el procesamiento (temporal, frecuencial), y cantidad de micrófonos en relación a la cantidad de fuentes ($N = M$ o determinado, $N > M$ o sobredeterminado, $N < M$ o subdeterminado, el caso más difícil). En este trabajo nos centraremos en el caso determinado, y principalmente en los enfoques en el dominio tiempo-frecuencia.



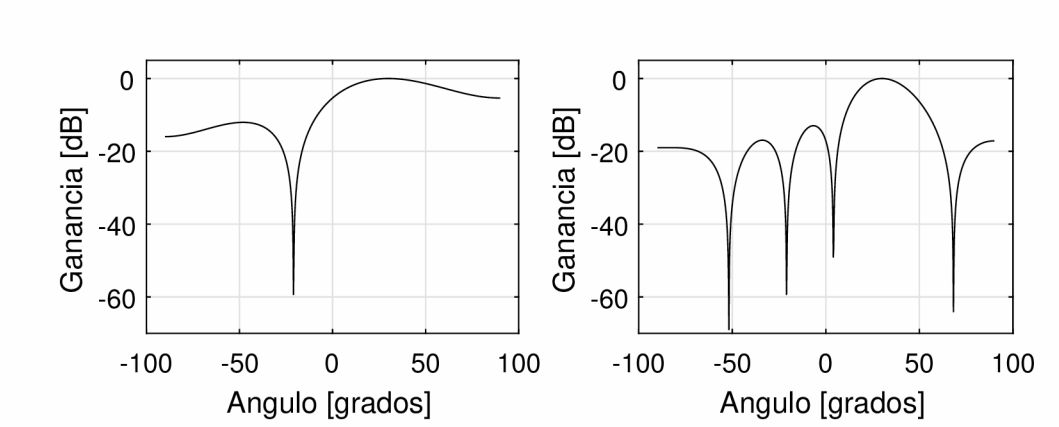


Fig. 2 – *Arreglo de micrófonos lineal uniforme (arriba) y patrones de directividad (abajo) de un beamformer para el caso de 5 micrófonos (izquierda) y 10 micrófonos (derecha)*

El enfoque tradicional para este problema estaba basado en lo que se conoce como *beamforming* [3]. En este tipo de procesamiento, se adquiere el campo sonoro con múltiples micrófonos omnidireccionales y las señales generadas por ellos son combinadas de alguna forma específica. Según el tipo de combinación, se logra transformar el arreglo de micrófonos omnidireccionales en un sensor direccional, que realza el sonido que arriba de una dirección específica y atenúa los que llegan desde otras direcciones. En la Fig. 2 puede verse un arreglo lineal de micrófonos y el patrón de directividad producido al "apuntar" el mismo hacia un ángulo específico. Existen muchas variantes de beamforming, pero la principal desventaja de este tipo de métodos es que la dirección de la o las fuentes de interés debe ser conocida *a priori*.

En algunas aplicaciones esto no es un problema. Por ejemplo, en el uso de computadoras personales para videoconferencia, lo usual es que el sonido de interés sea el habla de la persona sentada frente a la cámara. De tal forma, los diseñadores de notebooks suelen incorporar en la misma un arreglo de micrófonos ubicados apropiadamente, con los que producen un beamformer que apunta en la dirección hacia donde está apuntada la cámara, y de esta forma se supone que captará preponderantemente la voz del interlocutor que está siendo filmado. Esto es tecnología estándar disponible hoy en día.

Sin embargo, en otras aplicaciones de interés, como interfaces hombre-máquina específicas para control remoto de dispositivos, videoconferencia o teleconsulta con múltiples personas y en situaciones dinámicas,

sistemas de comunicación de manos libres, etc., es imposible conocer de antemano las posiciones que tendrá cada fuente que pueda ser de interés (y más aún, puede no saberse a priori cuales son las fuentes de interés). En estos escenarios, aparece el interés en algoritmos *ciegos*, es decir, aquellos que sean capaces de segregar o separar las fuentes de interés con la menor información posible del escenario sonoro implicado. Es aquí donde entran en juego los algoritmos de separación ciega de fuentes sonoras que se revisarán en el artículo [2].

En la siguiente sección se presentarán los métodos iniciales de separación de fuentes en el dominio temporal. Luego se presentarán los métodos de separación en el dominio tiempo-frecuencia, repasando desde sus inicios a los métodos más recientes, para el caso de mezclas completas. Luego se presentarán métodos más recientes basados en factorización de matrices nonegativas, aplicados a mezclas subcompletas y completas. Luego se discutirá cómo puede evaluarse la calidad de la separación obtenida mediante los métodos de separación. Finalmente se presentará un ejemplo de aplicación de algunos algoritmos seleccionados a un conjunto de señales obtenidas en ambientes reales, evaluando el desempeño en forma comparativa de los diferentes métodos. El artículo cerrará con conclusiones.

Primeros trabajos: separación en el dominio temporal

Los primeros métodos de separación de fuentes surgieron a partir de un modelo de mezclas más sencillo, llamado de mezclas instantáneas. En el mismo, se supone que cada fuente llega al sensor sin que haya retardos relativos, lo que es equivalente a suponer que no hay retardo en la trasmisión (de aquí el término instantáneo) lo cual equivale a que la velocidad de transmisión es muy rápida o las distancias implicadas muy cortas. En este modelo, cada fuente estará afectada por un coeficiente de atenuación α_{ij} que contemplará la pérdida de energía producida al viajar de la fuente j-ésima al sensor i-ésimo. El modelo generativo de la mezcla es

$$x_i(t) = \sum_{j=1}^{M} \alpha_{ij} s_j(t) , \qquad (3)$$

que agrupando las fuentes en el vector $\mathbf{s} = \left[s_1 s_2 \ldots s_M \right]^T$ y las señales de los sensores en el vector $\mathbf{x} = \left[x_1 x_2 \ldots x_N \right]^T$, puede escribirse como

$$\mathbf{x}(t) = A\mathbf{s}(t), \tag{4}$$

donde A es una matriz con elementos α_{ij} . El problema de separación se traduce entonces en encontrar una matriz de separación W tal que aplicada al vector $\mathbf{x}$ genere un vector de señales $\mathbf{y}$ que sea una buena estimación de las fuentes originales $\mathbf{s}$, es decir

$$\mathbf{y}(t) = W\mathbf{x}(t) \approx \mathbf{s}(t). \tag{5}$$

Debe notarse que, en las ecuaciones anteriores, tanto la matriz A como las señales fuentes en $\mathbf{s}$ son desconocidas, con lo cual estamos frente a un sistema de ecuaciones con menos ecuaciones que incógnitas. Para poder encontrar la matriz de separación W adecuada, se necesita recurrir a ecuaciones adicionales que permitan medir qué tan parecidas a $\mathbf{s}$ son las señales obtenidas $\mathbf{y}$. Pero como no conocemos $\mathbf{s}$, esto tampoco puede medirse, por lo que se recurre a hipótesis sobre sus propiedades. Por ejemplo, se puede suponer que las fuentes eran no correlacionadas, y por lo tanto buscar la matriz W que genere las señales lo menos correlacionadas entre si posibles. Una de las propiedades más populares para esto es la independencia estadística de las fuentes. Entonces se propone una función de costo que mide qué tan independientes estadísticamente son las estimaciones de las fuentes obtenidas, y se puede optimizar esta función buscando la W que maximice dicha independencia. Este método se denomina Análisis de Componentes Independientes (ICA, del inglés Independent Component Analysis [1]) y ha tenido numerosas aplicaciones en el procesamiento de imágenes [4], Electroencefalografía [5], Electrocardiografia [6], entre otras.

Un problema de este enfoque es que presenta dos indeterminaciones intrínsecas e imposibles de salvar. El orden en que aparecen las fuentes estimadas es arbitrario (es decir, aunque uno haya generado la mezcla con las fuentes específicamente en cierto orden, y pueda recuperarlas con exactitud, podrían estar en orden arbitrario), y con un escalado también arbitrario. En muchas aplicaciones esto podría no parecer de importancia, pero será crucial para los algoritmos que se desarrollarán más adelante.

El modelo de mezcla instantáneo, como ya se discutió, no es apropiado para señales de audio, ya que en un ambiente anecóico debería contemplar los retardos de transmisión relativos (ecuación (1)), y en un ambiente

reverberante debería utilizarse el modelo de mezclas convolutivas (ecuación (2)). Por esto, inicialmente a mediados de los 90 comenzaron a aparecer enfoques que generalizaban los métodos de ICA al caso de mezclas convolutivas, en el dominio temporal [7, 8]. El problema convolutivo también puede expresarse en forma compacta con un poco de abuso de notación de la siguiente forma

$$\mathbf{x}(t) = H * \mathbf{s}(t), \tag{6}$$

donde H ahora es una matriz en la que cada entrada es un filtro $h_{ij}(t)$, y la operación * es similar a un producto matriz-vector, pero reemplazando los productos por convoluciones. El problema de separación entonces se plantea como encontrar la matriz de filtros W tal que

$$\mathbf{y}(t) = W * \mathbf{x}(t) \approx \mathbf{s}(t), \tag{7}$$

y en forma análoga a los métodos de ICA explicados anteriormente, se puede optimizar una función que mida el grado de independencia obtenido por las fuentes estimadas para una matriz de filtros W, y buscar los filtros que maximicen dicha independencia.

El principal problema de estos enfoques es que las optimizaciones necesarias involucran convoluciones, y las ecuaciones de actualización implicadas resultan tanto más complejas cuanto más largas sean las respuestas al impulso h_{ij}. Esto hacía que el costo computacional fuera prohibitivo para filtros de apenas decenas de coeficientes. Ahora bien, una habitación promedio de una casa, por ejemplo un living, donde podría utilizarse un sistema de control remoto por voz de dispositivos, puede tener tiempos de reverberación del orden de 400 a 600 milisegundos, lo que muestreando las señales a 8000 Hz implica que las respuestas al impulso pueden tener entre 3200 y 4800 coeficientes, con lo cual los métodos de ICA en el dominio temporal resultaban inaplicables en la práctica. Esto dio lugar a los métodos que analizaremos en la siguiente sección.

BSS basado en ICA en el dominio frecuencial

El modelo de mezcla convolutivo de la ecuación (2) puede simplificarse si se aplica una transformada de Fourier de tiempo corto (STFT, del inglés short time Fourier Transform). Sea $S(\omega,\tau)$ la STFT del vector de

señales **s**(t), y suponiendo que las fuentes y los micrófonos están estáticos (o al menos no se mueven durante el tiempo de observación), podemos expresar la ecuación (2) en dominio frecuencial como

$$X(\omega,\tau) = H(\omega)S(\omega,\tau) \tag{8}$$

donde $H(\omega)$ es una matriz de mezcla que sólo depende de la frecuencia ω. Puede verse que, para una frecuencia fija, este modelo es simplemente el modelo de mezclas instantáneas de la ecuación (4). La única diferencia sería que las señales implicadas, y la matriz de mezcla, tienen valores complejos. Esto quiere decir que se puede aplicar cualquier algoritmo de ICA que permita trabajar con señales complejas (por ejemplo [9]), y de esta forma resolver el problema de separación de fuentes. En resumen, se reemplaza la solución de un problema de mezclas convolutivas en el dominio temporal, de muy alta complejidad y costo computacional, por la solución de un conjunto de problemas de ICA instantáneos (uno para cada frecuencia ω), de menor complejidad y costo computacional.

El gran problema de este enfoque ya fue comentado en la sección anterior: las indeterminaciones de amplitud y permutación. La Fig. 3 presenta un esquema de este problema. Supongamos una mezcla con dos fuentes y dos micrófonos. En una frecuencia ω se aplica un algoritmo de ICA complejo y se obtienen estimaciones de las dos fuentes.

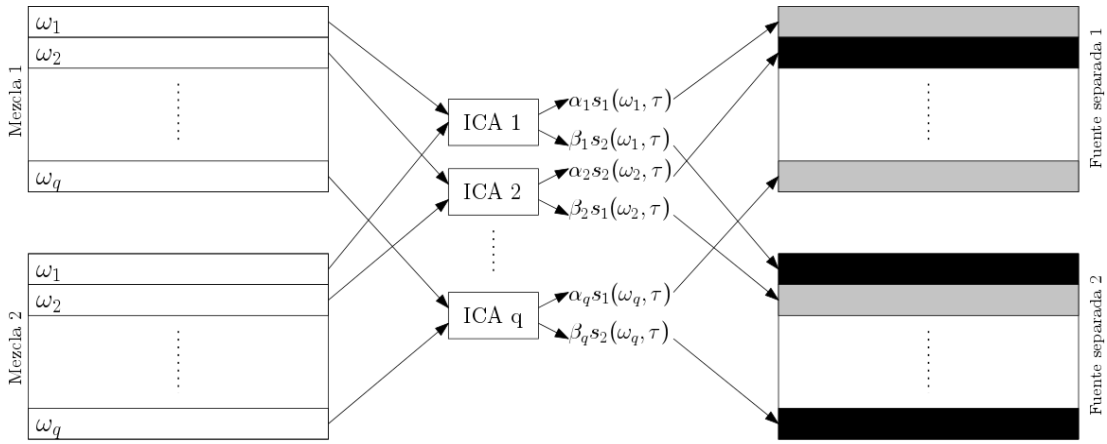


Fig.3 – *Ilustración del problema de permutaciones. Al aplicar ICA en una banda de frecuencias, las fuentes estimadas pueden tener una permutación arbitraria respecto de las fuentes extraídas para las demás frecuencias*

La primera estimación corresponde a la primera fuente y la segunda a la otra. Al pasar a la siguiente frecuencia y resolver el problema de ICA correspondiente, la primera estimación podría corresponder a la segunda fuente y la segunda estimación a la primera. Y así para todas las frecuencias, con permutaciones aleatorias e impredecibles. Más aún, cada estimación en cada frecuencia estaría escalada por un valor arbitrario, también desconocido, y distinto para cada frecuencia. Esto debería ser corregido antes de poder antitransformar para volver a tener las señales en el dominio temporal.

El primer algoritmo de valor práctico para esta tarea que pudo resolver razonablemente bien estas dificultades fue propuesto por Murata y colaboradores [10]. En su enfoque, para cada frecuencia se estima la envolvente de amplitud de las estimaciones a lo largo del tiempo. Los perfiles de variación de esta amplitud a lo largo del tiempo para dos frecuencias serán muy distintos para fuentes diferentes, y relativamente similares para la misma fuente. De esta manera, se va estimando una "envolvente de fuente" promediando las envolventes de las frecuencias ya clasificadas para cada fuente, y para una nueva frecuencia se calcula la correlación entre la envolvente de las estimaciones y dichas envolventes de fuente, asignando cada componente a la fuente con cuya envolvente tiene más correlación. También introducen un método de resolución de la indeterminación de amplitud. Sea $W(\omega)$ la matriz de separación estimada por ICA para una frecuencia. Las fuentes estimadas están dadas por $Y(\omega,\tau) = W(\omega)X(\omega,\tau)$. La matriz inversa W^{-1} es una estimación de la matriz de mezclas $H(\omega)$. Si todas las estimaciones menos una se anulan, y se aplica la matriz de mezclas estimada, lo que se obtiene es la proyección de esa fuente como fuera registrada por cada uno de los micrófonos, y la indeterminación de amplitud desaparece. Este método más adelante fue desarrollado en detalle, llamándose principio de mínima distorsión (MDP, del inglés Minimal Distortion Principle) [11]. Aun así, la evaluación del desempeño de este tipo de algoritmos mostró que la calidad obtenida con el mismo disminuía rápidamente al aumentar el tiempo de reverberación del recinto donde se obtuvieron las grabaciones.

En este tiempo también surgió un trabajo muy importante que explicó la principal limitación de este tipo de algoritmos en función de la reverberación. Araki y colaboradores [12] demostraron que el comportamiento de estos métodos en cada frecuencia es similar al de un conjunto de beamformers anuladores. Este es un tipo de beamformer en el cual, en lugar de combinarse

las señales para realzar el sonido de una dirección, se combinan para anular el sonido de una dirección. En el caso de mezclas de dos fuentes medidas con dos micrófonos, el algoritmo de ICA entonces estaría estimando en forma ciega las direcciones de las dos fuentes, y generando dos beamformers ciegos, donde cada uno está dedicado a anular el sonido proveniente de la dirección de una de las fuentes, dejando sólo el de la otra dirección. ¿Pero qué sucede en un ambiente con tiempos de reverberación alto? A los micrófonos llegan “rebotes” de las ondas sonoras que vienen de muchas direcciones, entonces para una de las fuentes la señal recibida por cada micrófono será una superposición de copias con diferentes retardos arribando de múltiples direcciones. Como el beamformer generado sólo elimina lo que llega de una dirección, quedarán ecos “residuales” provenientes de todos los demás ángulos de arribo, lo cual degradará la calidad de separación, dado que la señal obtenida no tendrá sólo la otra fuente, sino copias de la que se quería eliminar, atenuadas pero aún presentes.

A pesar de esta limitación, comenzó a desarrollarse un creciente interés por los métodos de separación ciega de fuentes en el dominio frecuencial, apareciendo en los años siguientes numerosos algoritmos que exploraban esta técnica, y que apuntaban a mejorar las debilidades de este método en tres aspectos: por un lado, en proponer mejores algoritmos de ICA complejo que generaran resultados de más calidad para la separación en cada frecuencia [13, 14]. En segundo lugar, algoritmos mejores para la resolución de las permutaciones, que permitieran un menor número de permutaciones residuales [15, 16]. En tercer lugar, el uso de post-procesamientos para incrementar la calidad de separación al tratar de eliminar los ecos residuales [17, 18]. Como ejemplo de esta etapa puede citarse el caso del algoritmo de Sawada y colaboradores [19] que combinó las mejoras en cada una de esas etapas para generar un resultado de muy buena calidad.

Un problema de este enfoque es el costo computacional. Si bien éste es manejable en relación con el de los métodos en el dominio temporal, y permite el manejo de tiempos de reverberación relativamente largos, sigue siendo alto, y en particular bastante complicado para aplicaciones en tiempo real como el comando por voz de dispositivos. Un algoritmo interesante que apunta a reducir el costo computacional a partir de simplificar el modelo de mezclas fue propuesto en 2009 [20]. En este trabajo se mostró experimentalmente que, si las señales se adquieren con micrófonos lo suficientemente cercanos, las respuestas al impulso desde la posición de una fuente a cada micrófono pueden aproximarse como

versiones escaladas y retardadas entre ellas. Es decir, la morfología de las respuestas al impulso desde una fuente a todos los micrófonos sería similar, sólo afectada por un escalado y retardo. La importancia de esta simplificación, que se denominó modelo de mezclas pseudoanecóico, es que produce que las matrices de mezcla para cada frecuencia queden relacionadas entre sí, dependiendo sólo de los parámetros de atenuación y escalado. Esto implica que ya no es necesario resolver por separado un problema de ICA para cada frecuencia: basta resolver un sólo problema de ICA, en forma robusta, a partir de esa solución estimar los parámetros de atenuación y retardo de las respuestas al impulso, y a partir de ahí se pueden sintetizar las matrices de separación para todas las frecuencias. Esto hace que, por ejemplo, si se usan ventanas de longitud 512 muestras en la STFT, en lugar de tener que resolver 257 problemas de ICA (sólo se resuelven para las frecuencias positivas, por la simetría conjugada de la transformada de Fourier para señales reales), se pueda realizar la separación resolviendo sólo 1 problema de ICA, por lo que el tiempo de cálculo será reducido aproximadamente en esa proporción. Por otro lado, al sintetizar las matrices de separación a partir de los mismos parámetros, desaparece la indeterminación de permutaciones. La clave aquí es que, al resolver sólo un problema de ICA para estimar los parámetros, esto debe ser muy confiable, porque una mala estimación generará malos resultados. Para generar una estimación confiable se utilizan datos de varias frecuencias que se empaquetan juntos para darle más robustez a la estimación. Otros aspectos claves son la selección de qué frecuencia usar para esta estimación, y el uso de un postfiltro de Wiener para eliminar el ruido residual [18]. El método de empaquetar datos de varias frecuencias para hacer una estimación más robusta también fue explorado como mecanismo para evitar las permutaciones. Al acoplar información de varias frecuencias, los métodos de ICA resueltos en cada frecuencia presentan menos permutaciones, lo que aumenta la calidad de separación. Esto fue aplicado para mejorar el método de Sawada, en lo que se llamó Multibin ICA [21].

Otra variante interesante que intenta eliminar el problema de las permutaciones se denomina análisis de vectores independientes (IVA, del inglés Independent Vector Analysis) [22,23]. En este enfoque las mezclas para cada frecuencia no se consideran independientes entre sí, sino que se modelan acopladas mediante una distribución de probabilidades. De esta forma la optimización no se hace frecuencia por frecuencia sino en forma acoplada para todas las frecuencias a la vez, lo que elimina las permutaciones. El principal problema de este enfoque es su convergencia lenta, y que el modelo de

acoplamiento probabilístico entre las frecuencias es una aproximación que reduce la calidad de los resultados.

Otro algoritmo que ha tenido mucho éxito, dando resultados muy buenos en una amplia variedad de escenarios, fue introducido por Nesta y colaboradores en el año 2013, fusiona muchos de los conceptos que se han vertido hasta este punto [24]. Por un lado, las frecuencias no son tratadas independientemente, sino que se estima la inicialización para una nueva frecuencia a partir de las matrices de separación obtenidas anteriormente, filtradas con un filtro estimado para garantizar suavidad en el determinante de la matriz de separación. Esta idea surge a partir de la observación de que las matrices de separación deberían ser parecidas entre frecuencias sucesivas y por lo tanto sus determinantes deberían variar suavemente a lo largo de las frecuencias. En segundo lugar, se utilizan pesos en los datos usados para resolver el algoritmo de ICA en cada frecuencia, de forma similar a filtros de Wiener. Es decir, la estimación de matrices de covariancia implicadas en el algoritmo se realiza teniendo más en cuenta los instantes en que está activa la fuente de interés, para evitar que información residual de otras fuentes interfiera en dicha estimación. En tercer lugar, las permutaciones se resuelven mediante un algoritmo avanzado que es robusto incluso a la presencia de aliasing espacial (producido por micrófonos demasiado espaciados) [16]. Si bien su costo computacional es elevado, la calidad de separación suele ser muy buena, en especial para mezclas de muy corta duración, donde los otros algoritmos fallan.

Dadas las limitaciones del método de ICA en el dominio frecuencial, y principalmente la necesidad de contar con tantos micrófonos como fuentes, el interés en este tipo de algoritmos ha ido disminuyendo en los últimos tiempos, cambiando hacia métodos que permitan tratar el caso subcompleto que veremos en la próxima sección.

BSS basada en factorización de matrices nonegativas.

Uno de los principales problemas de los métodos de ICA en el dominio frecuencial reseñados anteriormente es que se aplican al problema determinado ($M = N$) o como mucho al sobredeterminado ($N > M$, con una etapa adicional de reducción dimensional basados en PCA que produce denoising y los lleva a

$M = N$). En muchos casos la cantidad de micrófonos es un limitante. No es raro imaginar en un living donde debe funcionar un sistema de control remoto de un televisor, que además de la voz de la persona dando comandos estarán presentes el sonido de los parlantes del televisor, tal vez el ruido de un sistema de aire acondicionado, de la pava en el fuego, de otras personas presentes en el cuarto, etc. Es poco realista y demasiado exigente respecto del hardware pretender tener tantos micrófonos como potenciales fuentes sonoras se encuentren en el cuarto. Por ello, es de interés el problema subdeterminado ($N < M$). Es en este contexto que surgió la aplicación de Factorización de Matrices Nonegativas (NMF, del inglés Nonnegative Matrix Factorization) para esta tarea.

La idea también es trabajar en el dominio tiempo-frecuencia después de una STFT. La magnitud de la STFT de una señal resulta una matriz de valores no negativos. La técnica de NMF factoriza una matriz X de dimensión PxQ en dos matrices W de PxK y H de KxQ, ambas de valores también no negativos, tales que $D(X,WH)$ sea mínima, donde $D(,)$ representa alguna métrica o "distancia" que puede tomar distintas formas. Es decir, NMF aproxima X como el producto de dos matrices de rango K. Las columnas de la matriz W pueden interpretarse como un diccionario que permite, mediante los coeficientes de la correspondiente columna de la matriz H, sintetizar cada columna de la matriz X. Estas técnicas surgieron para el análisis de otro tipo de datos, y han sido usadas con éxito para procesamiento de imágenes [25], para inferencia de funciones de genes [26], para agrupamiento [27], entre muchas otras aplicaciones y dominios.

Los primeros trabajos de separación de fuentes sonoras en esta línea se deben a Smaragdis [28], que utilizó un enfoque supervisado. La idea es utilizar ejemplos de las fuentes que se desea separar (por ejemplo, la voz de cada hablante, o los sonidos de distintos instrumentos musicales), grabadas en solitario, para entrenar diccionarios W específicos para cada fuente. Luego, cuando se tiene una mezcla de dichas fuentes capturada con un solo micrófono, se usa un diccionario conjunto armado concatenando los diccionarios específicos de cada fuente, y se estiman los coeficientes H. Finalmente se particiona esta H en los coeficientes correspondientes a cada diccionario, y con cada diccionario y sus coeficientes se reconstruye una STFT sólo de esa fuente. Como se utilizó la amplitud de la STFT, se usa la fase de la mezcla para poder reconstruir señales individuales mediante la transformación inversa.

Este método fue extendido usando bases convolutivas [29], luego el método NMF fue generalizado para incluir la fase de señales complejas [30], y luego fue generalizado para mezclas multicanal [31], con lo que se pueden aplicar estos métodos a señales adquiridas con múltiples micrófonos. En este último caso se ha generalizado al caso no supervisado, donde se utilizan modelos estadísticos para estimar el número de fuentes, agrupar los elementos del diccionario y de esa forma particionar el diccionario aprendido sobre las mezclas directamente. Esta línea ha despertado un gran interés en los últimos años, está en pleno desarrollo y la mayoría de los algoritmos para separación de fuentes que están apareciendo están basados en este tipo de técnicas. Un ejemplo reciente y muy interesante es el de Kitamura [32], que combina el enfoque de NMF multicanal, con IVA, para generar un algoritmo híbrido.

Evaluación de la calidad de separación.

Este es un aspecto no menor para el estudio de este tipo de algoritmos. ¿Cómo poder determinar si un algoritmo generado es efectivamente mejor que otro para una tarea de separación? Esto está estrechamente ligado a dos aspectos adicionales: la aplicación de interés y la metodología de experimentación. Respecto del primer aspecto, por ejemplo, un algoritmo que produzca resultados mejores desde el punto de vista subjetivo para una persona que escucha el resultado, no necesariamente será mejor para una tarea de reconocimiento automático del habla. Respecto del segundo aspecto, la idea es comparar qué tan buena resultó la señal separada por un algoritmo respecto del ideal de referencia. Pero, ¿Cuál es el ideal de referencia? ¿Es la fuente original? ¿o es tal vez la señal que habría llegado a cada micrófono si la fuente hubiera estado sola (es decir, que incluye el efecto reverberante del cuarto, a esto se le llama *imágenes* de las fuentes)? Según sea la respuesta, cambia el protocolo de experimentación: en el primer caso se reproducen a la vez todas las fuentes, mientras que en el segundo cada fuente se reproduce por separado. Además, ¿cómo se obtienen las mezclas? ¿Se graban las fuentes simultáneamente en un cuarto real? ¿O se graban las fuentes por separado y luego se suman los resultados? ¿O lo que se registra son las respuestas al impulso desde la posición de las fuentes a la posición de cada micrófono y luego generan todas las mezclas sintéticas que sea necesario a partir de la convolución de ellas y las fuentes deseadas? Cada una de estas preguntas implica un marco experimental diferente, y por lo tanto requiere una forma de evaluación diferente.

Un trabajo inicial que ha establecido el estándar para esta evaluación es el de Vincent y colaboradores [33]. Estos investigadores han desarrollado un conjunto de herramientas para evaluar la separación en diferentes escenarios: tanto mezclas reales o simuladas donde el objetivo sea estimar las fuentes originales, como mezclas reales o simuladas donde el objetivo sea estimar las imágenes de las fuentes obtenidas por cada micrófono. El corazón del método de evaluación consiste en descomponer la señal obtenida por el algoritmo, mediante proyecciones, en subespacios que permiten estimar cuánto de lo obtenido es atribuible a la señal de referencia, cuánto es distorsión residual del algoritmo, cuánto es debido a artefactos y cuánto se debe a las fuentes competidoras que no pudieron ser eliminadas. En base a estas proyecciones se producen medidas llamadas SDR (del inglés Signal to Distortion Ratio), SAR (del inglés Signal to Artifact Ratio) y SIR (del inglés Signal to Interference Ratio), que miden la relación de potencias entre la componente de señal deseada y la potencia de la distorsión, de los artefactos, y de las fuentes de interferencia, respectivamente. Los autores han desarrollado un toolbox de Matlab que permite evaluar estas medidas para los distintos escenarios experimentales [34].

Si la aplicación de interés es el reconocimiento automático del habla, en [35] y [36] se presentó un estudio exhaustivo de la correlación entre diferentes medidas de calidad disponibles y la tasa de reconocimiento de un sistema de reconocimiento automático del habla. Se encontró que la medida que mejor correlación presenta como predictor de la tasa de reconocimiento de palabras es la medida PESQ, que corresponde al estándar ITU P.862 para la estimación de calidad de habla en canales telefónicos y es conocida por proveer una alta correlación también con la calidad subjetiva percibida por personas. Estas dos propiedades, predecir muy bien la calidad perceptual y a la vez predecir la tasa de reconocimiento, la hacen una medida que ha sido adoptada en la comunidad para evaluar algoritmos de separación de fuentes sonoras.

Un aspecto no menor de la experimentación en esta área es la disponibilidad de conjuntos de datos estandarizados de mezclas sonoras que permitan una comparación justa entre algoritmos. Una fuente muy importante en este sentido han sido los concursos SISEC (del inglés Signal Separation Evaluation Campaign). Ha habido seis ediciones de los mismos, en cada una de ellas se proponen diferentes tareas relacionadas a la separación en distintos tipos de escenarios (subdeterminado, determinado, sobredeterminado), para

mezclas reales o artificiales, con distinto tipo de ruidos competidores, con mezclas de distintos tipos de señales (habla, música, señales biológicas), con distintos tipos de mezclas (instantáneas, anecóicas, reales, simuladas) [37]. Además de proveerse de datos específicos para cada una de las tareas, se proveen scripts para aplicar la separación a todos los datos, y también se incluyen medidas para evaluar la calidad de los resultados obtenidos. Además se presentan tablas con los resultados generados por todos los participantes de la campaña de evaluación, lo que permite rápidamente comparar un nuevo método contra los ya existentes en el estado del arte para esa tarea. Todo esto hace que dichas competiciones sean una fuente importantísima de información, algoritmos y datos para cualquiera que esté interesado en trabajar en esta área.

Resultados

Para dar una idea de los desempeños comparativos de los distintos algoritmos referenciados en este trabajo, se trabajó con un conjunto de mezclas reales provistas en la competencia SISEC2010 [38], para la tarea “Robust blind linear/non-linear separation of short two-sources-two-microphones recordings”. Estas consisten en grabaciones realizadas en dos habitaciones (por lo tanto distintos tiempos de reverberación), y para cada habitación, tres diferentes distribuciones espaciales de fuentes y micrófonos. En cada una de esas 6 condiciones se han mezclado diferentes fuentes en 6 combinaciones, incluyendo siempre una voz (que puede ser masculina o femenina), con un ruido que puede ser a su vez otra voz (masculina o femenina), el ruido de un estornudo, el ruido de risas, el ruido de fondo de un televisor, o el ruido de un cristal al romperse, para generar entonces 36 mezclas incluyendo diferentes tiempos de reverberación, diferentes distribuciones espaciales y diferentes tipos de ruido competidor. Las señales, originalmente disponibles a 44 KHz, fueron remuestreadas a 8 KHz. La evaluación de calidad se hizo con las medidas PESQ, SDR, SIR y SAR, promediando los resultados obtenidos para las 36 mezclas. Se presenta también el tiempo de ejecución promedio de los algoritmos, para tener una idea del costo computacional. Para todas las medidas, cuanto más grande mejor, con excepción del tiempo de calculo que se desea mantener lo más bajo posible. Todos los métodos están programados en el lenguaje Matlab, en una computadora con procesador Intel I3 modelo 6100U de doble núcleo a 2.3 GHz, con 8 GB de memoria. Se comparan los algoritmos de Murata [10] (el primero en resolver el problema de permutaciones adecuadamente), Sawada [19], Di Persia

1 [21], Di Persia 2 [20] y Nesta [24], los cuales son variantes de ICA en dominio frecuencial, Ono [23] basado en IVA, Ozerov [31] basado en NMF, y Kitamura [32], que combina NMF con IVA. Los algoritmos de Ikeda, Nesta y Ozerov fueron obtenidos de sus autores, los restantes implementados por el autor de este trabajo. Para todos los casos se utilizó una ventana de 1024 muestras. La Tabla I presenta los resultados obtenidos para cada algoritmo según las medidas mencionadas.

Tabla I – Resultados promedio de los algoritmos evaluados sobre los datos de SISEC2010. En negrita el mejor resultado para cada medida, y en itálica el segundo mejor.

Algoritmo	PESQ	SDR	SIR	SAR	Tiempo
Murata [10]	2,11	-4,74	0,88	-0,83	*0,73*
Sawada [19]	*2,83*	**2,89**	**9,36**	10,39	2,74
Di Persia 1 [21]	**2,85**	2,20	*9,19*	*11,35*	3,86
Di Persia 2 [18, 20]	2,66	-0,08	7,78	**15,62**	**0,38**
Nesta [24]	2,42	-0,93	5,22	10,75	1,89
Ono [23]	2,38	-0,99	2,26	8,04	3,30
Ozerov [31]	2,28	*2,43*	4,35	8,64	4,67
Kitamura [32]	2,50	-3,11	3,88	8,61	13,22

Ninguno de los algoritmos de separación resulta el mejor para todas las medidas utilizadas. El algoritmo de Sawada es el mejor respecto de las medidas SDR y SIR, quedando en segundo lugar en PESQ. El algoritmo de Di Persia 1 resulta el mejor en PESQ, quedando en segundo lugar en SDR y SAR. El algoritmo de Di Persia 2 resulta el mejor para la medida SAR y en tiempo de ejecución. Los demás algoritmos representan compromisos intermedios para las medidas. Los algoritmos más recientes basados en NMF todavía se encuentran lejos del desempeño obtenido por los métodos basados en ICA en el dominio frecuencial, pero tienen la ventaja de poder ser adaptados a casos donde se utilizan menos micrófonos que fuentes a separar, por lo que presentan mayores posibilidades de desarrollo a futuro.

Agradecimientos

El autor desea reconocer los aportes del Dr. Diego Milone, del Dr. Masuzo Yanagida, del Dr. Hugo Leonardo Rufiner, y de numerosos estudiantes que han contribuido a los trabajos reportados en este artículo durante más de 16 años de trabajo. Los mismos han sido financiados por el proyecto “Speech recognition in a specified space” por el MEXT (Ministry of Education, Culture, Sports, Science and Technology, Japon), los proyectos PICT 11-12700, PAE-PICT-2007-00052, PAE-PID-2007-00113, PICT 2010-1730 y PICT 2014-2627 financiados por ANPCyT, los proyectos CAID 12/G407, CAID Tipo III R4-N14 y CAID 2011 58-00519 financiados por la UNL, el proyecto PPCP-006-2011 financiado por la Secretaría de Políticas Universitarias, y el proyecto PIP 201101 000284 financiado por CONICET.

Referencias

[1] A. Hyvärinen, J. Karhunen & E. Oja, *Independent Component Analysis*, John Wiley & Sons, Inc., New York, 2001.

[2] P. Comon & C. Jutten, *Handbook of Blind Source Separation*, Academic Press, Oxford, 2010.

[3] M. Brandstein & D. Ward, Microphone Arrays, En: *Signal Processing Techniques and Applications* (M. Brandstein & D. Ward, Editores), Springer, Berlin, 2001.

[4] P. Cavalcanti, J. Scharcanski, L. E. Di Persia & D. H. Milone, *Proceedings of the 33rd Annual International IEEE EMBS Conference (EMBC 2011)*, Boston, 2011.

[5] A. Delorme, T. Sejnowski & S. Makeig, *Neuroimage* 34,1443 (2007).

[6] R. J. Martis, U. R. Acharya & L. C. Min, *Biomed. Signal Process. Control,* 8, 437 (2013).

[7] S. I. Amari, S. C. Douglas, A. Cichocki & H. H. Yang *Proceedings of the IEEE workshop on signal processing advances in wireless communications*, Paris, 1997.

[8] A. Cichocki & S. I. Amari, *Adaptive Blind Signal and Image Processing*, John Wiley & Sons, Chichester, 2002.

[9] E. Bingham & A. Hyvärinen, *Internation. J. Neural Syst.* 10, 1 (2000).

[10] N. Murata, S. Ikeda & A. Ziehe, *Neurocomputing* 41, 1 (2001).

[11] K. Matsuoka, *Proceedings of the SICE annual conference*, Osaka, 2002.

[12] S. Araki, R. Mukai, S. Makino, T. Nishikawa & H. Saruwatari, *IEEE Transact. Speech Audio Process.* 11, 109 (2003).

[13] M. Novey & T. Adali, *IEEE Transact. Neural Networks* 19, 596 (2008).

[14] J.-F. Cardozo & T. Adali, *Proceedings of the 2006 International Conference on Acustics, Speech and Signal Processing, ICASSP 2006*, Toulouse, (2006).

[15] C. Servière & D. T. A. P. Pham, *EURASIP J. Appl. Signal Process.* 2006, 1 (2006).

[16] F. Nesta & M. Omologo, *IEEE Transact. Audio Speech Lang. Process.* 20, 246 (2012).

[17] R. Aichner, M. Zourub, H. Buchner & W. Kellermann, *Proceedings of the 2006 International Conference on Audio, Speech and Signal Processing, ICASSP 2006*, Toulouse, (2006).

[18] L. E. Di Persia, D. H. Milone & M. Yanagida, *J. Signal Process. Syst.* 63, *333* (2011).

[19] H. Sawada, S. Araki & S. Makino, *Proceedings of the IEEE International Conference on Circuits and Systems*, Marrakesh, 2007.

[20] L. E. Di Persia, D. H. Milone & M. Yanagida, *IEEE Transact. Audio Speech Lang. Process.* 17, 299 (2009).

[21] L. E. Di Persia & D. H. Milone, *Signal Process.*116, 162 (2016).

[22] I. Lee, T. Kim & T.-W. Lee, *Signal Process.*87,1859 (2007).

[23] N. Ono, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, 2011.

[24] F. Nesta, P. Svaizer & M. Omologo, *IEEE Transact. Audio Speech Lang. Process.* 19, 624 (2011).

[25] P. Cavalcanti, J. Scharcanski, C. E. Martinez & L. E. Di Persia, *Proceedings of the IEEE International Instrumentation and Measurement Technology Conference*, Montevideo, 2014.

[26] L. E. Di Persia, G. Leale, G. Stegmayer & D. H. Milone, *Proceedings of the 4th ISCB-LA Bioinformatics Conference*, Buenos Aires, 2016.

[27] C. Ding, T. Li, W. Peng & H. Park, *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, Filadelfia, 2006.

[28] P. Smaragdis & J. C. Brown, *Proceedings of the IEEE Workshop on*

Applications of Signal Processing to Audio and Acoustics, New Paltz, 2003.

[29] P. Smaragdis, *IEEE Transact. Audio Speech Lang. Process.* 15, 1 (2007).

[30] H. Kameoka, N. Ono, K. Kashino & S. Sagayama, *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing, ICASSP 2009*, Taipei, (2009).

[31] A. Ozerov & C. Fevotte, *IEEE Transact. Audio Speech Lang. Process.* 18, 550 (2009).

[32] D. Kitamura, N. Ono, H. Sawada, H. Kameoka & H. Saruwatari, *IEEE Transact. Audio Speech Lang. Process.,* 24,1626 (2016).

[33] E. Vincent, R. Gribonval & C. Fevotte,*IEEE Transact. Audi Speech Lang. Process.* 14,1462 (2006).

[34] E. Vincent, «BSSEval. A toolbox for performance measurement in (blind) source separation,» [En línea].Disponible: http://bass-db.gforge.inria.fr/bss_eval/. [Último acceso: 29 Marzo 2017].

[35] L. E. Di Persia, D. H. Milone, M. Yanagida & H. L. Rufiner,*Signal Process.* 87,1951 (2007).

[36] L. E. Di Persia, D. H. Milone, H. L. Rufiner & M. Yanagida, *Signal Process.* 88, 2578 (2008).

[37] «SISEC 2016,» [En línea]. Disponible: https://sisec.inria.fr/. [Último acceso: 29 Marzo 2017].

[38] «SISEC 2010,» [En línea]. Disponible: http://sisec2010.wiki.irisa.fr/tiki-index.html. [Último acceso: 29 Marzo 2017].

Manuscrito recibido el 29 de marzo de 2017.
Aceptado el 28 de abril de 2017.