

Bases convolutivas de voz y su aplicación para separación de voz supervisada

Leandro Di Persia

*Laboratorio de Señales e Inteligencia Computacional,
Facultad de Ingeniería y Ciencias Hídricas,
ldipersia@gmail.com*

I. INTRODUCCIÓN

La búsqueda de bases para la descomposición de señales tiene innumerables aplicaciones tanto para el análisis como para la extracción de características de las mismas. Entre las técnicas utilizadas caben mencionarse las basadas en Análisis de Componentes Principales (PCA, Principal Component Analysis), Análisis de Componentes Independientes (ICA, Independent Component Analysis), métodos de subespacios basados en Descomposición en Valores Singulares (SVD, Singular Value Decomposition), etc.

En particular, recientemente se ha desarrollado la técnica de Factorización en Matrices No-Negativas (NMF, Nonnegative Matrix Factorization). Este enfoque está restringido a matrices de datos positivas. Propuesta por Lee y Seung en 1999 [1], esta técnica factoriza una matriz de entrada en un par de matrices con entradas no negativas, mediante la optimización de una función de costo específica derivada de la divergencia de Kullback-Leibler, bajo una restricción de positividad. Esta optimización se realiza mediante un algoritmo del tipo Maximización de la Esperanza (EM, Expectation Maximization), y resulta en una regla de actualización multiplicativa para las matrices.

El trabajo analizado [2] presenta una generalización del método NMF para la obtención de bases convolutivas. La idea básica de la extensión a bases convolutivas es permitir la existencia de patrones repetitivos a lo largo de las columnas de la matriz a factorizar. Si se utilizan bases comunes, estos patrones repetitivos deberían ser generados a través de múltiples activaciones de algunos elementos de la base, mientras que al permitir la existencia de bases convolutivas se pueden representar mediante la activación de una sola o unas pocas de ellas. Esta idea es particularmente interesante para descomponer espectrogramas de señales de voz, donde se verifica la existencia de patrones espectrales que se repiten a lo largo de las ventanas de análisis.

Más aún, en el artículo se propone la utilización de dichas bases convolutivas para separación ciega de fuentes a partir de un solo canal de grabación, en modo supervisado. La idea es, a partir de grabaciones de múltiples hablantes, para cada uno de ellos generar sus correspondientes bases convolutivas. Luego armar un diccionario con estas bases convolutivas, y para una señal que contiene una mezcla de varios de esos hablantes, dejando fija la matriz de bases con las bases de todos, iterar para determinar los coeficientes óptimos para reconstrucción. Luego se reconstruyen señales independientes utilizando sólo las bases y los coeficientes encontrados para cada hablante.

El artículo presenta una serie de experimentos destinados a mostrar, por un lado, las características de las bases

convolutivas generadas. Además, se realiza un estudio de los parámetros óptimos a utilizar. Finalmente se demuestra la capacidad de separación utilizando medidas objetivas de calidad.

II. ANÁLISIS GENERAL

A. Estructura general

El trabajo presenta una nueva técnica de obtención de bases convolutivas. El aporte es relevante, dado que permite obtener bases que, por un lado, semejan estructuras básicas de la estructura del habla, y por otro, al aplicarse a mezclas tienen características propias de las voces independientes. El desarrollo de la extensión a mezclas convolutivas se presenta en forma empírica, sin demostrar su convergencia. Si bien sería deseable contar con una verificación teórica de dicha convergencia, el algoritmo parece funcionar adecuadamente en los extensos experimentos, con lo cual se supone que dichas demostraciones pueden dejarse para trabajos futuros.

Las conclusiones son interesantes, y si bien se trata de un método supervisado, puede tener aplicación en algunas áreas como control de acceso restringido e incluso como método de identificación del hablante.

B. Análisis de metodología

Un aspecto de la técnica no analizado por el autor es el costo computacional, que se presume elevado dado que, si bien el algoritmo converge en alrededor de 100 iteraciones, cada una de ellas implica multiplicaciones matriciales con matrices de gran dimensión. Esto se ve empeorado en el caso de separación presentado, donde se deben utilizar simultáneamente las bases para todas las fuentes de interés, lo que aparentemente limitaría el número de hablantes que el método podría separar.

Hay un error relativamente grave en la formulación teórica. En la definición de la función de costo (ecuación (1) del manuscrito original) se define la función de costo en base a la norma de Frobenius. Pero en el trabajo original sobre NMF (y las demostraciones de la convergencia del mismo que se darán) se define diferente, ya que la norma de Frobenius implicaría sumar todos los elementos de la matriz resultante *elevados al cuadrado*, mientras que en el trabajo original se los suma directamente, sin elevarlos al cuadrado.

En la sección de post-proceso, el filtro utilizado tiene la estructura de un filtro de Wiener tiempo-frecuencia, si bien esto no es analizado por el autor. A la vista de esto, se podría analizar dicho post procesamiento desde el punto de vista de la validez de cumplimiento de las hipótesis de tal filtro.

C. Bibliografía

La bibliografía aparece como un tanto escueta, sin embargo si se tiene en cuenta que el método es completamente original, y no hay otros enfoques dedicados a la obtención de bases convolutivas (al menos en conocimiento de este revisor), puede considerarse razonable.

Como evaluación global, el artículo resulta muy interesante, no solo por la aplicación desarrollada en el mismo, sino también por las aplicaciones potenciales de las bases convolutivas en el análisis y extracción de características.

III. DESARROLLO

En esta sección se detallará el algoritmo de entrenamiento de NMF, incluyendo una demostración de su convergencia, y luego se explayará en la generalización a mezclas convolutivas.

A. Algoritmo NMF

El algoritmo para NMF fue introducido por Lee y Seung en [1], con una demostración de la convergencia de los algoritmos en [3] Dada una matriz $V \in \mathbb{R}^{\geq 0, M \times N}$, se buscan dos matrices no negativas $W \in \mathbb{R}^{\geq 0, M \times R}$ y $H \in \mathbb{R}^{\geq 0, R \times N}$ tales que $V = WH$, donde R es el número de bases (tamaño del diccionario) a utilizar. Se minimiza para esto el error de reconstrucción, midiendo este error con una función de costo adecuada. La función de costo utilizada en este trabajo (y también una de las usadas originalmente) es una variante de la divergencia de Kullback-Leibler:

$$D(V, \hat{V}) = \sum_i \sum_j \left(V_{ij} \log \left(\frac{V_{ij}}{\hat{V}_{ij}} \right) - V_{ij} + \hat{V}_{ij} \right) \quad (1)$$

donde $\hat{V} = W * H$ es la reconstrucción a partir de la factorización obtenida. Esta ecuación se puede ver como la divergencia de Kullback-Leibler, si las matrices intervinientes estuvieran normalizadas de forma que la suma de todos sus elementos fuera 1 y por lo tanto se pudieran interpretar como densidades de probabilidad.

El problema se plantea entonces como minimizar $D(V, WH)$, sujeto a las restricciones $W \geq 0$ y $H \geq 0$.

Ahora, esta claro que se puede escribir $D(V, \hat{V}) = \sum_j D(\mathbf{v}^j, \hat{\mathbf{v}}^j)$, donde \mathbf{v}^j y $\hat{\mathbf{v}}^j$ denotan las j -ésimas columnas de V y \hat{V} , respectivamente. Esto implica que se puede minimizar $D(V, \hat{V})$ si se minimiza el criterio para cada columna individual. En las demostraciones siguientes se seguirá este criterio sobre cada columna, dejando de lado el superíndice j para evitar sobrecargar la notación.

Realizando esta optimización como se demostrará mas adelante, se llega a una estructura iterativa en dos pasos, primero actualizando W y luego actualizando H , con una regla de actualización de tipo multiplicativa. Dado el carácter multiplicativo de la actualización de pesos, basta con asegurar una inicialización con valores estrictamente positivos para asegurar el cumplimiento de las restricciones de positividad en la optimización. Las reglas de optimización están dadas en el siguiente teorema.

Teorema 1: La divergencia $D(V, WH)$ no se incrementa bajo las reglas de actualización:

$$H_{ij} = H_{ij} \frac{\sum_k W_{ki} V_{kj} / (WH)_{kj}}{\sum_l W_{li}} \quad (2)$$

$$W_{ij} = W_{ij} \frac{\sum_k W_{jk} V_{ik} / (WH)_{ik}}{\sum_l W_{jl}} \quad (3)$$

Más aún, la divergencia es invariante frente a estas actualizaciones si y solo si W y H están en un punto estacionario de la divergencia.

La demostración de este teorema no es trivial y se requieren ciertos pasos intermedios que se detallarán en la próxima sección.

B. Demostración de la convergencia

Para poder demostrar las ecuaciones de actualización se utilizará una técnica similar a la usada en el método EM, utilizando una función auxiliar derivada de la función objetivo. En este contexto, se define una función de los parámetros actuales y el vector de parámetros objetivo, y se optimiza esta función de costo. Para que esto funcione, la nueva función auxiliar debe estar relacionada con la función de costo original [3]. Esto queda plasmado en la siguiente definición.

Definición 1 (Función Auxiliar): Sea una función de costo $F(\mathbf{h})$. La función $G(\mathbf{h}, \mathbf{h}')$ es una *función auxiliar* de $F(\mathbf{h})$ si se verifica:

$$G(\mathbf{h}, \mathbf{h}') \geq F(\mathbf{h}), \quad G(\mathbf{h}, \mathbf{h}) = F(\mathbf{h}). \quad (4)$$

El siguiente lema permite asegurar que si en un paso se elige como nuevo vector \mathbf{h} aquel que minimice la función auxiliar, la función de costo será reducida (o en el peor de los casos será igual).

Lema 1: Si $G(\mathbf{h}, \mathbf{h}')$ es una función auxiliar, entonces la función de costo asociada $F(\mathbf{h})$ no se incrementa bajo la actualización $\mathbf{h}^{t+1} = \operatorname{argmin}_{\mathbf{h}} G(\mathbf{h}, \mathbf{h}^t)$.

Demostración: Dada la regla de actualización, se tiene que $G(\mathbf{h}^{t+1}, \mathbf{h}^t) \leq G(\mathbf{h}, \mathbf{h}^t) \forall \mathbf{h}$, en particular para $\mathbf{h} = \mathbf{h}^t$. Además, por la definición de función auxiliar se cumple que $G(\mathbf{h}^{t+1}, \mathbf{h}^t) \geq F(\mathbf{h}^{t+1})$. Uniendo estas dos, se tiene:

$$F(\mathbf{h}^{t+1}) \leq G(\mathbf{h}^{t+1}, \mathbf{h}^t) \leq G(\mathbf{h}^t, \mathbf{h}^t) = F(\mathbf{h}^t) \quad (5)$$

y por lo tanto $F(\mathbf{h})$ no se incrementa frente a la actualización. ■

Esto nos permite delinear la siguiente demostración para el teorema: si logramos definir una función auxiliar para la función de costo deseada (y demostramos que en efecto cumple con las condiciones requeridas para ser una función auxiliar), entonces podemos derivar la ecuación de actualización minimizando dicha función auxiliar. Esto lo tenemos que realizar para ambas matrices, W y H . Se toma W fijo y se actualiza H , y luego se toma H fijo y se actualiza W . El lema anterior garantiza que en cada actualización la función de costo subyacente no se incrementa, y por lo tanto el algoritmo converge. La demostración se realizará para la actualización de H , siendo la demostración para W exactamente análoga. Además, como ya se dijo, se trabajará con la actualización de cada columna de h .

Lema 2: Sea la función G :

$$G(\mathbf{h}, \mathbf{h}^t) = \sum_i (v_i \log v_i - v_i) + \sum_i \sum_a W_{ia} h_a \quad (6)$$

$$- \sum_i \sum_a v_i \frac{W_{ia} h_a^t}{\sum_b W_{ib} h_b^t} \left(\log W_{ia} h_a - \log \frac{W_{ia} h_a^t}{\sum_b W_{ib} h_b^t} \right)$$

Esta es una *función Auxiliar* para

$$F(\mathbf{h}) = \sum_i \left\{ v_i \log \left(\frac{v_i}{\sum_a W_{ia} h_a} \right) - v_i + \sum_a W_{ia} h_a \right\}. \quad (7)$$

Demostración: Para demostrar esto se deben probar las dos condiciones dadas en la definición 1, esto es, que $G(\mathbf{h}, \mathbf{h}) = F(\mathbf{h})$ y $G(\mathbf{h}, \mathbf{h}^t) \geq F(\mathbf{h})$. La primera parte se verifica fácilmente, haciendo $\mathbf{h}^t = \mathbf{h}$:

$$\begin{aligned}
G(\mathbf{h}, \mathbf{h}) &= \sum_i (v_i \log v_i - v_i) + \sum_i \sum_a W_{ia} h_a \quad (8) \\
&\quad - \sum_i \sum_a v_i \frac{W_{ia} h_a}{\sum_b W_{ib} h_b} \\
&\quad \left(\log W_{ia} h_a - \log W_{ia} h_a + \log \sum_b W_{ib} h_b \right) \\
&= \sum_i (v_i \log v_i - v_i) + \sum_i \sum_a W_{ia} h_a \\
&\quad - \sum_i \sum_a v_i \frac{W_{ia} h_a}{\sum_b W_{ib} h_b} \log \sum_b W_{ib} h_b \\
&= \sum_i (v_i \log v_i - v_i) + \sum_i \sum_a W_{ia} h_a \\
&\quad - \sum_i v_i \log \sum_b W_{ib} h_b \frac{\sum_a W_{ia} h_a}{\sum_b W_{ib} h_b} \\
&= \sum_i (v_i \log \frac{v_i}{\sum_b W_{ib} h_b} - v_i) + \sum_i \sum_a W_{ia} h_a \\
&= F(\mathbf{h}).
\end{aligned}$$

Luego queda demostrar que $G(\mathbf{h}, \mathbf{h}^t) \geq F(\mathbf{h})$. Para esto se utiliza la siguiente desigualdad:

$$-\log \sum_a W_{ia} h_a \leq -\sum_a \alpha_a \log \frac{W_{ia} h_a}{\alpha_a} \quad (9)$$

la cual es válida dada la convexidad del logaritmo negativo, para todo los α_a no negativos que verifiquen $\sum_a \alpha_a = 1$. Para ver esto, una función $f(x)$ es convexa, si se verifica que $f(E\{x\}) \leq E\{f(x)\}$, con $E\{\cdot\}$ el operador esperanza. Dado un conjunto de $\alpha_a \geq 0$ tales que $\sum_a \alpha_a = 1$, dichos coeficientes se pueden interpretar como una distribución de probabilidad, con lo cual la ecuación de convexidad resulta $f(\sum_a x_a \alpha_a) \leq \sum_a \alpha_a f(x_a)$. Ahora, tomando $x_a = W_{ia} h_a / \alpha_a$, y $f(x) = -\log(x)$ (la cual es convexa), se deduce la inecuación anterior. Así, usando

$$\alpha_a = \frac{W_{ia} h_a^t}{\sum_b W_{ib} h_b^t} \quad (10)$$

en (9), se obtiene finalmente

$$\begin{aligned}
-\log \sum_a W_{ia} h_a &\leq -\sum_a \frac{W_{ia} h_a^t}{\sum_b W_{ib} h_b^t} \quad (11) \\
&\quad \left(\log W_{ia} h_a - \log \frac{W_{ia} h_a^t}{\sum_b W_{ib} h_b^t} \right)
\end{aligned}$$

la cual implica que $G(\mathbf{h}, \mathbf{h}^t) \geq F(\mathbf{h})$ ya que estos son los términos en los que difieren las ecuaciones para G y F . ■

Completada la demostración de la existencia de una función auxiliar, resta demostrar la ecuación de reestimación. *Demostración:* [Teorema 1] Para encontrar el mínimo de G se toma el gradiente e iguala a cero:

$$\frac{\partial G(\mathbf{h}, \mathbf{h}^t)}{\partial h_a} = -\sum_i v_i \frac{h_a^t}{h_a} \frac{W_{ia}}{\sum_b W_{ib} h_b^t} + \sum_i W_{ia} = 0. \quad (12)$$

$$\Rightarrow h_a^{t+1} = h_a^{opt} = \frac{h_a^t}{\sum_k W_{ka}} \sum_i \frac{v_i}{\sum_b W_{ib} h_b^t} W_{ia}. \quad (13)$$

la cual es la ecuación buscada. ■

C. Generalización a mezclas convolutivas

Para completar esta sección, se detalla la generalización a mezclas convolutivas presentada en el trabajo. Antes de esto, utilizaremos una notación más compacta para las actualizaciones, expresándola en términos de matrices.

$$H = H \odot \frac{W^\top \frac{V}{\hat{V}}}{W^\top \mathbf{1}} \quad (14)$$

$$W = W \odot \frac{\frac{V}{\hat{V}} H^\top}{\mathbf{1} H^\top} \quad (15)$$

donde el símbolo \odot representa un producto elemento a elemento, y las divisiones involucradas también son elemento a elemento. El símbolo $\mathbf{1}$ representa una matriz de $M \times N$ con todos sus elementos iguales a 1, y el símbolo \top representa la transposición.

Para generalizar el modelo estándar de NMF al caso de bases convolutivas se redefine la ecuación de reconstrucción, de la siguiente forma:

$$\hat{V} = \sum_{t=0}^{T-1} W(t) \overset{t \rightarrow}{H} \quad (16)$$

donde V es la matriz a descomponer como antes, $W(t)$ es ahora un conjunto de T matrices que constituyen la base convolutiva, y H contiene los pesos de cada base en la combinación. El operador $\overset{i \rightarrow}{\cdot}$ es un operador de desplazamiento que mueve las columnas de la matriz a la que se le aplica en i unidades a la derecha, completando con columnas de ceros por la izquierda. Análogamente se utilizará el operador $\overset{\leftarrow{i}}{\cdot}$ para indicar el desplazamiento de i columnas a la izquierda.

Debe destacarse cómo están constituidas las bases convolutivas. En el método NMF estándar, cada columna de la matriz W es una base. En esta extensión, las T columnas i -ésimas de cada matriz $W(t)$ constituyen la i -ésima base convolutiva.

Utilizando la nueva forma de reconstrucción, la ecuación de costo se puede escribir como:

$$\begin{aligned}
D(V, \hat{V}) &= \sum_i \sum_j \left(V \log \left(\frac{V}{\hat{V}} \right) - V + \hat{V} \right)_{ij} \quad (17) \\
&= \sum_i \sum_j \left(V \left[\log V - \log \left(\sum_{t=0}^{T-1} W(t) \overset{t \rightarrow}{H} \right) \right] \right. \\
&\quad \left. - V + \sum_{t=0}^{T-1} W(t) \overset{t \rightarrow}{H} \right)_{ij}
\end{aligned}$$

Dada forma de esta función de costo, con términos que dependen de las matrices para cada t combinados aditivamente, se la puede descomponer en un conjunto de aproximaciones NMF simultáneas, una para cada tiempo t . Entonces para cada uno de los problemas NMF hay que adaptar la correspondiente $W(t)$ y H (después de un desplazamiento adecuado). Esto da lugar a las siguientes ecuaciones de actualización:

$$H = H \odot \frac{W(t)^\top \left[\frac{V}{\hat{V}} \right]}{W(t)^\top \mathbf{1}} \quad (18)$$

$$W(t) = W(t) \odot \frac{\frac{V}{\hat{V}} \overset{t \rightarrow}{H}}{\mathbf{1} \overset{t \rightarrow}{H}} \quad (19)$$

Lo que estas ecuaciones implican es que para cada tiempo t , se debe actualizar $W(t)$ y correspondientemente H . Como para cada t se tiene una $W(t)$ diferente pero todas comparten la misma H , cada $W(t)$ tiene efecto sobre una versión desplazada de H , $H^{\leftarrow t}$, es por eso que en la ecuación su actualización, el efecto primero es llevado hacia la izquierda antes de aplicarlo a H . Además, esto implica que habría múltiples actualizaciones a H , con algunos de sus elementos recibiendo múltiples mejoras, y por lo tanto la matriz estimada tendría un sesgo importante, con el elemento en $t = T - 1$ dominando sobre los demás. Es por esto que una mejor estrategia es, en lugar de aplicar todas las actualizaciones en cascada, hacer un promediado de las actualizaciones parciales sobre H anterior:

$$H = \frac{1}{T} \sum_{t=0}^{T-1} \left(H \odot \frac{W(t)^\top \begin{bmatrix} \leftarrow t \\ \mathbf{V} \\ \leftarrow t \end{bmatrix}}{W(t)^\top \mathbf{1}} \right). \quad (20)$$

D. Aplicación a separación ciega de fuentes

Utilizando las bases convolutivas así estimadas, se las puede utilizar en forma supervisada para aplicarla a separación de fuentes. Para esto, la factorización se aplica al espectrograma de un registro monofónico de audio. El espectrograma es no negativo, con lo cual cumple el requisito principal para esta factorización. Los autores aplican el método a dos voces individualmente, y luego a una mezcla de ellas. Por observación, encuentran que en las bases obtenidas de la mezcla hay algunas que son muy parecidas a las de uno de los hablantes, mientras que otras son muy parecidas al otro hablante. Por lo tanto utilizan la hipótesis de mezcla lineal de los espectros (la cual no necesariamente es realista, sobre todo en recintos reverberantes), y suponen que podrían usar las bases aprendidas individualmente para reconstruir la mezcla. Esto lleva al siguiente algoritmo de separación supervisado:

1. Dadas señales de audio pronunciadas por P hablantes individualmente, entrenar un conjunto $W(t)^p$, con $1 < p < P$, bases convolutivas, esto es, un conjunto de bases convolutivas para cada hablante.
2. Construir un diccionario mayor, juntando las bases obtenidas para todos los hablantes: $W(t) = W(t)^1 \cup W(t)^2 \cup \dots \cup W(t)^P$.
3. Dado el espectrograma de una mezcla de varios de estos hablantes, realizar un entrenamiento del NMF convolutivo, pero manteniendo constante la $W(t)$, en el diccionario global para todos los hablantes, adaptando sólo los pesos H . Es decir, se deja fijo el diccionario y se determinan los coeficientes óptimos para reconstruir la señal usando el mismo.
4. Particionar el diccionario y los pesos obtenidos, en partes correspondientes a cada uno de los hablantes. Con estos, sintetizar utilizando la (16) cada uno de los espectrogramas componentes.
5. Utilizar la fase de la mezcla en cada espectrograma para antitransformar y obtener las señales en dominio temporal.

Este último punto es un tanto chocante, porque se usa la fase original mezclada sin ningún procesamiento. Es sin embargo necesario para poder antitransformar. Es el mismo tipo de método que se usa en algoritmos de realce que

operan sobre el espectro de magnitud, como al aplicar un filtro de Wiener o substracción espectral.

IV. ANÁLISIS DE RESULTADOS

El algoritmo delineado en la sección anterior se ha probado en numerosos experimentos tendientes a evaluar el efecto de los distintos parámetros en la calidad de separación. Para evaluar la calidad de separación el autor utiliza algunas medidas derivadas de la correlación, que en definitiva representan variantes de la SNR. Como bien lo nota el autor, si bien estas medidas pueden ser razonables para tener una idea general de la calidad de separación, no son muy adecuadas dadas las características no lineales del procesamiento realizado. Sería interesante que se hubieran utilizado otro tipo de medidas de calidad, como medidas perceptuales (PESQ, WSS). Dada esta ausencia, es difícil evaluar el efecto sonoro de la separación obtenida (es decir, se reporta una mejora de 5 dB en el índice SR, esto implica que la señal se escucha mejor? que tanto mejor?).

En la parte experimental el autor se detiene para realizar un extenso número de validaciones, teniendo en cuenta los parámetros principales que afectan al algoritmo propuesto, a saber, la longitud de ventana utilizada en la FFT, el número de bases convolutivas a extraer (R), y la cantidad de valores temporales para cada una (T). Los experimentos están bien diseñados, utilizando varios valores para cada uno de estos parámetros. El único aspecto un tanto dudoso es el número de mezclas usadas, que son ocho solamente. Aún así, dada la gran cantidad de experimentos necesarios para evaluar todas las combinaciones de los parámetros (180 para cada par), puede considerarse una simplificación razonable destinada a encontrar un conjunto de parámetros adecuados.

Luego el autor se detiene en una serie de experimentos destinados a evaluar los restantes parámetros, que muestran tener una menor influencia sobre la calidad de separación. Los resultados son expresados en términos claros y las figuras están correctamente referenciadas y explicadas en el texto.

Un aspecto que no ha sido tenido en cuenta es que el framework ha sido propuesto para separar señales de múltiples fuentes, pero todos los experimentos reportan separación sobre sólo dos fuentes. Sería interesante ver evaluado el método con más fuentes activas. Por otro lado, en lo que respecta a separación de voces, siempre utiliza pares de voces donde uno es un hablante masculino y otro un hablante femenino. Dado que las características armónicas de estos dos serán claramente diferentes, es razonable que esto permita obtener buenos resultados mediante las bases individuales. Sería interesante poder evaluar si realmente se logra una separación para casos del mismo sexo. Si bien el autor trata algo de esto hacia el final de la sección V.D, solo lo hace superficialmente.

Por otro lado en el último experimento se utilizan no ya voces, sino ruidos de características muy diferentes a la voz, lo cual aparece como casos ideales para aplicar este método, con lo cual no es de extrañar que se reporte por ejemplo una mejora de más de 16 dB para el caso de campanillas (que no sólo tienen características espectrales muy diferentes a la de voz, sino que se presentan en zonas puntuales y aisladas del espectro, lo que favorece su remoción).

El autor también establece que este método podría usarse sin problemas para sonidos afectados por reverberación,

debido a las características de la convolución implícita de las bases. Si bien esto a priori suena razonable, este revisor tiene sus dudas al respecto, debido a que el fenómeno de reverberación afectaría a la principal hipótesis del algoritmo, que es que los espectrogramas de las fuentes se mezclan linealmente. Para pares de fuentes sin ecos, los términos con productos cruzados pueden ser de pequeño valor y no afectar mucho a la separación, pero indudablemente con más fuentes activas (y múltiples ecos provenientes de diferentes direcciones pueden interpretarse como múltiples fuentes correlacionadas), estos términos cruzados terminarán por llegar a magnitudes suficientes para invalidar el método. De todas maneras el tema debería ser sujeto de mayor investigación, dado que en este trabajo sólo se presentan resultados sobre mezclas instantáneas.

REFERENCIAS

- [1] Daniel D. Lee y H. Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.
- [2] Paris Smaragdis, "Convolutional speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.
- [3] Daniel D. Lee y H. Sebastian Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.