UNIVERSIDAD NACIONAL DEL LITORAL

DOCTORADO EN INGENIERÍA

Modelos de factorización en matrices no-negativas para procesamiento de audio

Francisco Javier Ibarrola

FICH
FACULTAD DE INGENIERÍA
Y CIENCIAS HÍDRICAS

INTEC
INSTITUTO DE DESARROLLO TECNOLÓGICO
PARA LA INDUSTRIA QUÍMICA

CIMEC
CENTRO DE INVESTIGACIÓN DE
MÉTODOS COMPUTACIONALES

sinc(i)
INSTITUTO DE INVESTIGACIÓN EN SEÑALES
SISTEMAS E INTELIGENCIA COMPUTACIONAL

Tesis de Doctorado **2019**

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Desarrollo Tecnológico para la Industria Química
Centro de Investigación de Métodos Computacionales
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional

MODELOS DE FACTORIZACIÓN EN MATRICES NO-NEGATIVAS PARA PROCESAMIENTO DE AUDIO

Francisco Javier Ibarrola

Tesis remitida al Comité Académico del Doctorado como
parte de los requisitos para la obtención del grado de
DOCTOR EN INGENIERÍA
Mención en Inteligencia Computacional, Señales y Sistemas
de la
UNIVERSIDAD NACIONAL DEL LITORAL

2019

Comisión de Posgrado, Facultad de Ingeniería y Ciencias Hídricas, Ciudad Universitaria, Paraje "El Pozo", S3000, Santa Fe, Argentina.

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Desarrollo Tecnológico para la Industria Química
Centro de Investigación de Métodos Computacionales
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional

MODELOS DE FACTORIZACIÓN EN MATRICES NO-NEGATIVAS PARA PROCESAMIENTO DE AUDIO

Francisco Javier Ibarrola

Lugar de Trabajo:

sinc(i)
Instituto de Señales, Sistemas e Inteligencia Computacional
Facultad de Ingeniería y Ciencias Hídricas
Universidad Nacional del Litoral

Director:

Dr. Leandro E. Di Persia	sinc(i), CONICET-UNL

Co-director:

Dr. Rubén D. Spies	IMAL, CONICET-UNL

Jurado Evaluador:

Dr. Hugo Aimar	IMAL, CONICET-UNL
Dra. Ana Georgina Flesia	CIEM, CONICET-UNC
Dr. Gastón Schlotthauer	IBB, CONICET-UNER

2019

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas

Santa Fe, 6 de Septiembre de 2019.

Como miembros del Jurado Evaluador de la Tesis de Doctorado en Ingeniería titulada ***"Modelos de factorización en matrices no-negativas para procesamiento de audio"***, desarrollada por el Lic. Francisco Javier IBARROLA, en el marco de la Mención "Inteligencia Computacional, Señales y Sistemas", certificamos que hemos evaluado la Tesis y recomendamos que sea aceptada como parte de los requisitos para la obtención del título de Doctor en Ingeniería.

La aprobación final de esta disertación estará condicionada a la presentación de dos copias encuadernadas de la versión final de la Tesis ante el Comité Académico del Doctorado en Ingeniería.

Dr. Hugo Aimar

Dr. Gastón Schlotthauer

Dra.Ana Georgina Flesia

Santa Fe, 6 de Septiembre de 2019

Certifico haber leído la Tesis, preparada bajo mi dirección en el marco de la Mención "Inteligencia Computacional, Señales y Sistemas" y recomiendo que sea aceptada como parte de los requisitos para la obtención del título de Doctor en Ingeniería.

Dr. Rubén Spies
Codirector de Tesis

Dr. Leandro Di Persia
Director de Tesis

DECLARACIÓN LEGAL DEL AUTOR

TESIS POR COMPILACIÓN

La presente tesis se encuentra organizada bajo el formato de Tesis por Compilación, aprobado en la resolución No 255/17 (Expte. No 888317-17) por el Comité Académico de la Carrera Doctorado en Ingeniería, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral (UNL). De dicha resolución:

"En el caso de optar por la Tesis por Compilación, ésta consistirá en una descripción técnica de al menos 30 páginas, redactada en español e incluyendo todas las investigaciones abordadas en la tesis. Se deberán incluir las secciones habituales indicadas a continuación en la Sección Contenidos de la Tesis. Los artículos científicos publicados por el autor, en el idioma original de las publicaciones, deberán incluirse en un Anexo con el formato unificado al estilo general de la Tesis indicado en la Sección Formato. El Anexo deberá estar encabezado por una sección donde el tesista detalle para cada una de las publicaciones cuál ha sido su contribución. Esta sección deberá estar avalada por su director de Tesis. El documento central de la Tesis debe incluir referencias explícitas a todas las publicaciones anexadas y presentar una conclusión que muestre la coherencia de dichos trabajos con el hilo conceptual y metodológico de la tesis. Los artículos presentados en los anexos podrán ser artículos publicados, aceptados para publicación (en prensa) o en revisión."

AGRADECIMIENTOS

A Rubén, que fue mi guía durante este proceso, y mucho antes también. Que me dio la oportunidad de transitar este camino cuando parecía que ya era tarde.

A Leandro, que sin conocerme y hablando idiomas diferentes se arriesgó a dirigirme, y me acompañó en cada paso.

A todos los que directa o indirectamente hayan hecho de este viaje algo un poquito mejor.

A Cristina, sin cuyo amor nada de esto hubiera sido posible.

Índice general

Índice de tablas

Índice de figuras

Resumen

A la par de los avances tecnológicos en cuanto a la capacidad de cómputo de los aparatos electrónicos portátiles, ha surgido en los últimos años la necesidad de agilizar la interacción entre hombre y máquina. Dado que el habla constituye la manera más natural de comunicación entre personas, este medio ha buscado ser extrapolado a la interacción entre personas y aparatos electrónicos.

Uno de los principales problemas a la hora de establecer una comunicación oral fluida tiene que ver con que en la vida real, el dispositivo de grabación no tiene oportunidad de registrar la señal deseada de manera directa. Varios tipos de distorsiones intervienen en el proceso, como ser la presencia de ruido, las reflexiones de la propia señal de audio en la habitación en que se registra y la presencia de otras fuentes sonoras.

Si bien estos fenómenos admiten un modelado en el dominio temporal, el costo computacional que implican algunos de los procedimientos necesarios para tratar los problemas asociados puede resultar prohibitivo a la hora de realizar implementaciones numéricas. Para evitar este inconveniente y a la vez ganar interpretabilidad, podemos hacer uso de la Transformada de Fourier de Tiempo Corto (STFT, por sus siglas en inglés), a través de la cual una señal de audio está unívocamente determinada por una matriz de elementos complejos. El espectrograma asociado a esta transformada es una matriz cuyos elementos son las amplitudes al cuadrado de la STFT. Es en este contexto que toma relevancia la factorización en matrices no negativas (NMF), técnica de modelado que permite una representación de los datos por partes y puramente aditiva. En otras palabras, este enfoque y sus variantes asocian matrices a elementos constitutivos de los datos.

Los enfoques clásicos de NMF aplicados a procesamiento de señales de audio normalmente presentan ciertas dificultades. Por un lado, la cantidad de variables suele ser mayor que la dimensión de los datos, por lo que no existe unicidad en las representaciones encontradas, lo que influye negativamente en la calidad de las restauraciones que pueden obtenerse. Por otro lado, la mayoría de los métodos de optimización del estado del arte están basados en algoritmos iterativos y multiplicativos, que no son inmediatamente adaptables a los casos en que el modelo contempla ciertas relaciones temporales o frecuenciales entre sus elementos.

En esta tesis se desarrollan nuevos métodos de representación en matrices no negativas para abordar los problemas de dereveberación y separación de fuentes de habla. En primer lugar, se introducen los fenómenos de reverberación y mezcla en su formulación temporal, y se explica su traslado al dominio tiempo-frecuencia. A partir de esto se formulan representaciones en matrices no negativas y a través de un enfoque bayesiano y sus funciones de penalización asociadas se imponen características particulares sobre sus elementos. Esta estructura impuesta permite superar algunas dificultades clásicas en este contexto, que tienen que ver con la falta de unicidad y de correlación en las representaciones. Además, se desarrollan algoritmos de optimización para resolver los problemas de minimización asociados a los funcionales construidos, que permiten incorporar ciertos tipos de penalizantes que por sus características no pueden ser definidos

elemento a elemento.

Los algoritmos desarrollados fueron puestos a prueba, comparándolos con métodos del estado del arte, en condiciones simuladas y reales. La evaluación de los resultados obtenidos acusó mejoras significativas en las restauraciones obtenidas con los métodos propuestos, con varias medidas de calidad apropiadas para cada tipo de problema.

Capítulo 1

Introducción

1.1. Procesamiento de señales sonoras

Los aparatos electrónicos con gran capacidad computacional son cada vez más comunes en nuestras vidas: desde teléfonos celulares, hasta relojes inteligentes y asistentes personales para el hogar. Aparejado a esto se ha incrementado la necesidad de una interacción hombre-máquina dinámica y ágil. En nuestra vida diaria, este tipo de interacción se da naturalmente con otras personas a través del habla, y en un intento de emular esto, mucho esfuerzo está actualmente avocado a posibilitar el establecimiento de una comunicación oral fluida con aparatos tecnológicos. Esta necesidad de mejora es inherente a una serie de temas en el área de procesamiento de señales, que incluye sistemas de traducción automático ([1]), reconocimiento de emociones y estados afectivos ([2]) y asistentes personales digitales ([3]), para nombrar sólo algunos, que requieren señales de habla como entrada.

Uno de los principales problemas dentro de este contexto tiene que ver con que en la vasta mayoría de las circunstancias, una señal capturada por un micrófono presentará distorsiones de varios tipos. El desafío consiste entonces en "limpiar" la señal de interés, y el problema asociado puede clasificarse como de *denoising*, dereverberación o separación de fuentes, según el tipo de distorsión presente en la grabación. Nos encontramos ante un problema de denoising cuando la grabación presenta, además de la señal de interés, componentes acústicas originadas en otras fuentes, en general difusas. Por ejemplo, el ruido del viento, de un motor funcionando o el murmullo en una habitación contigua. Lo que se entiende por "ruido" depende en cierta medida del contexto, pero en general dentro del procesamiento de señales sonoras hablaremos de denoising cuando el ruido no tenga relación con la fuente de la señal de interés.

El contexto de dereverberación está mucho más claramente delimitado. Cuando una señal de audio es grabada en un espacio cerrado, la grabación se ve afectada por las componentes reverberantes que se producen por la reflexión de las ondas sonoras en las paredes, el mobiliario, etcétera. Estos ecos pueden degradar significativamente las características de la señal obtenida (en particular cuando los micrófonos se encuentran apartados de la fuente, [4]), y resultan problemáticos sobre todo en el contexto de aplicaciones en procesamiento de habla ([5]).

Finalmente, hablaremos de separación de fuentes cuando en una grabación aparecen dos o más señales sonoras correspondientes a fuentes específicas. Por ejemplo, cuando disponemos de una grabación de una conversación entre varias personas cuyas voces se mezclan y queremos aislar la voz de algunas de ellas en particular (lo que se conoce como *cocktail party problem*, [6]), estamos hablando de un problema de separación de fuentes. Hasta cierto punto, esta definición de mezcla coincide con la de ruido (si consideramos a las señales secundarias como interferencia) y en general un problema se clasificará como de denoising o separación de fuentes según el enfoque adoptado.

Normalmente, los problemas mencionados se presentan acoplados. De una forma u otra, el ruido siempre está presente en cualquier grabación, por lo que un buen método de dereverberación o separación debe ser robusto con respecto al ruido. Además, si pensamos en separación de fuentes a partir de una grabación en una habitación cerrada, el problema de separación se mezcla con el de dereverberación.

Todos los problemas antes mencionados pueden clasificarse, según el contexto, como ciegos o supervisados. El problema se dice ciego cuando la única información disponible es la grabación en sí, y se dice supervisado si vamos a hacer uso de cierta información *a-priori*. Esta información típicamente consiste en aquella que permita inferir características de la habitación en problemas de dereverberación, o en características de las fuentes en problemas de separación.

Por otra parte, un problema de procesamiento de señales sonoras puede clasificarse en mono o multi-canal, dependiendo de la cantidad de micrófonos con que se hayan realizado las grabaciones. En general, decimos que un problema es subdeterminado si la cantidad de micrófonos es menor que la cantidad de fuentes, y determinado o sobredeterminado si hay, respectivamente, la misma o mayor cantidad de micrófonos que de fuentes.

Para atacar estos problemas, haremos uso de herramientas de factorización en matrices no negativas (NMF, [7]) y sus variantes. Si bien ya existen algunos enfoques basados en este tipo de técnicas ([8], [9]), presentan varios problemas, por lo general ligados a la falta de unicidad en las representaciones encontradas. Esto se traduce luego en alta sensibilidad a la inicialización, aparición de artefactos (ruidos artificiales indeseables) o limitación en la capacidad de dereverberación.

En trabajos anteriores ([10], [11]) hemos abordado el uso de métodos de regularización mixta en el contexto de problemas de procesamiento de imágenes. Este tipo de enfoque permite desfavorecer la aparición de ciertas características estructurales en una matriz asociada a una imagen reconstruida, mejorando así la calidad de los resultados. A partir de esta idea, haremos uso aquí de un enfoque análogo para encontrar representaciones NMF estables respecto a la inicialización, cuya estructura permita reproducir las características deseadas de las señales limpias.

A continuación, entraremos en más detalles en lo que respecta a los problemas de dereverberación y separación de fuentes. El problema de denoising será abordado dentro del contexto de esos otros dos problemas.

1.2. Dereverberación

El fenómeno de reverberación dentro de una habitación puede modelarse a partir de lo que se conoce como respuesta al impulso del cuarto (RIR, por sus siglas en inglés, *room impulse response*). Dadas una habitación determinada y las ubicaciones de una fuente y un micrófono, la RIR se define como la salida del sistema cuando la entrada es una delta de Dirac. Más informalmente, puede pensarse como la señal que sería grabada a partir de reproducir un pulso breve. En la Figura 1.1 puede observarse una RIR (simulada a partir del método de las imágenes de fuente, [12]) para una habitación de 5[m] × 4[m] × 6[m]. Las posiciones del micrófono y parlante son $(2[\text{m}], 1.5[\text{m}], 1[\text{m}])$ y $(2[\text{m}], 3.5[\text{m}], 2[\text{m}])$, respectivamente, y el *tiempo de reverberación*[1] de 450[ms]. Puede observarse que la RIR consta de un primer impulso de gran amplitud, que corresponde a la llegada de la señal sonora de forma directa, y luego picos más pequeños, cuyas amplitudes disminuyen con el tiempo, y que corresponden a los diferentes ecos que llegan al micrófono.

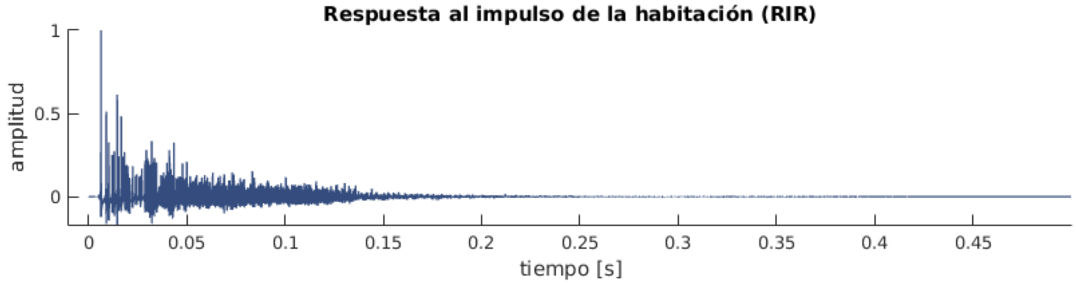


Figura 1.1: Señal de respuesta al impulso de una habitación simulada.

Ahora que hemos definido la respuesta al impulso, podemos proceder al modelado del fenómeno de reverberación.

Sean $s, h, x : \mathbb{R} \to \mathbb{R}$, con soporte en $[0, \infty)$, las funciones asociadas a una señal limpia, la respuesta al impulso del cuarto y la señal reverberante, respectivamente. El fenómeno de reverberación puede modelarse apropiadamente a través de un sistema lineal e invariante en el tiempo, que se representa a través de la ecuación

$$x(t) = (h * s)(t), \quad t \in \mathbb{R}, \tag{1.1}$$

donde "$*$" denota convolución. El uso de esta representación está fundado en dos hipótesis principales: la primera, que la fuente y el micrófono están fijos, en una habitación que no sufre cambios, y la segunda, que las componentes no lineales del fenómeno son lo bastante pequeñas como para poder ser despreciadas sin perder precisión. La segunda hipótesis suele cumplirse con frecuencia, mientras que la primera depende del contexto: en general, si la fuente corresponde a un hablante, la persona puede estar en movimiento mientras habla. Empero, en tanto el movimiento sea moderado, podremos suponer que la hipótesis se cumple por tramos, en el peor de los casos.

[1] Tiempo que tarda el sonido en "apagarse" en una habitación cerrada una vez que la fuente se ha detenido. Normalmente se considera el tiempo que tarda la presión sonora en caer 60dB.

Dado que en la práctica trabajaremos con señales discretas, escribimos la versión discreta del modelo (1.1), como

$$x[n] = \sum_{m=-\infty}^{\infty} s[n-m]\, h[m], \quad n \in \mathbb{Z}. \tag{1.2}$$

El efecto de la reverberación en la señal de audio puede visualizarse en la Figura 1.2, en donde se ilustran una señal de habla limpia y su versión reverberante, obtenida mediante una convolución discreta con la respuesta al impulso ilustrada en la Figura 1.1. Puede observarse que el fenómeno tiene efectos importantes distinguibles a simple vista: ciertas pausas o silencios presentes en la señal original han desaparecido, mientras que la morfología de algunos segmentos ha cambiado de manera significativa.

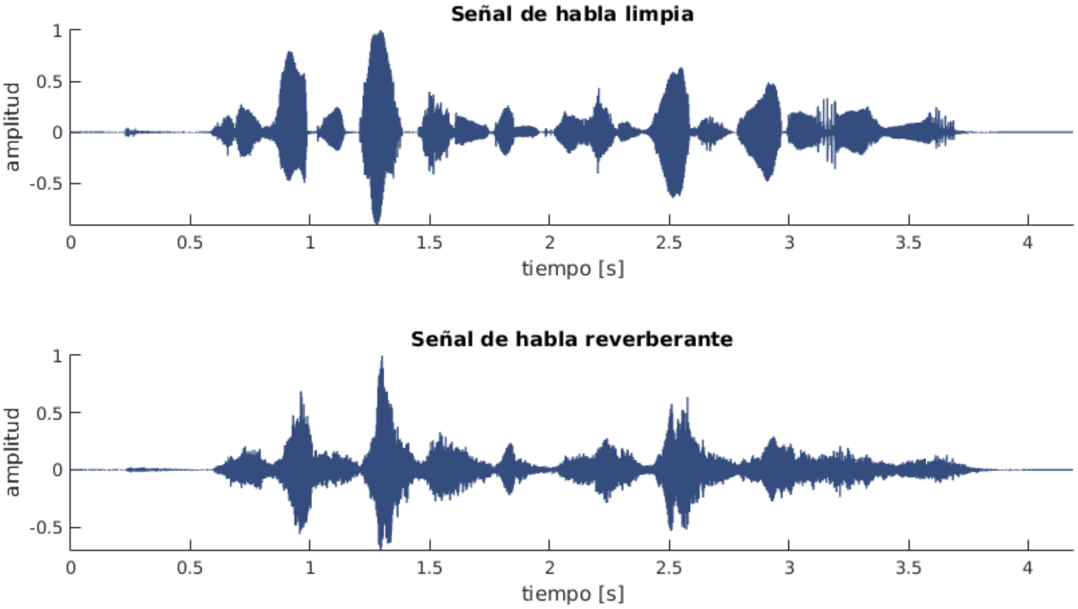


Figura 1.2: Señales de habla limpia y reverberante (convolución discreta con una respuesta al impulso).

A continuación, proveemos un marco teórico para el problema de separación de fuentes sonoras.

1.3. Separación de fuentes sonoras

En su versión más simplificada, el problema de mezcla puede modelarse como una combinación lineal aditiva de las señales correspondientes a cada fuente, desfasadas en el tiempo de acuerdo a las distancias entre las fuentes y los micrófonos. En la práctica, empero, es frecuente encontrar condiciones de reverberación y presencia de ruido.

Consideremos entonces un contexto con I fuentes sonoras (hablantes) y R micrófonos. Para poder modelar la mezcla reverberante, comenzamos por definir las funciones

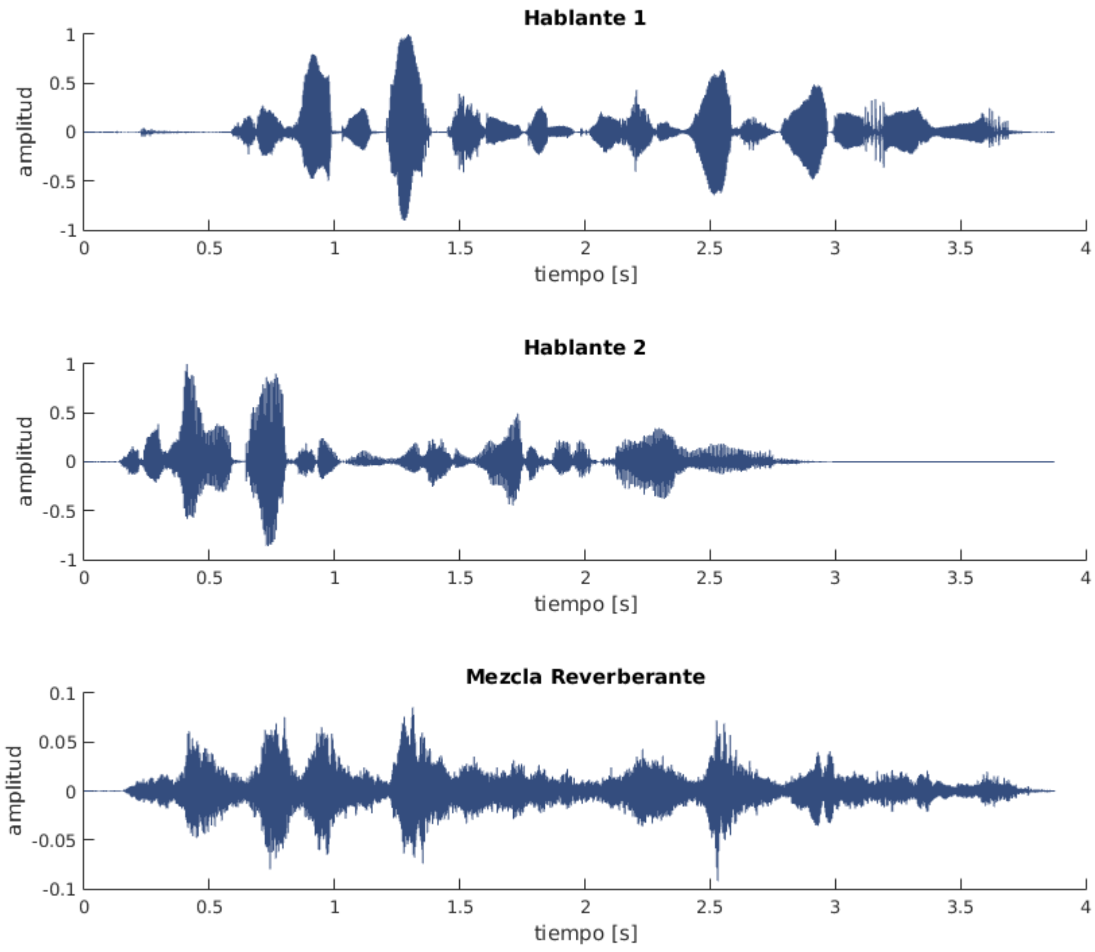


Figura 1.3: Ejemplo de dos señales limpias y el resultado de una mezcla reverberante.

a tiempo continuo y de soporte compacto $s_i, h_{r,i} : \mathbb{R} \to \mathbb{R},\ i = 1, \ldots, I,\ r = 1, \ldots R$, donde s_i es la señal proveniente de la i-ésima fuente y $h_{i,r}$ es la respuesta al impulso medida desde la i-ésima fuente al micrófono r. Entonces, bajo las mismas hipótesis de linealidad e invariancia temporal de la sección anterior, la señal x_r captada por el micrófono r satisface:

$$x_r(t) \doteq \sum_{i=1}^{I} (h_{r,i} * s_i)(t), \quad r = 1, \ldots, R. \tag{1.3}$$

En la Figura 1.3 puede verse cómo luce una mezcla reverberante de señales de habla. Es claro que la información de las señales limpias es indiscernible a simple vista a partir de la mezcla, al menos en el dominio temporal. No obstante, una representación en el dominio tiempo-frecuencia puede proveer otra perspectiva. Introducimos entonces a continuación el concepto de Transformada de Fourier de Tiempo Corto (STFT).

1.4. Representación en tiempo-frecuencia

Dado que las señales de audio en general no son estacionarias y a menudo presentan rápidas oscilaciones en su amplitud, suele ser conveniente utilizar una transformación apropiada para pasar del dominio temporal al dominio tiempo-frecuencia. Aquí nos centraremos en la transformación más utilizada, la STFT ([13]), que introducimos a continuación.

Sea $x : \mathbb{R} \to \mathbb{R}$ una función en $L^2(\mathbb{R})$. Se define la STFT de x en tiempo t y frecuencia k como

$$\mathbf{x}_k(t) \doteq \int_{-\infty}^{\infty} x(u)w(u-t)e^{-2\pi iuk}du, \quad t, k \in \mathbb{R},$$

donde $w : \mathbb{R} \to \mathbb{R}_0^+$ es una función par y de soporte compacto tal que $\|w\|_1 = 1$. Esta función se llama *ventana*.

En su versión discreta (que es lo que nos compete dado que nuestras señales están discretizadas), la STFT está definida como

$$\mathbf{x}_k[n] \doteq \sum_{m=-\infty}^{\infty} x[m]w[m-n]e^{-2\pi imk}, \quad n, k \in \mathbb{Z}. \tag{1.4}$$

La idea consiste entonces en desplazar la ventana w hasta cubrir todo el soporte de la señal de interés, obteniendo así el espectrograma asociado a la señal. Esto resulta en una matriz compleja cuyos elementos son las componentes tiempo-frecuencia de la señal. En la Figura 1.4 pueden verse los espectrogramas asociados a las señales introducidas en la Figura 1.3, en donde se ponen de manifiesto las diferencias entre la estructura espectral característica de ambos hablantes. Se grafica solamente el espectro para las frecuencias positivas, dada la simetría existente con las frecuencias negativas para toda señal a valores reales en el dominio temporal. La STFT fue calculada utilizando una ventana Hann ([14]) de 512 muestras, con un solapamiento de 256. Las amplitudes (al cuadrado) de los elementos de cada espectrograma se muestran en escala logarítmica.

En el dominio tiempo-frecuencia introducido mediante la STFT, ya no estamos lidiando con vectores asociados a las señales, sino con matrices. A continuación introducimos algunas herramientas que nos permitirán abordar más adelante los problemas de procesamiento de señales ya introducidos.

1.5. Factorización en matrices no negativas

Decimos que una matriz $Y \in \mathbb{R}^{K \times N}$ es no negativa cuando sus elementos $Y_{k,n} \geq 0$, $\forall k = 1, \ldots, K$, $n = 1, \ldots, N$. En la Figura 1.4 puede observarse cierta estructura que caracteriza a los espectrogramas de cada hablante. Es deseable entonces construir una representación que permita capturar el espaciado frecuencial de dicha estructura. Un claro ejemplo de este tipo de representación es la Factorización en Matrices Nonegativas (NMF, [7]): dada $Y \in \mathbb{R}_{0,+}^{K \times N}$, decimos que $W \in \mathbb{R}_{0,+}^{K \times J}$ y $U \in \mathbb{R}_{0,+}^{J \times N}$ (con $J << \text{máx}\{K, N\}$) conforman una NMF de Y si

$$Y \approx WU, \tag{1.5}$$

donde el sentido de esta aproximación puede depender del enfoque adpotado, y será especificado más adelante. En este contexto, W es llamada *diccionario* y U *matriz de coeficientes*. La idea es que Y puede representarse (aproximadamente) como una combinación lineal de los *átomos* del diccionario (columnas de W), que caracterizan la estructura de Y.

La ventaja de este tipo de representación es que Y se modela como una combinación lineal puramente aditiva. Es decir, todos los átomos son nonegativos, y contribuyen nula o positivamente a Y, y por lo tanto puede asociárseles cierta noción de "partes constitutivas".

Existen otros tipos de representaciones en matrices no negativas asociadas a otro tipo de modelos. Por ejemplo, si Y quiere aproximarse a partir de cierta noción de dependencia temporal en sus componentes, puede utilizarse un NMF de tipo convolutivo (CNMF, [15]). Esto es, dada una matriz $Y \in \mathbb{R}_{0,+}^{K\times N}$, se construyen dos matrices $S \in \mathbb{R}_{0,+}^{K\times N}$ y $H \in \mathbb{R}_{0,+}^{K\times M}$ (donde M está asociado a la duración del fenómeno temporal) de manera que

$$Y_{k,n} \approx \sum_m S_{k,n-m+1} H_{k,m}, \quad k = 1,\ldots,K,\ n = 1,\ldots,N. \tag{1.6}$$

La forma en que estas representaciones se construyen depende en primera instancia de cómo "medimos" el error de aproximación en las ecuaciones (1.5) y (1.6). Además, dependiendo del contexto del problema en que estemos trabajando, puede ser deseable forzar ciertas características estructurales no sobre la aproximación de Y, sino sobre los elementos que la componen. Esto puede hacerse, por ejemplo, mediante métodos de penalización (que anteriormente hemos utilizado en el contexto de procesamiento de imágenes: [10], [11]), los que introduciremos en el Capítulo 2 a través de un enfoque bayesiano.

1.6. Objetivo general

El objetivo general de esta tesis es el desarrollo de modelos basados en técnicas de NMF para abordar la resolución de los problemas de dereverberación y separación de fuentes sonoras. Este tipo de procesamiento es un paso necesario en la mayoría de las aplicaciones basadas en interacción hombre-máquina a través del audio, y debe ser lo suficientemente robusto para adaptarse a la variabilidad de las circunstancias en que este tipo de interacciones suelen darse en la vida cotidiana.

A continuación se detallan los objetivos específicos de la presente investigación.

1.7. Objetivos específicos

- Determinar las medidas de similitud y las restricciones a aplicar en algoritmos de NMF, que permitan obtener representaciones útiles para problemas de procesamiento de señales de audio: dereverberación y separación de fuentes sonoras.

- Desarrollar métodos que permitan resolver los problemas de optimización planteados en términos de las funciones de costo y restricciones asociadas.
- Desarrollar métodos para ajustar los parámetros que regulan el compromiso entre la fidelidad de la aproximación y los términos de penalización utilizados.
- Desarrollar algoritmos eficientes para resolver problemas de dereverberación y separación de fuentes sonoras en función de los modelos determinados.
- Evaluar los resultados obtenidos a partir de los métodos desarrollados en comparación con algoritmos del estado del arte para la misma tarea, mediante medidas objetivas de calidad y desempeño.

1.8. Organización de la tesis

En el capítulo siguiente se describirán los métodos propuestos en esta tesis, que incluyen el modelado de los problemas de dereverberación y separación de fuentes, la obtención de funciones de costo y desarrollo de algoritmos de optimización. En el Capítulo 3 se exponen los resultados obtenidos a partir de estos métodos y se los compara con aquellos obtenidos con métodos del estado del arte. En el Capítulo 4 se discuten los resultados, presentando conclusiones y trabajos futuros. Finalmente, en el Capítulo 5 se listan las publicaciones asociadas a este trabajo de tesis, que pueden encontrarse en el Anexo junto a una breve síntesis de las contribuciones correspondientes a cada una de ellas.

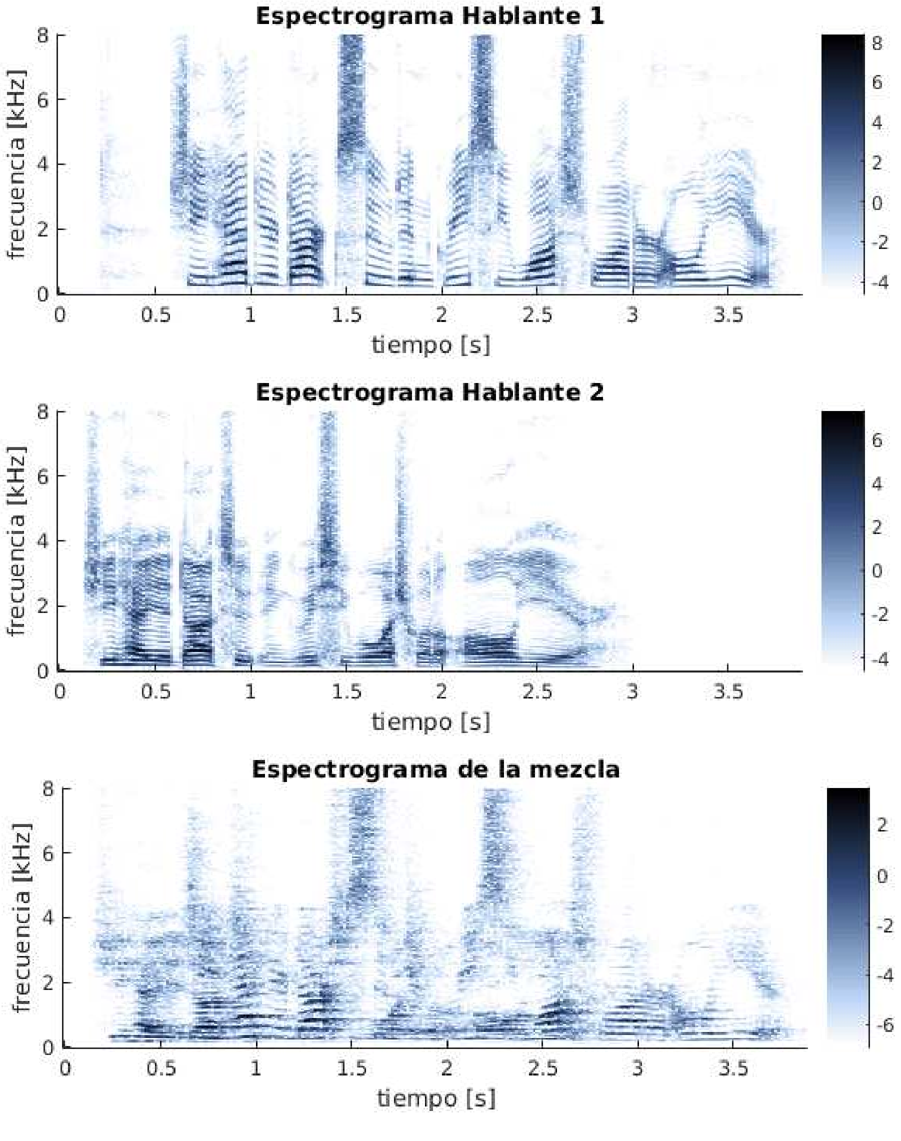


Figura 1.4: Espectrogramas limpios de dos hablantes y una mezcla reverberante artificial.

Capítulo 2

Métodos propuestos

2.1. Modelado

Siguiendo las ideas presentadas previamente, introducimos los modelos propuestos para la aplicación de métodos de NMF a los problemas de dereverberación y separación de fuentes sonoras.

2.1.1. Reverberación

Como hemos visto en la Sección 1.2, el modelo de reverberación en el dominio temporal puede expresarse como

$$x[n] = \sum_{m=-\infty}^{\infty} s[n-m]\, h[m], \tag{2.1}$$

donde x es una aproximación a la señal observada, s es la señal limpia y h la RIR. A partir de la definición (1.4) y la ecuación (2.1), podemos escribir el modelo en términos de la STFT como

$$\mathbf{x}_k[n] \approx \tilde{\mathbf{x}}_k[n] \doteq \sum_{m=0}^{M-1} \mathbf{s}_k[n-m]\, \mathbf{h}_k[m],$$

con $\mathbf{x}_k[n], \tilde{\mathbf{x}}_k[n], \mathbf{s}_k[n], \mathbf{h}_k[m] \in \mathbb{C}, k = 1, \ldots, K, n = 1, \ldots, N, m = 1, \ldots, M$. El parámetro M está asociado a la duración acotada del fenómeno de reverberación, que transforma la suma en (2.1) en una suma finita. Dado que los ángulos de fase son extremadamente sensibles a pequeñas variaciones en las condiciones de reverberación ([16]), es razonable tratarlos como realizaciones de variables aleatorias con distribución uniforme en $(-\pi, \pi]$, lo que deviene en el modelo

$$X_{k,n} = \sum_m S_{k,n-m+1} H_{k,m}, \tag{2.2}$$

donde $X_{k,n} \doteq E(|\mathbf{x}_k[n]|)^2, S_{k,n} \doteq |\mathbf{s}_k[n]|^2$ y $H_{k,m} \doteq |\mathbf{h}_k[m]|^2, k = 1, \ldots, K, n = 1, \ldots, N,$ $m = 1, \ldots, M$ (ver [17, Anexos]). Esto corresponde justamente al modelo de NMF convolutivo introducido en (1.6).

2.1.2. Dereverberación: enfoque bayesiano simple

Consideremos un espectrograma observado $Y \in \mathbb{R}_{0,+}^{K \times N}$ y una aproximación X, definida como en (2.2). Para modelar la fidelidad de esta representación, recurrimos a un enfoque probabilístico ([18, Anexos]), que consiste en suponer que X e Y son realizaciones de variables aleatorias $\mathcal{X}$ e $\mathcal{Y}$, respectivamente, donde

$$\mathcal{Y}_{k,n} = \mathcal{X}_{k,n} + \mathcal{E}_{k,n},$$

y $\mathcal{E} \in \mathbb{R}^{K \times N}$ es una matriz aleatoria que modela de manera conjunta el error de representación y el ruido. Dado que no tenemos información disponible acerca del error, supondremos $\forall k, n$, que $\mathcal{E}_{k,n}$ tiene distribución normal de media cero. Esto corresponde a suponer que $\mathcal{Y}_{k,n}$ tiene la siguiente distribución, condicionada a $X_{k,n}$:

$$\pi_{like}(Y_{k,n}|X_{k,n}) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(Y_{k,n} - X_{k,n})^2}{2\sigma^2}\right),$$

donde $\sigma > 0$ es un parámetro a estimar, que corresponde a la desviación estándar. Esta distribución suele ser llamada de verosimilitud o *likelihood*, pues está asociada a la probabilidad de observar los datos dado el modelo subyacente.

Es importante destacar que el hecho de que el dato Y esté bien representado por X no implica directamente que S sea una buena representación del espectrograma de la señal limpia. Esto se ilustra en el ejemplo de juguete en la Figura 2.1, donde el dato Y está representado exactamente por X, definida como en la ecuación (2.2), a partir de los pares $\{S^{(0)}, H^{(0)}\}$, $\{S^{(1)}, H^{(1)}\}$ o $\{S^{(2)}, H^{(2)}\}$. En analogía con el problema de reverberación, queremos recuperar el par $\{S^{(0)}, H^{(0)}\}$, pero es claro que a partir de la exactitud en la representación del dato no podemos establecer una preferencia por este par con respecto a los demás.

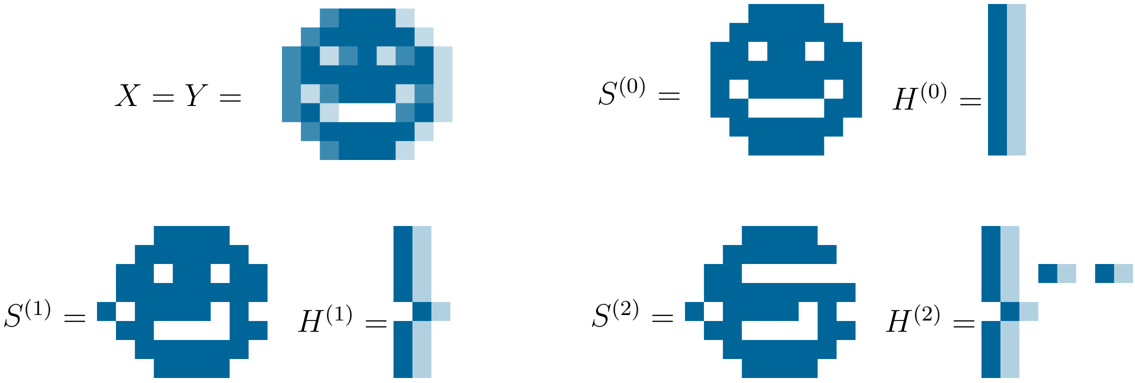


Figura 2.1: Ilustración conceptual de indeterminación en las representaciones.

La mayoría de los métodos del estado del arte ([8], [19]) buscan establecer una preferencia sobre la estructura de S a través de la imposición de raleza (predominancia de elementos nulos). Si bien cierto grado de raleza es una característica esperable en un espectrograma limpio, es claro que esto es insuficiente si observamos que $S^{(2)}$ tiene más elementos nulos que $S^{(0)}$ en la Figura 2.1.

Como un primer enfoque para solucionar esto, comenzaremos por imponer cierta estructura sobre la matriz H (además de sobre S) para evitar que exhiba las características observadas en $H^{(2)}$. Más adelante incorporaremos otro tipo de representación al modelo que ayudará a prevenir desplazamientos como el observado en $S^{(1)}$, valiéndonos de la estructura espectral de las señales de habla.

En primer lugar, esperamos que el espectrograma S de la señal limpia exhiba una estructura rala, lo que significa que una cantidad significativa de sus elementos son nulos, como puede verse en la Figura 1.4. Esta raleza puede favorecerse mediante la suposición de que los elementos de S son realizaciones de una variable aleatoria con distribución exponencial con media 2ν:

$$\pi_{prior}(S_{k,n}) = \frac{1}{2\nu} \exp\left(-\frac{1}{2\nu} S_{k,n}\right) \chi_{[0,\infty)}(S_{k,n}),$$

donde χ denota la función característica.

Esta distribución suele llamarse distribución *a priori*, o prior, pues es una suposición sobre las características de las variables aleatorias subyacentes, previa a la observación de los datos.

Finalmente, dado que los ecos de una RIR se apagan gradualmente, es esperable que el espectrograma H exhiba cierto grado de suavidad. Una manera sencilla de inducir suavidad en la estructura de H es a través de la hipótesis de que sus elementos son realizaciones de la siguiente distribución:

$$\pi_{prior}(H_{k,m}) = \frac{2}{\sqrt{2\pi}\eta} \exp\left(-\frac{H_{k,m}^2}{2\eta^2}\right) \chi_{[0,\infty)}(H_{k,n}).$$

Para poder obtener matrices cuyos elementos sean representativos de las funciones de distribución antes mencionadas, podemos hallar el argumento que maximiza la distribución a posteriori:

$$\pi_{post}(X|Y) \propto \pi_{like}(Y|X)\pi_{prior}(S)\pi_{prior}(H),$$

donde $\propto$ indica proporcionalidad con respecto a variables determinadas por los datos. Esta maximización se conoce como estimación *maximum-a-posteriori* (MAP), y equivale a hallar

$$\arg\min -\log \pi_{post}(X|Y) = \arg\min -\log[\pi_{like}(Y|X)\pi_{prior}(S)\pi_{prior}(H)],$$

donde X depende de S y H como en la ecuación (2.2).

Finalmente, bajo la hipótesis de que todas las variables aleatorias subyacentes son independientes, el problema consiste en minimizar

$$f_1(S,H) \doteq \sum_{k,n}(Y_{k,n} - X_{k,n})^2 + \frac{\sigma^2}{\nu}\sum_{k,n} S_{k,n} + \frac{\sigma^2}{\eta^2}\sum_{k,m} H_{k,m}^2, \tag{2.3}$$

sujeto a $S_{k,n}, H_{k,m} \geq 0 \; \forall k, m, n$.

El primer término de f_1 es lo que comúnmente se denomina *término de fidelidad*, que mide cuánto se parece la representación X al dato Y. Los demás términos corresponden a penalizantes sobre S y H, y son los que desfavorecerán ciertas características estructurales sobre S y H, propiciando otras.

Este primer enfoque, no obstante, presenta una desventaja. Los parámetros σ, λ y η de las distribuciones están asociados de alguna manera a la relación de tamaño entre los sumandos de f_1. Esta relación en general resulta variable dependiendo de la banda frecuencial, sobre todo en espectrogramas de señales de habla, donde las zonas de alta energía suelen concentrarse en las bajas frecuencias. Una manera de solucionar esto consiste en elegir ν y η dependientes de k (ver [20, Anexos]), con lo que la función f_1 se transforma ahora en

$$f_2(S, H) \doteq \sum_{k,n} (Y_{k,n} - X_{k,n})^2 + \sum_{k,n} \frac{\sigma^2}{\nu_k} S_{k,n} + \sum_{k,m} \frac{\sigma^2}{\eta_k^2} H_{k,m}^2. \tag{2.4}$$

Notar que agregar una dependencia de σ con respecto a k no agregaría ningún grado de libertad al modelo, y es por eso que se conserva independiente como en la función de costo anterior.

2.1.3. Dereverberación: enfoque bayesiano con priors correlacionadas

Hasta ahora hemos explorado el enfoque bayesiano más sencillo, que supone independencia entre todas las variables aleatorias subyacentes. No obstante, si bien la distribución de probabilidad *a-priori* asignada a H implica cierto grado de suavidad sobre la observación, no modela realmente el decaimiento temporal en H. Esto puede incorporarse al modelo mediante la suposición de una distribución de probabilidad no sobre H sino sobre su gradiente temporal ([17, Anexos]). Para hacer esto, consideramos una matriz $L \in \mathbb{R}^{M \times (M-1)}$ de diferencias finitas, y luego definimos $V \in \mathbb{R}^{K \times (M-1)}$ de modo que $V_{k,m} \doteq [HL]_{k,m} = H_{k,m+1} - H_{k,m}$. Ahora, suponiendo que los elementos de V son realizaciones de variables aleatorias con distribución normal de media cero y varianza η_k^2, tenemos

$$\pi_{prior}(V_{k,m}) = \frac{1}{\sqrt{2\pi}\, \eta_k} \exp\left(-\frac{V_{k,m}^2}{2\eta_k^2} \right).$$

Utilizando fórmula de Bayes y tomando el logaritmo de la distribución a posteriori, esto resulta en la función de costo

$$f_3(S, H) \doteq \sum_{k,n} (Y_{k,n} - X_{k,n})^2 + \sum_{k,n} \frac{\sigma^2}{\nu_k} S_{k,n} + \sum_{k} \frac{\sigma^2}{\eta_k^2} \|H_k L\|^2, \tag{2.5}$$

donde H_k denota la k-ésima fila de H.

Resulta oportuno mencionar que este tipo de enfoque permite modelar los parámetros de la función de costo a partir del uso de hyperparámetros. Esto es, podemos considerar a ν_k o $\eta_k, k = 1, \dots, K$, como realizaciones de variables aleatorias, e incorporar las distribuciones supuestas a la función de costo a través de la fórmula de Bayes. No obstante, hemos observado que el modelo propuesto es robusto con respecto a los parámetros, por lo que no ahondaremos aquí en detalles, que pueden encontrarse en [17, Anexos].

2.1.4. Dereverberación con NMF mixto

Aunque algunos enfoques iniciales basados en NMF, como los introducidos en [8], y otros más recientes, como el introducido en la sección anterior ([17, Anexos]), han demostrado producir resultados satisfactorios (como se verá en el Capítulo 3), no tienen en cuenta ninguna relación entre las componentes frecuenciales. De hecho, una hipótesis subyacente es que los elementos pertenecientes a distintas bandas de frecuencia en el espectrograma de una señal limpia no están correlacionados, lo que resulta ser una simplificación excesiva cuando se trata de señales de voz, dado que su estructura incluye componentes armónicas, múltiplos de la frecuencia fundamental de la voz (ver Figura 1.4). Esta estructura espectral puede incorporarse en un modelo de NMF convolutivo utilizando un enfoque basado en *diccionarios*, como se propone en [19]. Dicho enfoque ha mostrado un buen rendimiento dentro de un marco supervisado, pero no así en un entorno ciego. Esto tiene que ver con el hecho que hay demasiadas variables que deben aprenderse simultáneamente a partir de los escasos datos disponibles, lo que a menudo resulta en una buena representación del espectrograma reverberante con elementos que no permiten una buena representación del espectrograma limpio asociado.

Para explorar cómo solucionar esto, comenzamos por introducir el modelo mixto. Este consiste básicamente en reemplazar la matriz asociada al espectrograma limpio S, en el modelo de reverberación convolutivo (1.6) por una aproximación NMF, como la introducida en (1.5). Esto resulta en el modelo

$$X_{k,n} \doteq \sum_{m=1}^{M} \sum_{j=1}^{J} W_{k,j} U_{j,n-m+1} H_{k,m} \tag{2.6}$$

Ahora bien, para permitirle al modelo aprender todas estas variables, proponemos abordar el problema en dos etapas. La primera está destinada a, partiendo de los datos, construir un diccionario que permita una buena representación de la señal limpia, mientras que la segunda está avocada a usar dicho diccionario para obtener una representación adecuada del espectrograma reverberante. El principal desafío en este contexto consiste en que el diccionario debe ser aprendido sin tener acceso a la señal limpia, dado que toda la información disponible consiste en la señal reverberante.

Dado que X, definido como en (2.6) es una aproximación de la observación Y, necesitamos una forma de medir el error de representación. La correlación intrínseca entre las componentes frecuenciales mediante el modelo (2.6) permite entonces la utilización de una medida de fidelidad computable punto a punto sin perder esta relación. Como veremos en la Sección 2.3, esto representa una ventaja para el proceso de optimización.

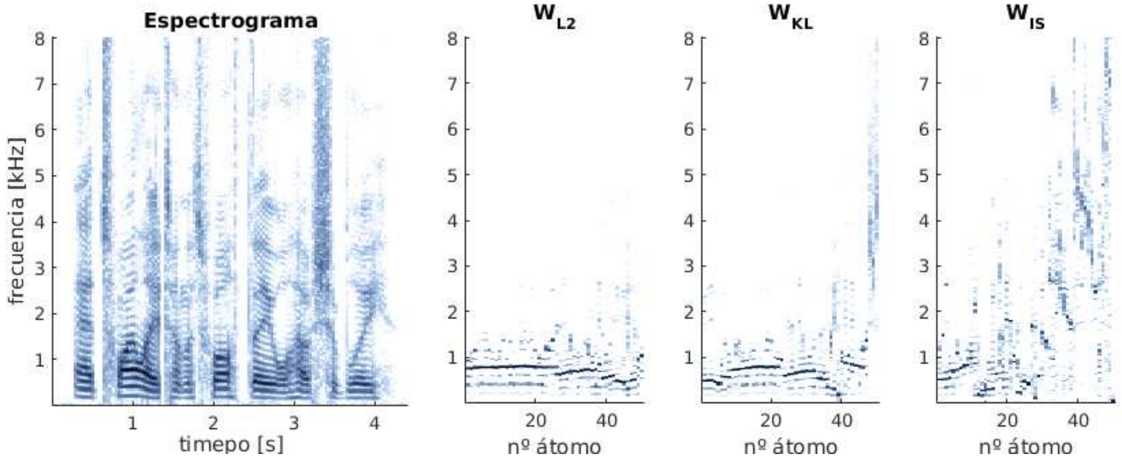


Figura 2.2: Diccionarios aprendidos con diferentes medidas de divergencia.

Introducimos entonces la divergencia β, definida en [21] como

$$D_\beta(Y||X) \doteq \sum_{k,n} \left(Y_{k,n} \frac{Y_{k,n}^{\beta-1} - X_{k,n}^{\beta-1}}{\beta(\beta-1)} + X_{k,n}^{\beta-1} \frac{X_{k,n} - Y_{k,n}}{\beta} \right), \quad \beta \in (0,2] \setminus \{1\}.$$

Una particularidad de esta manera de "medir" cómo X aproxima a Y tiene que ver con que, dependiendo del valor de β, se otorga más o menos importancia a las discrepancias en zonas de baja energía ([21]). Resulta oportuno señalar que esta medida de divergencia es una generalización de otras medidas más populares. De hecho, puede verse que $D_2(\cdot||\cdot)$ es igual a la mitad del cuadrado de la norma de Frobenius de $Y - X$, utilizada en la sección anterior, mientras que $D_\beta(\cdot||\cdot)$ aproxima la divergencia de Kullback-Leibler ([19]) cuando $\beta \to 1$ y la divergencia de Itakura-Saito ([22]) cuando $\beta \to 0^+$. En la Figura 2.2 se ilustra el resultado de aprender diccionarios utilizando las tres medidas de divergencia recién mencionadas a partir de un mismo espectrograma limpio. Rápidamente puede observarse que cuando el valor de β es pequeño (divergencia de Itakura-Saito, W_{IS}), las zonas de alta frecuencia cobran más relevancia que utilizando un valor de β cercano a 2 (norma L^2, W_{L2}).

La idea (desarrollada en profundidad en [23, Anexos]) consiste entonces en explotar esta particularidad para construir un diccionario óptimo a partir del espectrograma reverberante. Para ello, comenzamos por definir una función de costo utilizando la divergencia β, que partiendo de la misma idea desarrollada en la sección anterior toma la siguiente forma:

$$f_4(W,U,H) \doteq D_\beta(Y||X) + \sum_{j,n} \lambda_n^{(u)} U_{j,n} + \sum_k \lambda_k^{(h)} \|H_k L\|_2^2, \tag{2.7}$$

donde los parámetros $\lambda_n^{(u)}$ y $\lambda_k^{(h)}$ son análogos a los obtenidos en las secciones anteriores a partir de la estimación MAP. Notar que ahora la raleza no se fuerza sobre los elementos del espectrograma reconstruido, sino sobre la matriz de coeficientes U.

A partir de lo observado en la Figura 2.2 con respecto a la capacidad de la divergencia para aprender diccionarios en función del parámetro β, planteamos el siguiente procedimiento:

1. El primer paso es construir un diccionario W que permita una buena representación del espectrograma limpio, minimizando f_4 con respecto a W y U bajo ciertas condiciones. Esto es, fijar $H_{k,1} = 1 \; \forall k = 1, \ldots, K$, y $H_{k,m} = 0 \; \forall m > 1$, de manera de impedirle a H modelar la reverberación, tomar $\lambda_n^{(u)} = 0 \; \forall n = 1, \ldots, N$, y elegir $\beta = \beta_1$ apropiadamente.

2. Tomar $\lambda_n^{(u)} > 0$, $\lambda_k^{(h)} > 0$ y $\beta = \beta_2$ (elegidos como en [23, Anexos]). Minimizar f_4 con respecto a U y H, dejando fijo el diccionario W construido en el paso anterior. En este paso se busca entonces una buena aproximación del espectrograma reverberante utilizando un diccionario diseñado para obtener una buena representación del espectrograma limpio.

El proceso de cómo elegir β_1 no es trivial, y se explica en la Sección 3.2, mediante un experimento en el que simultáneamente se corrobora la ventaja de utilizar una divergencia variable.

2.1.5. Separación de fuentes

El modelo utilizado para mezcla de fuentes sonoras es análogo al utilizado para reverberación en (2.6), pues la mezcla de señales consiste en la suma de señales reverberantes. Si recordamos el modelo (1.3), la aditividad de la STFT (bajo la hipótesis de independencia estadística de las fuentes) deriva inmediatamente en el modelo no negativo:

$$X_{k,n,r} = \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{j=1}^{J} W_{k,j,i} U_{j,n-m+1,i} H_{k,m,i,r}, \tag{2.8}$$

donde $n = 1, \ldots, N$ y $k = 1, \ldots, K$ son las componentes correspondientes a tiempo y frecuencia, respectivamente, y $i = 1, \ldots, I$ y $r = 1, \ldots, R$ son los índices correspondientes a fuentes y micrófonos, también respectivamente. Nótese que en este caso se trata de un enfoque multicanal, donde se dispone de R micrófonos para realizar la captura del campo sonoro.

En este caso utilizaremos un enfoque supervisado para encarar el problema. Esto es, supondremos que tenemos datos de las fuentes que nos permitan construir un diccionario $W_i \in \mathbb{R}_{0,+}^{K \times J}$ característico para cada cada hablante $i = 1, \ldots, I$, concatenados en un tensor $W \in \mathbb{R}_{0,+}^{K \times J \times I}$. No nos centraremos en la forma de construir esta matriz a partir de los datos, pero en caso de contar con señales limpias, esto puede hacerse a partir de un método de NMF estándar ([7]), y en caso de que las señales provengan de entornos ruidosos o reverberantes, podemos recurrir al paso 1 del método presentado en [23, Anexos].

A partir de lo anterior, el problema de separación de fuentes queda planteado en términos de la minimización de la siguiente función de costo:

$$f_5(U, H) \doteq D_\beta(Y||X) + \sum_{j,n,i} \lambda_n^{(u)} U_{j,n,i} + \sum_{k,i,r} \lambda_k^{(h)} \|\underline{H}_{k,i,r} L\|_2^2, \tag{2.9}$$

donde $\underline{H}_{k,i,r}$ es el vector fila con componentes $H_{k,m,i,r}$, $m = 1, \ldots, M$. Notar que ahora la función de costo no depende de los diccionarios, que suponemos conocidos de antemano. El problema consiste entonces simplemente en minimizar f_5 con respecto a los tensores U y H.

Minimizar este tipo de funciones de costo no resulta un problema trivial, dada la alta dimensión de sus dominios y la no convexidad. En la siguiente sección presentamos los métodos optimización para la minimización de las funciones de costo asociadas a los problemas de dereverberación y separación.

2.2. Optimización

Una forma eficiente de minimizar las funciones de costo propuestas puede derivarse a partir de la técnica de la función auxiliar, que introducimos a continuación.

Sean $\Omega \subset \mathbb{R}^P$ y $f : \Omega \to \mathbb{R}_0^+$. Entonces, se dice que $g : \Omega \times \Omega \to \mathbb{R}_0^+$ constituye una *función auxiliar* para f si $g(\omega, \omega) = f(\omega)$ y $g(\omega, \omega') \geq f(\omega)$, $\forall \omega, \omega' \in \Omega$.

Si suponemos f y g como en la definición anterior, con $\omega^0 \in \Omega$ arbitrario y

$$\omega^t \doteq \arg\min_{\omega} g(\omega, \omega^{t-1}),\ t \in \mathbb{N} \tag{2.10}$$

entonces, bajo hipótesis de existencia y unicidad sobre este minimizante, se puede probar ([7]) que la sucesión $\{f(\omega^t)\}_{t\geq 1}$ es monótona no creciente. Notar que estas hipótesis no son muy restrictivas en la práctica, dado que la función g puede definirse de manera heurística. No obstante, esto no implica la convergencia de la sucesión $\{\omega^t\}_{t\in\mathbb{N}_0}$, tema que no ha sido abordado en la literatura para el tipo de funciones de costo que utilizaremos. A continuación demostramos un resultado que garantiza esta convergencia bajo ciertas condiciones sobre g.

Proposición 2.2.1. *Sean $\Omega \subset \mathbb{R}^P$, $f : \Omega \to \mathbb{R}_0^+$, y $g : \Omega \times \Omega \to \mathbb{R}_0^+$ una función auxiliar para f. Sea además $\{\omega^t\}_{t\in\mathbb{N}_0}$ definida como en (2.10), con ω^0 arbitrario. Entonces, si g es diferenciable y uniformemente fuertemente convexa en la primera variable, $\{\omega^t\}_{t\in\mathbb{N}_0}$ converge.*

Demostración. Recordemos que g es fuertemente convexa en la primera variable si $\forall \eta \in \Omega$, $\exists\, m_\eta > 0$ tal que

$$g(\omega', \eta) \geq g(\omega, \eta) + \nabla_1^T g(\omega, \eta)(\omega' - \omega) + \frac{m_\eta}{2}\|\omega' - \omega\|_2^2, \quad \forall \omega, \omega' \in \Omega,$$

donde ∇_1 denota el gradiente con respecto a la primera variable. La uniformidad en la convexidad fuerte implica que $\exists\, m > 0$ tal que $m < \text{ínf}_{\eta\in\Omega}\{m_\eta\}$. Entonces, la desigualdad anterior vale en particular para $\eta = \omega^{t-1}$. Si además tomamos $\omega = \omega^t$ y $\omega' = \omega^{t-1}$, tenemos

$$\|\omega^{t-1} - \omega^t\|_2^2 \leq \frac{2}{m}\left[g(\omega^{t-1}, \omega^{t-1}) - g(\omega^t, \omega^{t-1}) - \nabla_1^T g(\omega^t, \omega^{t-1})(\omega^{t-1} - \omega^t)\right].$$

Dado que ω^t es el minimizante de $g(\cdot, \omega^{t-1})$, se sigue que $\nabla_1^T g(\omega^t, \omega^{t-1}) = 0$, y como además $g(\omega^t, \omega^{t-1}) \geq f(\omega^t)$, tenemos

$$\begin{aligned} \|\omega^{t-1} - \omega^t\|_2^2 &\leq \frac{2}{m} \left[g(\omega^{t-1}, \omega^{t-1}) - g(\omega^t, \omega^{t-1}) \right] \\ &\leq \frac{2}{m} \left[f(\omega^{t-1}) - f(\omega^t) \right]. \end{aligned} \tag{2.11}$$

Notemos que

$$f(\omega^t) \leq g(\omega^t, \omega^{t-1}) \leq g(\omega^{t-1}, \omega^{t-1}) = f(\omega^{t-1}).$$

Puesto que $\{f(\omega^t)\}_{t \in \mathbb{N}_0}$ es monótona no creciente y $f(\omega) \geq 0$ $\forall \omega \in \Omega$, la sucesión converge y entonces es de Cauchy. Finalmente, la desigualdad (2.11) implica que $\{\omega^t\}_{t \in \mathbb{N}_0}$ también es de Cauchy y por lo tanto converge. □

La idea es formular una técnica de minimización basada en funciones auxiliares, de la siguiente manera: dada una función de costo f, construir una función auxiliar g con respecto a cada uno de sus argumentos individualmente, y luego minimizar cada g de forma iterativa para minimizar f.

La construcción de una función auxiliar debe realizarse de manera específica, acorde a la función de costo que se pretende minimizar. Funciones auxiliares para los funcionales de dereverberación f_1, f_2, f_3 y f_4, definidos respectivamente en (2.3), (2.4), (2.5) y (2.7) pueden encontrarse en [18, Anexos], [20, Anexos], [17, Anexos] y [23, Anexos]. Además, una función auxiliar para minimización del funcional de separación definido en (2.9) puede encontrarse en [24, Anexos].

En la próxima sección introducimos los algoritmos desarrollados para dereverberación y separación basados en la minimización de las funciones aquí propuestas.

2.3. Algoritmos

En primer lugar, introducimos un algoritmo de dereverberación mediante un enfoque bayesiano simple (Algoritmo 1). Es decir, un método para minimizar el funcional (2.4). Para evitar sobrecargar esta sección, evitamos especificar cómo se llevó a cabo la construcción del algoritmo, pero todos los detalles pueden consultarse en [20, Anexos]. Notar que este procedimiento es directamente aplicable a la minimización de (2.3), simplemente considerando los parámetros de f_2 independientes del índice de frecuencia k.

El Algoritmo 1 consta esencialmente de tres partes, comenzando con una inicialización en la que se lleva la señal del dominio temporal al dominio tiempo-frecuencia y se eligen apropiadamente los parámetros iniciales. Luego, un proceso iterativo reestima los parámetros del modelo alternativamente y de manera multiplicativa hasta un punto de convergencia. Este proceso multiplicativo tiene la ventaja de ser computacionalmente eficiente y conservar la nonegatividad de los elementos. Una vez que el proceso iterativo se estaciona, la etapa de reconstrucción retorna la señal del dominio tiempo-frecuencia al dominio temporal.

Algoritmo 1 Dereverberación: Enfoque Bayesiano Simple

```
Inicialización
$Y_{k,n} = |\mathrm{STFT}(y)_{k,n}|^2$, $\forall n, k$.
$H_{k,n} = \exp(1-n)$, $\forall n, k$.
$S_{k,n} = Y_{k,n}$, $\forall n, k$.
Minimización
while $\|S - S^{(-1)}\|_F^2 > \delta$ do        ▷ $S^{(-1)}$ es el resultado de la iteración previa
    $X_{k,n} \leftarrow \sum_m S_{k,n-m+1} H_{k,m}, \quad \forall k, n.$
    $S_{k,n} \leftarrow S_{k,n} \frac{\sum_m H_{k,m-n+1} Y_{k,n}}{\sum_m H_{k,m-n+1} X_{k,n} + \frac{\sigma^2}{\nu_k} p |S_{k,n}|^{p-1}}, \quad \forall k, n.$
    $H_{k,m} \leftarrow H_{k,m} \frac{\sum_n S_{k,n-m+1} Y_{k,n}}{\sum_n S_{k,n-m+1} X_{k,n} + \frac{\sigma^2}{\eta^2} H_{k,m}}, \quad \forall k, m.$
    $H_{k,m} \leftarrow H_{k,m} / \sum_m H_{k,m}, \quad \forall k, m.$
end while
Reconstrucción
$Z_{k,n} = \sqrt{\hat{S}_{k,n}} \arg(\mathrm{STFT}(y)_{k,n})$.
$\hat{s} = \mathrm{ISTFT}(Z)$.
```

El proceso de actualización de los elementos del modelo incluye un paso de normalización de H. Esto se hace para evitar la ambigüedad en la escala de los elementos (ver [20, Anexos] para más detalles). El mismo proceso se realiza en todos los algoritmos aquí presentado

A continuación, describimos el algoritmo desarrollado para la minimización de (2.5). Nuevamente, los detalles de cómo se derivan las ecuaciones en el Algoritmo 2 pueden encontrarse en [17, Anexos]. Si bien es similar al proceso anterior, la correlación temporal entre los elementos de la matriz H asociada a la respuesta al impulso no permite utilizar eficientemente un proceso iterativo para su actualización. Se requiere entonces la resolución de un sistema lineal para cada banda frecuencial. No obstante, el sistema es de orden pequeño (~ 30) y la forma en que lo hemos construido garantiza la unicidad y no negatividad en las soluciones.

Finalmente, en el Algoritmo 3 se detallan los pasos para minimizar (2.7). Este proceso parte de pasar la señal al dominio tiempo-frecuencia, y su estructura consta de dos etapas centrales. La primera de ellas está relacionada a la construcción de un diccionario W e involucra la elección de parámetros apropiados y un proceso iterativo multiplicativo mediante el cual se actualizan las matrices W y U hasta obtener convergencia. Una segunda etapa central incorpora elementos de penalización y está destinada a hallar matrices U y H que permitan una buena representación manteniendo fijo el diccionario W hallado en la etapa anterior. Aquí de nuevo la correlación en los elementos de H no permite un proceso de actualización multiplicativo, por lo que se adopta una estrategia

Algoritmo 2 Dereverberación: Enfoque Bayesiano Correlacionado

```
Inicialización
$Y_{k,n} = \left|\mathrm{STFT}(y)_{k,n}\right|^2$, $\forall n, k$.
$H_{k,n} = \exp(1-n)$, $\forall n, k$.
$S_{k,n} = Y_{k,n}$, $\forall n, k$.
Minimización
while $\|S - S^{(-1)}\|_F^2 > \delta$ do        ▷ $S^{(-1)}$ es el resultado de la iteración previa
    $X_{k,n} \leftarrow \sum_m S_{k,n-m+1} H_{k,m}, \quad \forall k, n.$
    $S_{k,n} \leftarrow S_{k,n} \frac{\sum_m H_{k,m-n+1} Y_{k,n}}{\sum_m H_{k,m-n+1} X_{k,n} + \frac{\sigma^2}{\nu_k} p |S_{k,n}|^{p-1}}, \quad \forall k, n.$

    for k = 1, . . . , K do
        Construir las matrices diagonales $A, B \in \mathbb{R}^{M \times M}$ :
            $A_{m,m} = \sum_n S_{k,n-m+1} X_{k,n},$
            $B_{m,m} = H_{k,m}.$
        Construir el vector $\zeta \in \mathbb{R}^M$ :
            $\zeta_m = \sum_n S_{k,n-m+1} Y_{k,n}$
        Resolver para $h$ el sistema lineal:
            $\left(A + \frac{\sigma_k^2}{\eta_k^2} B L^T L\right) h = B\zeta.$
        $H_{k,m} = h_m / \|h\|_1, \quad \forall m.$
    end for
end while
Reconstrucción
$Z_{k,n} = \sqrt{\hat{S}_{k,n}} \arg(\mathrm{STFT}(y)_{k,n}).$
$\hat{s} = \mathrm{ISTFT}(Z).$
```

similar a la utilizada en el Algoritmo 2. Finalmente, en la etapa de reconstrucción se hace uso de la representación obtenida como función de ganancia sobre el dato para disminuir la influencia del error de representación NMF sobre la calidad de la señal restaurada. Los detalles correspondientes a la construcción de este algoritmo pueden encontrarse en [23, Anexos].

Algoritmo 3 Dereverberación con NMF mixto

```
Preliminaries
Y_{k,n} = |STFT(y)_{k,n}|^2, ∀n, k.   ($Y_{k,n} = \left|\mathrm{STFT}(y)_{k,n}\right|^2,\ \forall n, k.$)
Etapa 1
Inicializar $\beta = \beta_1$, $\eta = 1/(2-\beta+(\beta-1)\chi_{\beta>1}(\beta))$ y $\lambda_n^{(u)} = 0,\ \forall n.$
Inicializar $H_{k,n} = 1$ si $n = 1$ y $H_{k,n} = 0, \forall n \geq 2, \forall k.$
Inicializar $W$ y $U$ de forma aleatoria (con $W_{k,j}, U_{j,n} > 0, \forall k, j, n.$)
while $\|W - W^{(-1)}\|_F^2 > \delta$ do        ▷ $W^{(-1)}$ es el resultado de la iteración previa
    $X_{k,n} \leftarrow \sum_m \sum_j W_{k,j} U_{j,n-m+1} H_{k,m},\ \ \forall k, n.$
    $W_{k,j} \leftarrow W_{k,j} \frac{\left[\left(\sum_{m,n} (X_{k,n})^{\beta-2} Y_{k,n} U_{j,m} H_{k,n-m+1}\right)^{\eta}\right]_{\epsilon}}{\left(\sum_{m,n} (X_{k,n})^{\beta-1} U_{j,m} H_{k,n-m+1}\right)^{\eta}},$
    $U_{j,m} \leftarrow U_{j,m} \frac{\left[\left(\sum_{k,n} (X_{k,n})^{\beta-2} Y_{k,n} W_{k,j} H_{k,n-m+1}\right)^{\eta}\right]_{\epsilon}}{\left(\sum_{k,n} (X_{k,n})^{\beta-1} W_{k,j} H_{k,n-m+1}\right)^{\eta}}.$
end while
▷ $[\,\cdot\,]_\epsilon \doteq \text{máx}\{\,\cdot\,,\epsilon\}$, donde $\epsilon \sim 10^{-10}$ es una operación que evita que los elementos de $W$ y $U$ tomen valores nulos.
```

Con respecto al problema de separación de fuentes, los pasos están detallados en el Algoritmo 4. Si bien [24, Anexos] provee una idea general de la derivación del proceso de minimización de (2.9), está planteado para el caso particular de $\beta = 2$. No obstante, los pasos utilizados para construir el Algoritmo 4 son análogos a aquellos introducidos en [23, Anexos]. Es oportuno recordar que en este contexto hemos supuesto el tensor de diccionarios W conocido, y por lo tanto no entramos en detalles en lo que respecta a su construcción.

```
Etapa 2
Inicializar $\beta = \beta_2$, $\eta = 1/(2 - \beta + (\beta - 1)\chi_{\beta>1}(\beta))$, $\lambda_n^{(u)}$, y $\lambda_k^{(h)}$ $\forall n, k$.
Inicializar $H_{k,n} = \exp(1 - n)$, $\forall n, k$.
while $\|WU - WU^{(-1)}\|_F^2 > \delta$ do        ▷ $U^{(-1)}$ es el resultado de la iteración previa
    $X_{k,n} \leftarrow \sum_m \sum_j W_{k,j} U_{j,n-m+1} H_{k,m}, \quad \forall k, n.$
    $U_{j,m} \leftarrow U_{j,m} \frac{\left[\left(\sum_{k,n} (X_{k,n})^{\beta-2} Y_{k,n} W_{k,j} H_{k,n-m+1} - \lambda_m^{(u)}\right)^{\eta}\right]_{\epsilon}}{\left(\sum_{k,n} (X_{k,n})^{\beta-1} W_{k,j} H_{k,n-m+1}\right)^{\eta}}..$

    for k = 1, ..., K do
        Construir la matriz diagonal $A \in \mathbb{R}^{M \times M}$ :
            $A_{m,m} = \sum_{n,j} W_{k,j} U_{j,n-m+1} X_{k,n}^{\alpha_1} / H_{k,m},$        ▷ $\alpha_1 = (\beta - 1)\chi_{\beta>1}(\beta)$
        Construir el vector $\zeta \in \mathbb{R}^M$ :
            $\zeta_m = \sum_{n,j} W_{k,j} U_{j,n-m+1} Y_{k,n} X_{k,n}^{\alpha_2},$        ▷ $\alpha_2 = (\beta - 2)\chi_{\beta\leq 2}(\beta)$
        Resolver para $h$ el sistema lineal:
            $\left(A + \frac{\sigma_k^2}{\eta_k^2} L^T L\right) h = \zeta.$
        $H_{k,m} = h_m / \|h\|_1, \quad \forall m.$
    end for
end while
Reconstrucción
Definir $G_{k,n} = \sum_j \hat{W}_{k,j} \hat{U}_{j,n} / \left(\sum_{j,m} \hat{W}_{k,j} \hat{U}_{j,n-m+1}, \hat{H}_{k,m}\right).$
Definir $Z \in \mathbb{C}^{K \times N}$ con $Z_{k,n} = \sqrt{G_{k,n} Y_{k,n}} \arg(\mathrm{STFT}(y)_{k,n}).$
$\hat{s} = \mathrm{ISTFT}(Z).$
```

Algoritmo 4 Separación NMF penalizado

```
Preliminaries
$Y_{k,n,r} = \left|\mathrm{STFT}(y_r)_{k,n}\right|^2$, $\forall n, k, r$.
Minimización
Inicializar $\beta = \beta_1$, $\eta = 1/(2 - \beta + (\beta - 1)\chi_{\beta>1}(\beta))$ y $\lambda_n^{(u)} = 0, \ \forall n$.
Inicializar $H_{k,n} = 1$ si $n = 1$ y $H_{k,n} = 0, \forall n \geq 2, \forall k$.
Inicializar $W$ y $U$ de forma aleatoria.
while $\|WU - WU^{(-1)}\|_F^2 > \delta$ do        ▷ $U^{(-1)}$ es el resultado de la iteración previa
    $X_{k,n,r} \leftarrow \sum_i \sum_m \sum_j W_{k,j,i} U_{j,n-m+1,i} H_{k,m,i,r}, \quad \forall k, n, r.$
    $U_{j,m,i} \leftarrow U_{j,m,i} \frac{\left[\left(\sum_{k,n,r} (X_{k,n,r})^{\beta-2} Y_{k,n,r} W_{k,j,i} H_{k,n-m+1,i,r}\right)^{\eta}\right]_{\epsilon}}{\left(\sum_{k,n,r} (X_{k,n,r})^{\beta-1} W_{k,j,i} H_{k,n-m+1,i,r}\right)^{\eta}}.$

    for k = 1, ..., K, i = 1, ..., I, r = 1, ..., R do
        Construir la matriz diagonal $A \in \mathbb{R}^{M \times M}$ :
            $A_{m,m} = \sum_{n,j} W_{k,j,i} U_{j,n-m+1,i} X_{k,n,r}^{\alpha_1} / H_{k,m,i,r},$        ▷ $\alpha_1 = (\beta - 1)\chi_{\beta>1}(\beta)$
        Construir el vector $\zeta \in \mathbb{R}^M$ :
            $\zeta_m = \sum_{n,j} W_{k,j,i} U_{j,n-m+1,i} Y_{k,n,r} X_{k,n,r}^{\alpha_2},$        ▷ $\alpha_2 = (\beta - 2)\chi_{\beta\leq 2}(\beta)$
        Resolver para $h$ el sistema lineal:
            $\left(A + \frac{\sigma_k^2}{\eta_k^2} L^T L\right) h = \zeta.$
        $H_{k,m,i,r} = h_m / \|h\|_1, \quad \forall m.$
    end for
end while
Reconstrucción
Definir $Z \in \mathbb{C}^{K \times N \times I}$ con $Z_{k,n,i} = \sqrt{\sum_j W_{k,j,i} U_{j,n,i}} \arg\left(\mathrm{STFT}(y)_{k,n}\right)$.
$\hat{s} = \mathrm{ISTFT}(Z)$.
```

Capítulo 3

Resultados

Antes de proceder a poner a prueba los algoritmos desarrollados en la sección anterior, haremos una breve introducción respecto a cómo medir la calidad de los procesos de dereverberación o de separación.

3.1. Evaluación de calidad

La evaluación de la calidad de una señal de audio restaurada no es una cuestión trivial. Dados dos vectores $s_1, s_2 \in \mathbb{R}^N$ asociados a dos señales de audio discretizadas, podríamos pensar en compararlos, por ejemplo, calculando $\|s_1 - s_2\|_2$. Supongamos, no obstante, que s_2 es una versión desplazada una muestra de s_1 (*i.e.* $s_1[n] = s_2[n-1] \, \forall n = 2, \ldots, N.$). Entonces, dadas las rápidas oscilaciones que presentan las señales de audio con respecto a la frecuencia de muestreo, $\|s_1 - s_2\|_2$ ostentará un valor muy grande, cuando en realidad las señales son casi iguales a los efectos prácticos.

Por otra parte, cuando tratamos con una señal de habla, la inteligibilidad suele ser preponderante. Es decir, lo importante no es tanto la representación exacta de la forma de onda de la señal, sino que la persona (o máquina) que la escucha pueda entender qué es lo que se está diciendo. Es decir, que se preserve adecuadamente la información contenida en la señal.

Teniendo esto en cuenta, para la evaluación de la calidad de los resultados se seleccionaron algunas medidas propuestas por diferentes autores.

3.1.1. Calidad de un proceso de dereverberación

Para evaluar la calidad de la restauración de una señal reverberante, se seleccionaron tres medidas, de acuerdo a los criterios establecidos en [25] y [26]:

- Relación Señal Ruido segmental pesada por frecuencia (fwsSNR): esta es una medida basada en la diferencia entre los espectrogramas de las señales a comparar, donde el peso que se le da a la diferencia entre los elementos depende de la banda frecuencial. Al medirse sobre los espectrogramas no se tiene en cuenta la información de fase local, que está relacionada a pequeños retardos temporales entre las señales. De esta forma, se le da más importancia a la preservación de la información relacionada al contenido frecuencial, con pesos específicos relacionados a la importancia de cada banda frecuencial en la percepción humana (ver [25]).

- Distancia Cepstral: es básicamente la norma euclídea de la diferencia entre los vectores de coeficientes cepstrales asociados a las señales. Esta medida resulta de importancia porque varios métodos de reconocimiento de habla están basados en las características de una señal en el dominio cepstral (ver [25]).
- Relación Reverberación-Energía de Modulación (SRMR): es una medida a través de la cual se cuantifica el grado de reverberación remanente en la señal luego del procesamiento. Esta medida es no intrusiva, en el sentido de que se toma directamente sobre la señal a medir sin tener en cuenta la señal limpia asociada. Si bien esto es una ventaja desde el punto de vista de la aplicabilidad, debe tenerse presente que esta medida no provee una noción de la preservación de la información de habla. Es decir, esta medida nos dice cuan reverberante "parece" ser la señal obtenida, pero no nos dice si contiene la misma información que la señal limpia original, debiendo considerarse siempre en conjunto con medidas que den una idea de la preservación de la información (ver [26]).

3.1.2. Calidad de separación

Por otra parte, para evaluar la calidad de una separación de una señal reverberante, se utilizan otras medidas de desempeño, definidas en [27]. Estas medidas están basadas en una descomposición de la señal restaurada en cuatro componentes:

$$\hat{s} = s + e_{interf} + e_{noise} + e_{artif},$$

en donde e_{interf} es un término asociado a la interferencia entre señales, e_{noise} es una componente asociada al ruido y e_{artif} a los artefactos que aparecen, por ejemplo, a raíz de las aproximaciones de modelado inherentes al proceso de separación. A partir de esto, se pueden definir tres medidas de calidad, que son análogas a la definición usual de la relación señal-ruido, pero donde se consideran las distintas componentes del ruido:

- Relación Señal-Distorsión (SDR).
- Relación Señal-Interferencia (SIR).
- Relación Señal-Artefactos (SAR).

Con estas medidas de calidad, podemos ahora evaluar el desempeño de los algoritmos para dereverberación y separación aquí desarrollados.

3.2. Dereverberación: experimentos

En esta sección presentamos algunos experimentos numéricos que dan cuenta del funcionamiento de los algoritmos propuestos en la sección anterior.

Hemos elegido dos configuraciones diferentes para los experimentos de validación. El primer experimento fue realizado mediante simulaciones para poder disponer de un gran número de ensayos, y el segundo mediante grabaciones para analizar la aplicabilidad de los métodos propuestos en condiciones reales.

Con el fin de evaluar el rendimiento de los algoritmos propuestos a lo largo de este trabajo, se realizaron comparaciones con dos métodos del estado del arte que se aplican en las mismas condiciones del problema: ciego y monocanal. El primer método elegido es el introducido por Wisdom *et al* en [28], a razón de su excelente desempeño en el Reverb Challenge 2016 ([29]). El segundo, presentado en [19] por Mohammadiha *et al*, constituye un primer intento de incorporar una representación basada en diccionarios a un enfoque convolutivo de NMF.

Antes de mostrar los resultados globales, presentamos algunos resultados que servirán para ilustrar el funcionamiento de los métodos introducidos. En primer lugar, explicamos brevemente cómo afecta el cambio de parámetros al método de reverberación NMF mixto de dos etapas (Algoritmo 3). En segundo lugar, ilustramos con un ejemplo el funcionamiento de los principales algoritmos propuestos, en comparación con los del estado del arte.

Cambio de divergencias en NMF mixto

En lo sucesivo, presentamos un experimento que da cuenta de la ventaja del cambio de divergencia que constituye la base del método de dereverberación con NMF mixto (Algoritmo 3), al tiempo que construimos un método para optimizar el parámetro de divergencia en la Etapa 1. Los detalles pueden encontrarse en [23, Anexos].

Para poder evaluar si un cierto parámetro β_1 es adecuado para construir un diccionario, tomamos un espectrograma reverberante Y, construimos un diccionario $W^{(\beta_1)}$ minimizando $D_{\beta_1}(Y||WU)$, y luego procedemos a corroborar cuán bien puede $W^{(\beta_1)}$ representar el espectrograma limpio S correspondiente. Para corroborar esto, minimizamos $D_{\beta^*}(S||W^{(\beta_1)}U)$ con respecto a U y calculamos la distancia cepstral entre S y su estimación. Es importante señalar en este punto que en este segundo paso, β^* no necesariamente es igual a β_1, por lo que nos proponemos hallar un par (β_1, β^*) óptimo, repitiendo este proceso sobre una grilla bidimensional de posibles valores para β_1 y β^*.

Resumiendo, dados un espectrograma reverberante Y y cada par admisible (β_1, β^*), hacemos lo siguiente:

1. Construir un diccionario $W^{(\beta_1)}$ minimizando $D_{\beta_1}(Y||WU)$ con respecto a W y U.

2. Utilizar $W^{(\beta_1)}$ para hallar una representación $\hat{S} = W^{(\beta_1)}\hat{U}$ del espectrograma limpio asociado S, donde $\hat{U} = \arg\min_U D_{\beta^*}(S||W^{(\beta_1)}U)$.

3. Comprobar la fidelidad de la representación $\hat{S}$ calculando la distancia cepstral con respecto a S.

Para llevar a cabo este experimento, tomamos cinco señales de la base de datos TIMIT y las hicimos reverberantes mediante convolución discreta con tres RIRs diferentes (450[ms], 600[ms] y 750[ms]). Los resultados obtenidos se muestran en la Figura

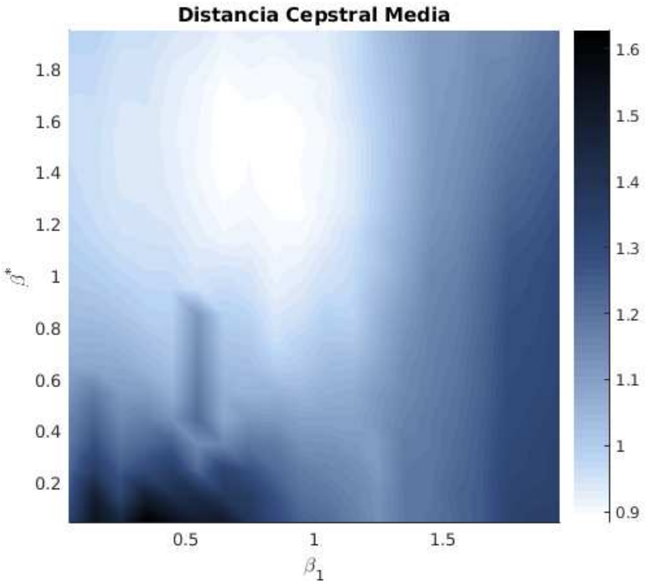


Figura 3.1: Distancia cepstral media como función de un parámetro de aprendizaje β_1 y uno de representación β^*.

3.1, que ilustra la distancia cepstral media en función de los parámetros β_1 y β^*. El minimizante se alcanza al rededor de $(0.75, 1.45)$, mostrando que $\beta_1 \approx 0.75$ resulta la mejor elección para la Etapa 1 del Algoritmo 3. Notar que esto no necesariamente significa que $\beta_2 \approx 1.45$ es la mejor elección para la segunda etapa, dado que aquí estamos minimizando $D_\beta(S||\hat{S})$, mientras que el segundo paso del algoritmo requiere minimizar la función definida en (2.7) con respecto a U y H.

En la Figura 3.1 puede observarse que los valores (β_1, β^*) están lejos de la diagonal, lo cual corrobora la ventaja de utilizar parámetros diferentes para las etapas de aprendizaje del diccionario y de representación. Esto tiene que ver con que valores pequeños de β_1 producen diccionarios que tienen en cuenta todo el espectro frecuencial, mientras que valores altos de β^* promueven fidelidad en las zonas de alta energía del espectrograma representado.

Ilustración del desempeño

Antes de presentar los resultados de las medidas de desempeño, ilustramos con un ejemplo el funcionamiento de algunos de los métodos. En la Figura 3.2 puede observarse un espectrograma limpio con su correspondiente versión reverberante, grabada en una de nuestras oficinas (oficina 1 en la Tabla 3.3). Más abajo, se muestran los espectrogramas obtenidos a partir de cuatro métodos de dereverberación diferentes.

Puede observarse que el método Wisdom reconstruye muy bien las zonas de alta energía, pero se observa un deterioro en las zonas de alta frecuencia y la correlación temporal no está bien marcada. Por otra parte, el método de Mohammadiha reconstruye mejor las zonas de alta frecuencia, pero la dereverberación es pobre y se generan artefactos. El método bayesiano correlacionado (Algoritmo 2) funciona de manera similar al Wisdom, y aunque introduce leves distorsiones debido a la falta de correlación

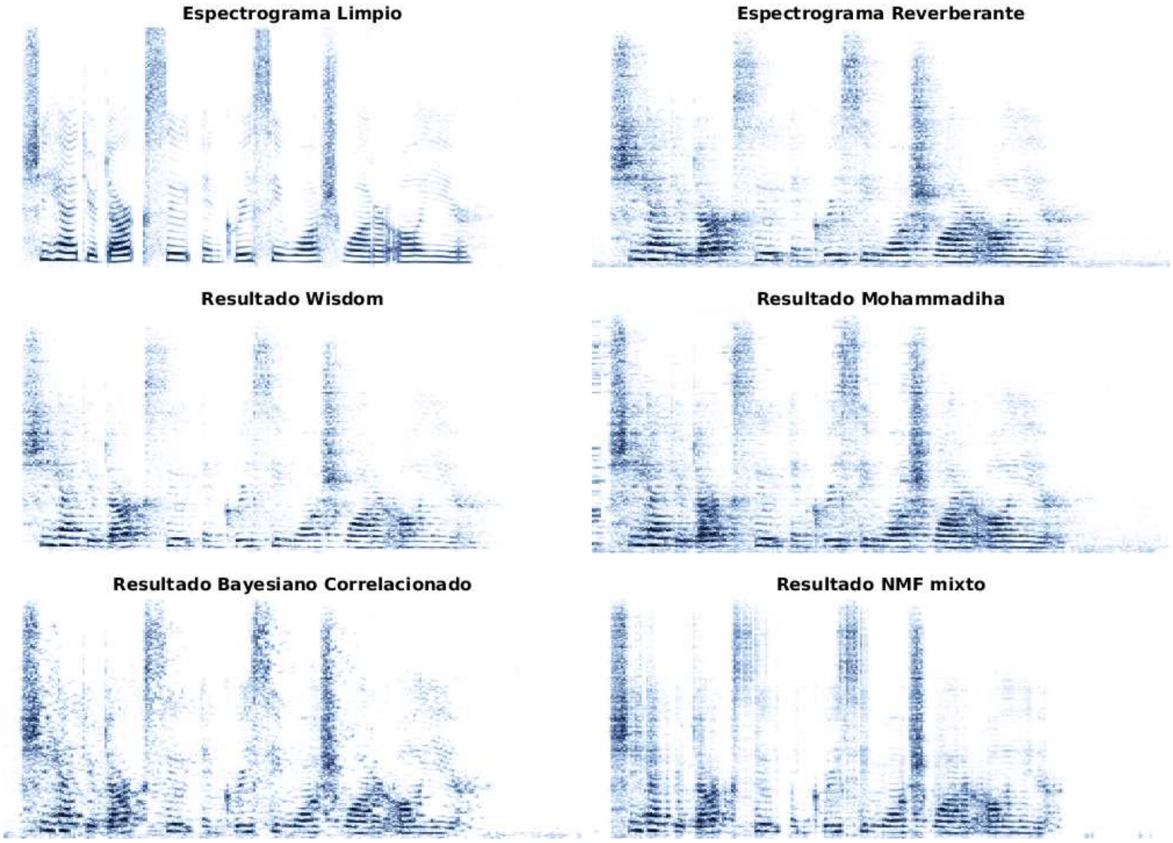


Figura 3.2: Ilustración del funcionamiento de los distintos métodos de dereverberación.

frecuencial, las frecuencias altas están mejor definidas. Finalmente, el método NMF mixto logra capturar muy bien las relaciones temporales entre componentes frecuenciales gracias a la representación NMF, mientras que la penalización permite un alto nivel de dereverberación con mínimas distorsiones.

3.2.1. Experimentos con simulaciones

Para los experimentos, tomamos 110 señales de habla de la base de datos TIMIT ([30]), grabadas a 16 kHz, y artificialmente las hicimos reverberantes mediante convolución con respuestas al impulso generadas con el software Room Impulse Response Generator[1], basado en el modelo presentado en [12]. Cada señal fue degradada bajo diferentes condiciones de reverberación: tres tamaños de habitación diferentes, cada uno con tres posiciones de micrófonos y tres tiempos de reverberación diferentes, lo que nos da un total de 2970 señales para las pruebas. La Tabla 3.1 da cuenta de las dimensiones de las salas y de las posiciones de las fuentes y los micrófonos elegidas.

En la Tabla 3.2 se muestran los resultados obtenidos con cada medida de rendimiento y cada uno de los métodos, ilustrados a su vez en la Figura 3.3. Es oportuno recordar en esta instancia que valores altos del fwsSNR y SRMR se asocian a mejor calidad en

[1] https://github.com/ehabets/RIR-Generator

	Largo	Ancho	Altura
Dimensiones habitación 1	5.00 [m]	4.00 [m]	6.00 [m]
Dimensiones habitación 2	4.00 [m]	4.00 [m]	3.00 [m]
Dimensiones habitación 3	10.0 [m]	4.00 [m]	5.00 [m]
Posición fuente	2.00 [m]	3.50 [m]	2.00 [m]
Posición micrófono 1	2.00 [m]	1.50 [m]	1.00 [m]
Posición micrófono 2	2.00 [m]	2.00 [m]	1.00 [m]
Posición micrófono 3	2.00 [m]	2.00 [m]	2.00 [m]

Tabla 3.1: Características de las habitaciones simuladas.

Método \ Medida	fwsSNR	distancia cepstral	SRMR
Señal reverberante	5.377 (1.70)	5.308 (0.61)	2.470 (1.01)
Wisdom	5.593 (1.67)	5.279 (0.60)	2.898 (1.14)
Mohammadiha	4.431 (1.48)	5.172 (0.78)	3.627 (1.00)
Bayesiano simple	5.987 (1.54)	5.035 (0.52)	4.062 (1.29)
Bayesiano correlacionado	7.604 (1.60)	4.614 (0.52)	4.423 (1.48)
NMF mixto	8.153 (1.51)	4.573 (0.48)	3.751 (1.21)

Tabla 3.2: Media y desviación estándar de desempeño con cada método en simulaciones.

la reconstrucción, mientras que con la distancia cepstral ocurre lo opuesto.

Puede observarse que el método NMF mixto (Algoritmo 3) presenta el mejor desempeño en cuanto a fwsSNR y distancia cepstral, seguido del método bayesiano correlacionado (Algoritmo 2). Por otra parte, en cuanto al SRMR, los métodos bayesianos (Algoritmos 1 y 2) presentan un mejor desempeño, pero es oportuno recordar que siendo el SRMR una medida no intrusiva, no nos da una idea de qué tanto se parece la señal restaurada a la señal limpia asociada.

3.2.2. Experimentos con grabaciones

Para comprobar si nuestro método funciona en situaciones de la vida real, realizamos grabaciones en dos de nuestras propias oficinas, durante horas normales de trabajo y con el acondicionador de aire y las computadoras encendidas. Las dimensiones de las oficinas se muestran en la Tabla 3.3, junto con las posiciones de los parlantes y micrófonos. Los tiempos de reverberación de las salas resultaron ser de 460[ms] en la Oficina 1 y de 440[ms] en la Oficina 2, medidos utilizando barridos sinusoidales ([31]). Se seleccionaron al azar cuatro hablantes (dos hombres y dos mujeres) de la base de datos TIMIT, y se grabaron 10 señales de voz de cada uno de ellos en cada habitación, con una frecuencia de muestreo de 16[kHz].

Como es habitual, las señales de habla limpias tenían filtradas sus componentes de baja frecuencia. Por lo tanto, preprocesamos nuestras grabaciones reverberantes usando un filtro pasa-alto FIR de tamaño 5000, con una frecuencia de corte de 30[Hz] para mitigar el ruido de baja frecuencia. Para que las comparaciones fueran justas, todos los métodos fueron puestos a prueba después de este pre-procesamiento.

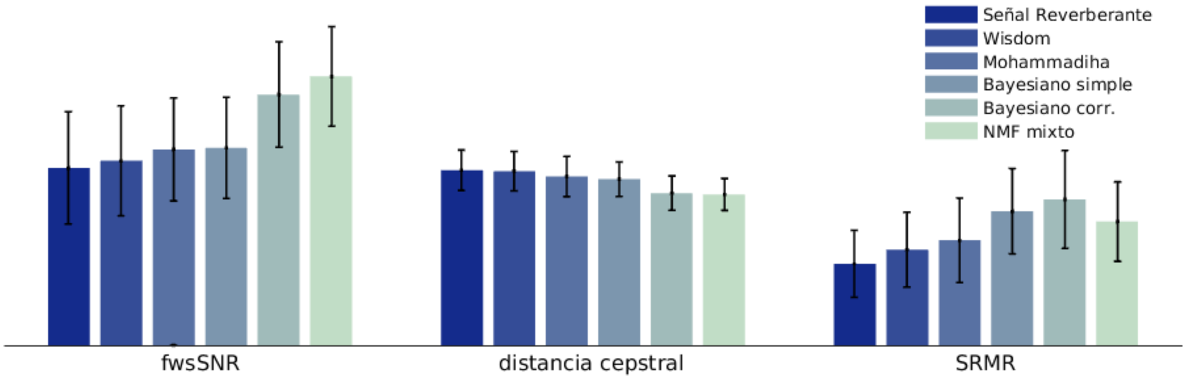


Figura 3.3: Visualización del desempeño obtenido con cada método en simulaciones.

	Largo	Ancho	Altura
Dimensiones oficina 1	4.15 [m]	3.00 [m]	3.00 [m]
Posición fuente 1	3.60 [m]	1.50 [m]	1.50 [m]
Posición micrófono 1	1.10 [m]	1.50 [m]	1.50 [m]
Dimensiones oficina 2	5.85 [m]	4.55 [m]	3.00 [m]
Posición fuente 2	1.10 [m]	1.50 [m]	1.50 [m]
Posición micrófono 2	1.10 [m]	4.00 [m]	1.50 [m]

Tabla 3.3: Características de oficinas

Los resultados obtenidos con cada medida de rendimiento y cada uno de los métodos se muestran en la Tabla 3.4 y están ilustrados en la Figura 3.4. Al igual que con los datos simulados, puede observarse que el método NMF mixto (Algoritmo 3) presenta el mejor desempeño en cuanto a fwsSNR y distancia cepstral, aunque la diferencia en distancia cepstral con el método de Wisdom no es estadísticamente significativa ($p > 0.01$). Por otra parte, en cuanto al SRMR, los métodos bayesianos (Algoritmos 1 y 2) mostraron nuevamente el mejor desempeño.

3.3. Separación: experimentos

Para poder evaluar las ventajas de la utilización de un enfoque bayesiano o de penalización sobre un modelo de mezcla no negativo, hemos optado por realizar comparaciones con un enfoque que está basado en una representación similar.

Murata *et al* proponen en [9] una generalización del modelo (2.8) que tiene en cuenta correlaciones temporales en los diccionarios de los hablantes y posibles desplazamientos de las fuentes. El problema con este tipo de enfoque es que demanda al modelo la representación de los espectrogramas observados mediante una cantidad muy grande de parámetros, que deben ser aprendidos simultáneamente. Para que este modelo funcione razonablemente bien, hemos elegido el caso particular en que las fuentes son estáticas, como lo son en nuestro caso.

Antes de mostrar los valores de desempeño obtenidos, ilustramos el funcionamiento

Método \ Medida	fwsSNR	distancia cepstral	SRMR
Señal reverberante	3.613 (1.52)	4.994 (0.56)	2.756 (0.75)
Wisdom	4.919 (1.37)	4.575 (0.43)	3.224 (0.77)
Mohammadiha	4.431 (1.48)	5.172 (0.78)	3.627 (1.00)
Bayesiano simple	5.486 (1.28)	4.807 (0.48)	4.487 (1.00)
Bayesiano correlacionado	6.526 (1.28)	4.827 (0.57)	4.859 (1.12)
NMF mixto	6.678 (1.18)	4.524 (0.53)	4.036 (0.84)

Tabla 3.4: Media y desviación estándar de desempeño con cada método en grabaciones.

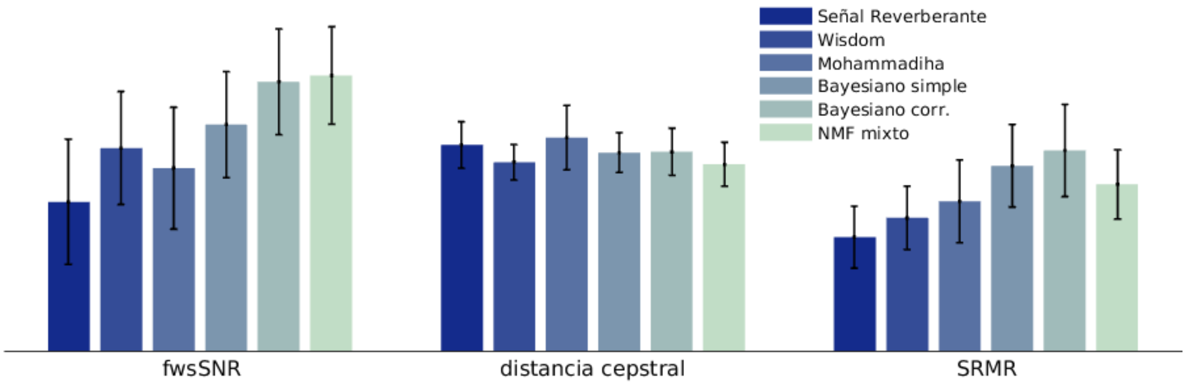


Figura 3.4: Visualización del desempeño obtenido con cada método en grabaciones.

de los métodos mediante un ejemplo. En la Figura 3.5 pueden verse los espectrogramas correspondientes a señales de habla limpia correspondientes a una mujer y un hombre, junto con los correspondientes a las señales captadas por los micrófonos. Lo primero que notamos es que las señales recibidas por los micrófonos son muy similares, y por lo tanto es esperable que el problema resulte bastante difícil de resolver. Un análisis visual permite observar que las separaciones obtenidas por el método de Murata son ruidosas, y no están bien definidas, mientras que las obtenidas con el método de NMF penalizado son más claramente discernibles. Si bien la separación está lejos de ser perfecta, se reconocen mejor las diferencias entre las frecuencias fundamentales predominantes en cada espectrograma. Auditivamente, la separación es notoria, aunque se percibe cierta degradación en la calidad del audio.

Para las pruebas se utilizaron señales de audio de ocho hablantes, cuatro femeninos y cuatro masculinos. Para cada par mixto de hablantes, se generaron seis mezclas distintas. Se generaron respuestas al impulso artificiales a partir de una habitación simulada, con dos micrófonos y cuatro posibles ubicaciones para los hablantes, como se especifica en la Tabla 3.5. Las mezclas artificiales fueron creadas sumando las convoluciones discretas de cada señal de habla con la respuesta al impulso correspondiente a las posiciones del hablante y micrófono apropiadas. Esto nos dio un total de 96 señales para poner a prueba nuestro método.

Dado que nos encontramos en un contexto supervisado, hemos construido los diccionarios característicos para los hablantes mediante la minimización de una divergencia

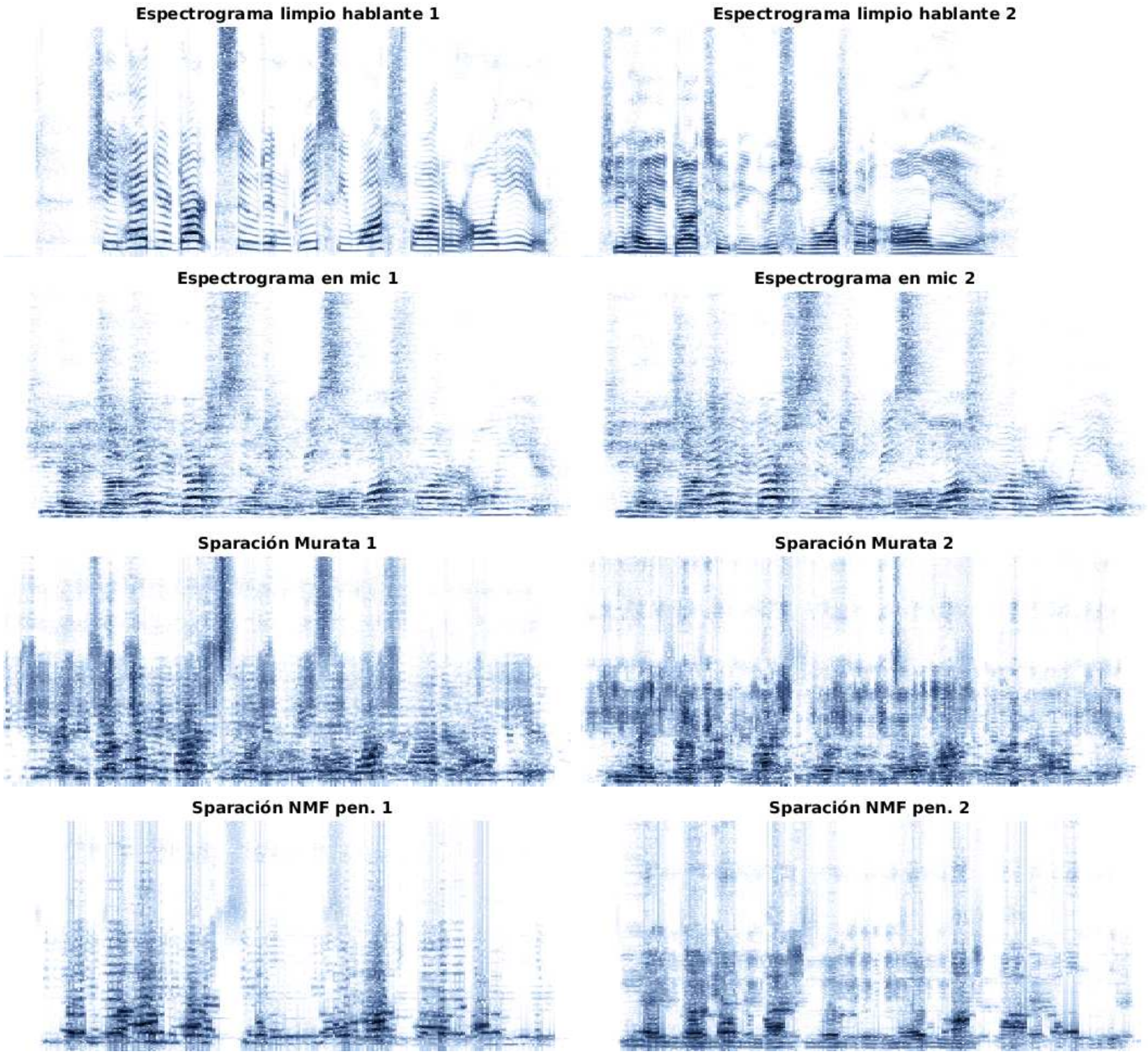


Figura 3.5: Ilustración del funcionamiento de los distintos métodos de separación.

β ($\beta = 0.75$, [21], [23, Anexos]). Luego, utilizamos los mismos diccionarios y las mismas inicializaciones para poner a prueba ambos métodos.

En la Figura 3.6 pueden verse los resultados obtenidos, especificados en la Tabla 3.6. Puede observarse que el enfoque aquí propuesto (NMF penalizado) mejora la separación (SIR) con respecto al propuesto por Murata *et al*, mientras disminuye la aparición de artefactos (SAR) y distorsión (SDR).

	Largo	Ancho	Altura
Dimensiones habitación	6.00 [m]	5.00 [m]	3.00 [m]
Posición micrófono 1	3.00 [m]	2.00 [m]	1.50 [m]
Posición micrófono 2	4.00 [m]	3.00 [m]	1.00 [m]
Posición fuente 1	1.00 [m]	3.00 [m]	1.70 [m]
Posición fuente 2	2.00 [m]	2.00 [m]	1.60 [m]
Posición fuente 3	3.00 [m]	4.00 [m]	1.80 [m]
Posición fuente 4	5.00 [m]	4.00 [m]	1.70 [m]

Tabla 3.5: Características de habitación y posiciones simuladas

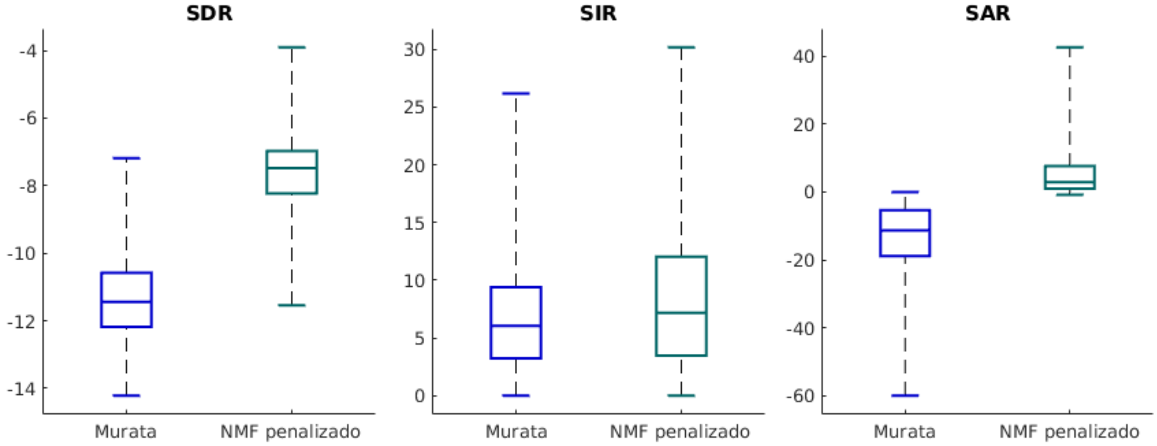


Figura 3.6: Visualización del desempeño obtenido con cada método de separación.

Método \ Medida	SDR	SIR	SAR
Mezcla	-4.728 (1.13)	2.462 (1.39)	5.424 (1.73)
Murata *et al*	-11.272 (1.39)	7.103 (5.64)	-14.820 (13.07)
NMF penalizado	-7.559 (1.17)	8.767 (7.39)	6.216 (8.77)

Tabla 3.6: Media y desviación estándar de desempeño con cada método de separación.

Capítulo 4

Discusión y conclusiones

En esta tesis doctoral presentamos un nuevo enfoque para los problemas de dereverberación y separación de fuentes en el contexto de procesamiento de habla, basado en modelos bayesianos y de penalización sobre representaciones no negativas.

En primer lugar, se realizó una exhaustiva revisión del estado del arte para analizar los fenómenos físicos que intervienen en los problemas a tratar y cómo modelarlos. A continuación, el problema se llevó al dominio tiempo-frecuencia a través de la transformada de Fourier de tiempo corto. Se estudió entonces la influencia del traslado de estos fenómenos del dominio temporal al dominio transformado. Por otra parte, se exploraron distintas representaciones no negativas, así como las diferentes maneras de utilizarlas para el modelado de espectrogramas de señales de audio.

A partir de lo anterior, se desarrollaron modelos basados en enfoques bayesianos y de penalización mixta sobre una representación en el dominio tiempo-frecuencia de las señales de audio. Diferentes modelos no negativos se propusieron y combinaron para representar los datos según las características de los fenómenos a modelar, incluyendo NMF en sus versiones clásica y convolutiva.

Para el problema de dereverberación, inicialmente se planteó un modelo de penalización sobre una representación NMF convolutiva de los datos. Se introdujeron penalizantes sobre el espectrograma asociado a la respuesta al impulso de la habitación, lo que permitió mejorar la representación obtenida. Más adelante, se modificó este modelo para incluir dependencia temporal desde un enfoque bayesiano. Finalmente, se incorporó un enfoque mixto, combinando NMF estándar y convolutivo, incorporando al nuevo método la capacidad de modelar la estructura espectral de las señales.

Dado que el desarrollo de un modelo como los aquí propuestos deriva en un problema de optimización de una función de costo, se procedió a estudiar los distintos métodos de optimización aplicables a este tipo de problemas. A partir del método de la función auxiliar, se desarrollaron algoritmos iterativos para abordar el problema de minimización asociado a cada función de costo. Puesto que en la mayoría de los casos las técnicas estándar no eran aplicables directamente, debieron desarrollarse reglas de actualización particulares *ad hoc.*

Finalmente, los algoritmos propuestos fueron probados en simulaciones y en señales reales grabadas. Se realizaron comparaciones con métodos del estado del arte que funcionan en las mismas condiciones que los propuestos, elegidos por su buen desempeño o su similitud con las nuevas propuestas y la consecuente capacidad de poner de manifiesto las ventajas introducidas.

Los resultados indicaron una mejora en las restauraciones obtenidos con la mayoría de los métodos propuestos, variando en función de la medida de desempeño utilizada. Con respecto al problema de dereverberación, los nuevos métodos superaron a los del estado del arte en cuanto a fwsSNR y SRMR, tanto en los experimentos simulados como en las grabaciones. En cuanto a la distancia cepstral, si bien se logró un mejor desempeño en las simulaciones, sólo el método NMF mixto logró una mejora significativa en las señales grabadas. Por otro lado, también se observó una mejora en el desempeño del algoritmo propuesto para separación de fuentes. Los resultados obtenidos con las mezclas simuladas son muy prometedores, quedando pendiente para trabajos futuros el estudio en mezclas reales.

Adicionalmente, se desarrolló un método para la optimización de parámetros en el modelo de NMF mixto, que resulta fácilmente extrapolable a otros métodos que conlleven un proceso en dos etapas. Además, se obtuvieron condiciones suficientes para la convergencia de los algoritmos iterativos basados en el método de la función auxiliar.

Si bien puede observarse que los objetivos propuestos han sido alcanzados, aún quedan muchos caminos por explorar en el contexto del uso de NMF para procesamiento de audio. Por un lado, la incorporación al modelo de las relaciones temporales existentes en las señales de voz podría mejorar la calidad de las restauraciones obtenidas. Estas relaciones también podrían tenerse en cuenta incorporando dependencias temporales a la estructura de los diccionarios. Por otra parte, si bien este trabajo está enfocado principalmente en un contexto ciego, los procedimientos aquí propuestos son directamente extrapolables a un problema supervisado. En este sentido, encontrar la manera "óptima" de incorporar la información disponible cuando el contexto en el que el problema de dereverberación o separación esté definido lo permita resulta un problema no trivial, que se pretende abordar en el futuro.

Bibliografía

[1] S. Yun, Y. J. Lee, and S. H. Kim, “Multilingual speech-to-speech translation system for mobile consumer devices,” *IEEE Transactions on Consumer Electronics*, vol. 60, no. 3, pp. 508–516, 2014.

[2] L. D. Vignolo, S. R. M. Prasanna, S. Dandapat, H. L. Rufiner, and D. H. Milone, “Feature optimisation for stress recognition in speech,” *Pattern Recognition Letters*, vol. 84, pp. 1–7, 2016.

[3] R. Sarikaya, P. A. Crook, A. Marin, M. Jeong, J.-P. Robichaud, A. Celikyilmaz, Y.-B. Kim, A. Rochette, O. Z. Khan, X. Liu *et al.*, “An overview of end-to-end language understanding and dialog management for personal digital assistants,” in *Spoken Language Technology Workshop (SLT), 2016 IEEE.* IEEE, 2016, pp. 391–397.

[4] I. J. Tashev, *Sound capture and processing: practical approaches.* John Wiley & Sons, 2009.

[5] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development.* Prentice hall PTR Upper Saddle River, 2001, vol. 95.

[6] S. Haykin and Z. Chen, “The cocktail party problem,” *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[7] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Processing Systems*, 2001, pp. 556–562.

[8] H. Kameoka, T. Nakatani, and T. Yoshioka, “Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms,” in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 45–48.

[9] N. Murata, H. Kameoka, K. Kinoshita, S. Araki, T. Nakatani, S. Koyama, and H. Saruwatari, “Reverberation-robust underdetermined source separation with non-negative tensor double deconvolution,” in *Signal Processing Conference (EUSIPCO), 2016 24th European.* IEEE, 2016, pp. 1648–1652.

[10] F. Ibarrola and R. Spies, “A two-step mixed inpainting method with curvature-based anisotropy and spatial adaptivity,” *Inverse Problems & Imaging*, vol. 11, no. 2, 2017.

[11] F. Ibarrola, G. Mazzieri, R. Spies, and K. Temperini, “Anisotropic BV-L^2 regularization of linear inverse ill-posed problems,” *Journal of Mathematical Analysis and Applications*, vol. 450, no. 1, pp. 427–443, 2017.

[12] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[13] J. B. Allen and L. R. Rabiner, “A unified approach to short-time fourier analysis and synthesis,” *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[14] F. J. Harris, “On the use of windows for harmonic analysis with the discrete fourier transform,” *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[15] P. Smaragdis, “Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs,” *Proceedings of the 5th Conference on Independent Component Analysis and Blind Signal Separation*, pp. 494–499, 2004.

[16] B. Yegnanarayana, P. S. Murthy, C. Avendaño, and H. Hermansky, “Enhancement of reverberant speech using lp residual,” in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1. IEEE, 1998, pp. 405–408.

[17] F. Ibarrola, L. Di Persia, and R. Spies, “A bayesian approach to convolutive nonnegative matrix factorization for blind speech dereverberation,” *Signal Processing*, vol. 151, pp. 89–98, 2018.

[18] ——, “Blind speech dereverberation using convolutive nonnegative matrix factorization with mixed penalization.” in *Proceedings of VI Congreso de Matemática Aplicada, Computacional e Industrial*, 2017, pp. 404–407.

[19] N. Mohammadiha, P. Smaragdis, and S. Doclo, “Joint acoustic and spectral modeling for speech dereverberation using non-negative representations,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on.* IEEE, 2015, pp. 4410–4414.

[20] F. Ibarrola, L. Di Persia, and R. Spies, “On the use of convolutive nonnegative matrix factorization with mixed penalization for blind speech dereverberation,” in *Computer Conference (CLEI), 2017 XLIII Latin American.* IEEE, 2017, pp. 1–4.

[21] R. Kompass, “A generalized divergence measure for nonnegative matrix factorization,” *Neural computation*, vol. 19, no. 3, pp. 780–791, 2007.

[22] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis,” *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[23] F. Ibarrola, L. Di Persia, and R. Spies, "Switching divergences for spectral learning in blind speech dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 5, pp. 881–891, 2019.

[24] ——, "Penalized nonnegative representations for speech separation," in *VII Congreso de Matemática Aplicada, Computacional e Industrial*. MACI, 2019.

[25] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[26] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[27] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[28] S. Wisdom, T. Powers, L. Atlas, and J. Pitton, "Enhancement of reverberant and noisy speech by extending its coherence," in *Proceedings of REVERB Challenge Workshop*, 2014, pp. 1–8.

[29] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[30] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[31] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Audio Engineering Society Convention 122*. Audio Engineering Society, 2007.

Anexo

Contribuciones

Se listan a continuación los trabajos relacionados con el uso de herramientas de NMF para procesamiento de señales de audio realizados durante el desarrollo del doctorado. El autor de esta tesis es primer autor en todas ellas, siendo el principal contribuyente en el desarrollo de modelos matemáticos, algoritmos, implementaciones numéricas, experimentos y escritura.

- **F. Ibarrola**, L. Di Persia, and R. Spies, "Blind speech dereverberation using convolutive nonegative matrix factorization with mixed penalization", *Proceedings of The VI Congreso de Matemática Aplicada, Computacional e Industrial*, pp. 404-407, 2017.

- **F. Ibarrola**, R. Spies, and L. Di Persia, "On the use of convolutive nonnegative matrix factorization with mixed penalization for blind speech dereverberation", *XLIII Latin American Computer Conference (CLEI)*, IEEE, pp. 1-4, 2017.

- **F. Ibarrola**, L. Di Persia, and R. Spies, "A Bayesian approach to convolutive nonnegative matrix factorization for blind speech dereverberation", *Signal Processing*, vol. 151, pp. 89-98, 2018.

- **F. Ibarrola**, L. Di Persia, and R. Spies, "Switching divergences for spectral learning in blind speech dereverberation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 5 , pp.8 81 - 891, 2019.

- **F. Ibarrola**, R. Spies, and L. Di Persia, "Penalized nonnegative representations for speech separation", *Proceedings of The VII Congreso de Matemática Aplicada, Computacional e Industrial*, 2019. En prensa.

Blind speech dereverberation using convolutive nonnegative matrix factorization with mixed penalization

Blind speech dereverberation using convolutive nonnegative matrix factorization with mixed penalization

Francisco J. Ibarrola *[1], Ruben D. Spies[2], and Leandro E. Di Persia[1]

[1]Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL, CONICET, Ciudad Universitaria UNL, (3000) Santa Fe, Argentina.
[2]Instituto de Matemática Aplicada del Litoral, IMAL, CONICET-UNL, Centro Científico Tecnológico CONICET Santa Fe, Colectora Ruta Nac. 168, km 472, Paraje "El Pozo", 3000, Santa Fe, Argentina and Departamento de Matemática, Facultad de Ingeniería Química, Universidad Nacional del Litoral, Santa Fe, Argentina.

Abstract

When a signal is recorded in an enclosed room, it typically gets affected by reverberation. This degradation represents a problem when dealing with audio signals, particularly in the field of speech applications, such as automatic speech recognition. Although there are some approaches to deal with this issue that are quite satisfactory under certain conditions, constructing a method that works well in a general context still poses a significant challenge. As an effort in this direction, we propose a method based on convolutive nonnegative matrix factorization that mixes two penalizers in order to impose certain characteristics over the time-frequency components of the restored signal and the reverberant components. An algorithm for finding such a solution is described and tested. The results show a significant improvement on the quality of the restored signals.

Keywords: signal processing, dereverberation, regularization
2000 AMS Subject Classification: 65F22 - 65T50

1 Introduction

When captured in enclosed rooms, audio recordings will most certainly be affected by reverberant components due to reflections of the sound waves in the walls, ceiling,

*fibarrola@sinc.unl.edu.ar

floor or furniture. This can severely degrade the characteristics of the recorded signal, generating difficult problems for processing such a signal, particularly when required for certain speech applications. The goal of any dereverberation technique is to remove or attenuate the reverberant components to obtain a cleaner signal. The dereverberation problem is called "blind" when the available data consists only of the reverberant signal itself, and this is the problem we shall address on this work.

Depending on the problem, our observation might consist of a single or multi-channel signal. That is, we might have a signal recorded by one or more microphones. For the latter case, there are several proposed methods that work quite well ([1]). For the case of single-channel, although some methods perform reasonably well ([2], [3], [4]), there is still much room for improvement.

In this work we present a dereverberation method for single channel data based on the idea of penalizing different characteristics of the components of a convolutive nonnegative matrix factorization (NMF) representation model for the reverberation phenomenon.

1.1 A Reverberation Model

Let $s, x : \mathbb{R} \to \mathbb{R}$ with support in$[0, \infty)$ be the functions associated to the clean and reverberant signals, respectively. Then, the reverberation model can be written as

$$x(t) = (h * s)(t), \tag{1}$$

where $h : \mathbb{R} \to \mathbb{R}$ is the room impulse response (RIR) signal, and "$*$" denotes convolution.

When dealing with sound signals (particularly speech signals), it is often convenient to work with the associated spectrograms rather than the signals themselves. Thus, we make use of the short time Fourier transform (STFT) on the discretized version of (1) to obtain the corresponding complex time-frequency components, resulting in the model

$$\mathbf{x}_k[t] = \sum_{\tau=0}^{T_h-1} \mathbf{s}_k[t-\tau]\mathbf{h}_k[\tau], \tag{2}$$

where $t = 1 \ldots T$, is a discretized time variable, $k = 1, \ldots K$, denotes the frequency subband and T_h is a parameter of the model associated to the maximum expected duration of the reverberation phenomenon.

Now, let us write $\mathbf{h}_k[\tau] = |\mathbf{h}_k[\tau]|e^{j\phi_k[\tau]}$. To overcome the problems derived from the well known sensitivity of the phase angle $\phi_k[\tau]$ with respect to variations of the reverberation conditions, we shall proceed as in [2], treating $\phi_k[\tau]$ as a random variable with distribution $\mathcal{U}[-\pi, \pi)$. Denoting the complex conjugate as * and the Kronecker delta as δ_{ij}, the latter assumption yields

$$\begin{aligned}
E|\mathbf{x}_k[t]|^2 &= E\sum_{\tau,\tau'} \mathbf{s}_k[t-\tau]\mathbf{s}_k^*[t-\tau']\mathbf{h}_k[\tau]\mathbf{h}_k^*[\tau'] \\
&= \sum_{\tau,\tau'} \mathbf{s}_k[t-\tau]\mathbf{s}_k^*[t-\tau']\,|\mathbf{h}_k[\tau]|\,|\mathbf{h}_k[\tau']|\,Ee^{j(\phi_k[\tau]-\phi_k[\tau'])} \\
&= \sum_{\tau,\tau'} \mathbf{s}_k[t-\tau]\mathbf{s}_k^*[t-\tau']\,|\mathbf{h}_k[\tau]|\,|\mathbf{h}_k[\tau']|\,\delta_{\tau\tau'} \\
&= \sum_{\tau} |\mathbf{s}_k[t-\tau]|^2\,|\mathbf{h}_k[\tau]|^2.
\end{aligned}$$

Finally, let us define $S_k[t] \doteq |\mathbf{s}_k[t]|^2$, $H_k[t] \doteq |\mathbf{h}_k[t]|^2$ and $X_k[t] \doteq E|\mathbf{x}_k[t]|^2$. Then, our model reads

$$X_k[t] = \sum_{\tau} S_k[t-\tau]H_k[\tau], \tag{3}$$

and the square magnitude of the observed spectrogram components can be written as

$$Y_k[t] = X_k[t] + \epsilon_k[t],$$

where $\epsilon_k[t]$ denotes the representation error. As shown in [2], this model is equivalent to a convolutive NMF with diagonal basis.

2 Mixed Penalization

As a way of measuring the representation error, we will use the square of the Frobenius norm $||Y-X||_F^2$, where Y and X are the matrices whose (k,t) components are $Y_k[t]$ and $X_k[t]$, respectively.

Since we are dealing with a blind dereverberation problem, we have no information on the structure of the matrix H (with elements $H_k[t]$). Hence, we must impose some conditions on the representation (3) in order to ensure that S and H will provide a satisfactory representation for our dereverberation problem.

For clean speech signals, the spectrogram is expected to have some sparse structure, which is not preserved under reverberant conditions (see Figure 1). This sparsity can be regained by introducing a penalization term over the matrix S. In a similar fashion, certain regularity conditions over the matrix H can be imposed to improve its correspondence with a room impulse response (RIR) signal.

Following these ideas, we propose the following cost function:

$$f(H,S) \doteq \sum_{t,k}(Y_k[t]-X_k[t])^2 + \lambda_1\sum_{t,k}|H_k[t]|^{p_1} + \lambda_2\sum_{t,k}|S_k[t]|^{p_2},$$

where $\lambda_1, \lambda_2 \geq 0$ are penalization parameters that quantify the weights of both penalizers relative to the fidelity term, whereas the exponents $p_1, p_2 \in (0,2)$ are tunning

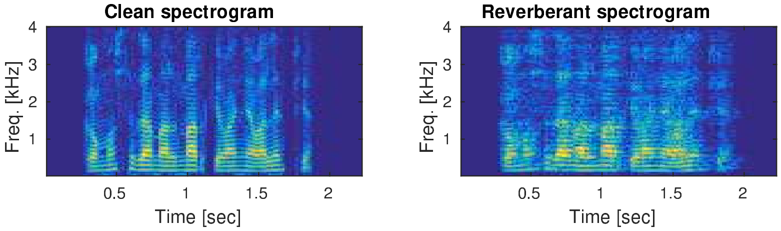


Figure 1: Spectrograms for a clean speech signal (left) and the corresponding reverberant speech signal (right).

parameters. Note that small values of these parameters will promote sparsity, whereas values close to 2 will promote smoothness. Since there is a clear scale indeterminacy in the representation (3), the additional constraint $\sum_{\tau=1}^{T_h} H_k[\tau] = 1 \ \forall k$ shall be imposed.

Next, we present an algorithm for approximating the matrices H and S that minimize f.

3 Updating rules

We shall build an iterative algorithm following the idea in [2], which is based on the auxiliary function technique.

Let $\Omega \subset \mathbb{R}$ and $f : \Omega \to \mathbb{R}_0^+$. Then, $g : \Omega \times \Omega \to \mathbb{R}_0^+$ is called an *auxiliary function* for f if $\forall w, w' \in \Omega$, $g(w, w') \geq f(w)$ and $g(w, w) = f(w)$. With this definition, it can be shown ([5]) that the sequence $\{f(w^j)\}_j$ is non-increasing under the update rule

$$w^j = \arg\min_w g(w, w^{j-1}). \tag{4}$$

We will use this approach to alternatively update the matrices H and S. Let us begin by fixing $H = H'$, where H' is an arbitrary $K \times T_h$ matrix. Then, it can be shown that an auxiliary function for f with respect to S is given by

$$\begin{aligned} g_s(S, S') \doteq & \sum_{k,t,\tau} \frac{S'_k[\tau] H'_k[t-\tau]}{X'_k[t]} \left(Y_k[t] - \frac{S_k[\tau]}{S'_k[\tau]} X'_k[t] \right)^2 + \lambda_1 \sum_{k,t} |H'_k[t]|^{p_1} \\ & + \lambda_2 \sum_{k,t} \left(\frac{p_2}{2} S'_k[t]^{p_2 - 2} S_k[t]^2 + |S'_k[t]|^{p_2} - \frac{p_2}{2} |S'_k[t]|^{p_2} \right), \end{aligned}$$

where $X'_k[t] = \sum_\tau S'_k[\tau] H'_k[t-\tau]$. In an analogous way, fixing $S = S'$, an auxiliary function for f with respect to H is given by

$$\begin{aligned} g_h(H, H') \doteq & \sum_{k,t,\tau} \frac{S'_k[t-\tau] H'_k[\tau]}{X'_k[t]} \left(Y_k[t] - \frac{H_k[\tau]}{H'_k[\tau]} X'_k[t] \right)^2 + \lambda_2 \sum_{k,t} |S'_k[t]|^{p_2} \\ & + \lambda_1 \sum_{k,t} \left(\frac{p_1}{2} H'_k[t]^{p_1 - 2} H_k[t]^2 + |H'_k[t]|^{p_1} - \frac{p_1}{2} |H'_k[t]|^{p_1} \right). \end{aligned}$$

Now, since g_s is quadratic with respect to S and g_h is quadratic with respect to H, we can use the first order necessary conditions to find the minimizers complying with the update rule (4). This leads to the following updating rules:

$$S_k[\tau] = S'_k[\tau]\frac{\sum_t H'_k[t-\tau]Y_k[t]}{\sum_t H'_k[t-\tau]X'_k[t] + \frac{\lambda_2 p_2}{2}|S'_k[\tau]|^{p_2-1}}, \qquad H_k[\tau] = H'_k[\tau]\frac{\sum_t S'_k[t-\tau]Y_k[t]}{\sum_t S'_k[t-\tau]X'_k[t] + \frac{\lambda_1 p_1}{2}|H'_k[\tau]|^{p_1-1}}.$$

Every updating step must be followed by a normalization of the rows of H to avoid the aforementioned scale indeterminacy issue. In principle, the algorithm is run until $\|S - S'\|_F^2$ reachs an established threshold value, but it is worth noting that other stopping criteria might also be suitable.

4 Experimental results

We begin by showing an example of the performance of our method. Starting from a clean speech signal sampled at 8kHz, we have artificially constructed a reverberant version by discrete convolution with a RIR signal from a simulated enclosed room with 400ms of reverberation time. The spectrogram was then computed using STFT with 256 window length and overlapping of 128 samples. Figure 2 shows the clean speech spectrogram, together with its reverberant version and a restoration using our method with parameters $p_1 = 1.8$ and $p_2 = 1.2$, meaning we impose some sparsity to S and a mild smoothness to H.

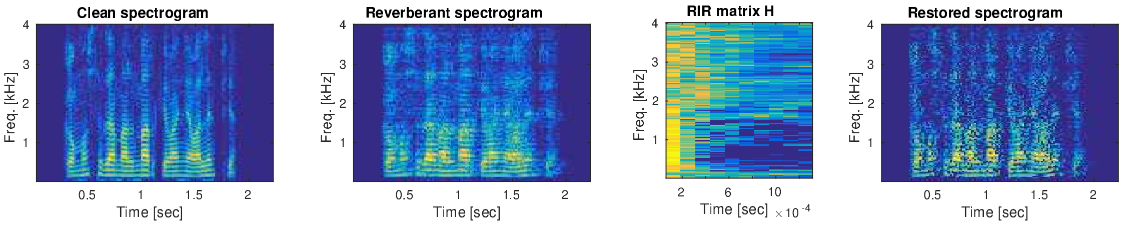


Figure 2: Spectrograms for a clean speech signal, its reverberant version, the RIR matrix and the obtained restoration.

To measure the performance of the method, we used the *frequency weighted segmental SNR* (fwsSNR) for its relevance for speech applications such as automatic speech recognition ([6]). The fwsSNR values are 16.20 for the reverberant signal and 17.41 for the restored example, indicating a significant improvement.

Next, we compare our method with the one proposed by Kameoka *et. al.* ([2]), which essentially consists of single penalization based on a Bayesian approach. To do so, both methods were run on artificially constructed reverberant signals with six different RIRs (three different microphone/source positions and two reverberation times) from a database of 20 clean speech signals. The parameters of the model were set as: $T_h = 10$, $p_1 = 1.8$ and $p_2 = 1.2$. For the sake of comparison, λ_2 and the maximum number of iterations (set as 20) were chosen as in [2] and λ_1 was chosen as $\lambda_2 \times 10^3$. The results of the experiment are summarized in Table 1, where improvements on the mean fwsSNR values (over the 20 signals) can be seen.

	RIR1	RIR2	RIR3	RIR4	RIR5	RIR6
Reverberant signal	16.93	14.19	17.86	15.19	18.00	15.76
Kameoka's restoration	17.38	14.58	18.38	15.66	18.49	16.25
Mixed pen. restoration	17.60	14.76	18.53	15.88	18.52	16.48

Table 1: Experimental results measures: mean fwsSNR values for speech dereverberation.

5 Conclusions

In this work we presented a model for signal dereverberation based on convolutive NMF with mixed penalization. An iterative updating algorithm was introduced and its performance was tested and compared with a state of the art method. The results show that our mixed penalization improves the quality of the restorations

Although these preliminary results are promising, there is still room for improvement. For instance, other types of penalizing terms can be used, and different ways to optimize the model parameters can be sought.

.

Acknowledgements

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET through PIP 2014-2016 N^{o}11220130100216-CO, the Air Force Office of Scientific Research, AFOSR/SOARD, through Grant FA9550-14-1-0130 and by Universidad Nacional del Litoral, UNL, through CAID-UNL 2011 N^{o} 50120110100519 "Procesamiento de Señales Biomédicas."

References

[1] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani *et al.*, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge," in *Proceedings of REVERB Challenge Workshop*, 2014.

[2] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 45–48.

[3] S. Xizhong and M. Guang, "Complex cepstrum based single channel speech dereverberation," in *2009 4th International Conference on Computer Science & Education.* IEEE, 2009, pp. 7–11.

[4] M. Moshirynia, F. Razzazi, and A. Haghbin, “A speech dereverberation method using adaptive sparse dictionary learning,” in *Proceedings of REVERB Challenge Workshop*, 2014.

[5] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Processing Systems*, 2001, pp. 556–562.

[6] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

On the use of convolutive nonnegative matrix factorization with mixed penalization for blind speech dereverberation

On the use of convolutive nonnegative matrix factorization with mixed penalization for blind speech dereverberation

Francisco J. Ibarrola *[1], Ruben D. Spies[2], and Leandro E. Di Persia[1]

[1]Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL, CONICET, Ciudad Universitaria UNL, (3000) Santa Fe, Argentina.
[2]Instituto de Matemática Aplicada del Litoral, IMAL, CONICET-UNL, Centro Científico Tecnológico CONICET Santa Fe, Colectora Ruta Nac. 168, km 472, Paraje "El Pozo", 3000, Santa Fe, Argentina and Departamento de Matemática, Facultad de Ingeniería Química, Universidad Nacional del Litoral, Santa Fe, Argentina.

Abstract

When a signal is recorded in an enclosed room, it typically gets affected by reverberation. This degradation represents a problem when dealing with audio signals, particularly for applications involving automatic speech and/or speaker recognition. There are some approaches to deal with this issue that are quite satisfactory when multi-channel recordings or learning data are available, but this is not the general case in most human-computer interaction applications, and constructing a method that works well in a general context still poses a significant challenge. In this article, we propose a method based on convolutive nonnegative matrix factorization that mixes two penalizers in order to impose certain characteristics over the time-frequency components of the restored signal and the reverberant components. An algorithm for finding such a solution is described and tested. Comparisons of the results against state of the art methods are presented, showing significant improvement.

Keywords: signal processing, dereverberation, regularization.

1 Introduction

When captured in enclosed rooms, audio recordings will most certainly be affected by reverberant components due to reflections of the sound waves in the walls, ceiling, floor or furniture. This can severely degrade the characteristics of the recorded signal ([1]), generating difficult problems for processing such a signal, particularly when required for certain speech applications ([2]). The goal of any dereverberation technique is to remove or attenuate the reverberant components to obtain a cleaner signal. The dereverberation problem is called "blind" when the available data consists only of the reverberant signal itself, and this is the problem we shall address on this work.

Depending on the problem, our observation might consist of a single or multi-channel signal. That is, we might have a signal recorded by one or more microphones. For the latter case, there are several proposed methods that work quite well ([3]). For the case of single-channel, although some methods perform reasonably well ([4], [5], [6]), there is still much room for improvement.

[*] fibarrola@sinc.unl.edu.ar

In this work we present a dereverberation method for single channel data based on the idea of penalizing different characteristics of the components of a convolutive nonnegative matrix factorization (NMF) representation model for the reverberation phenomenon.

May 27, 2019

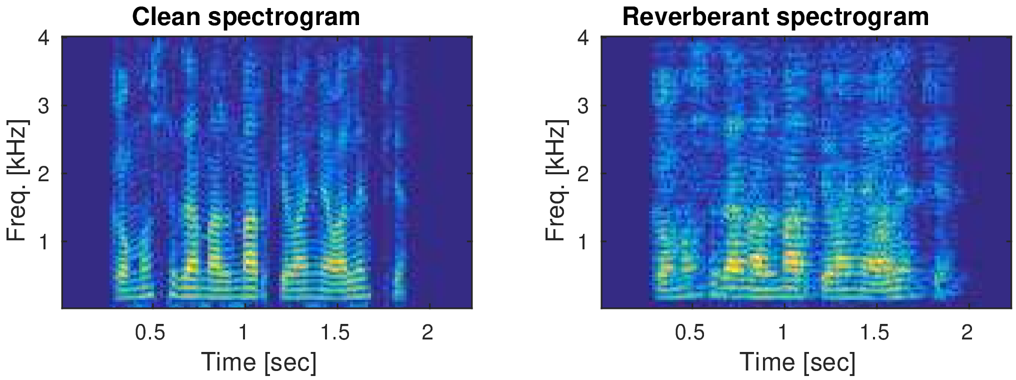


Figure 1: Spectrograms for a clean speech signal (left) and the corresponding reverberant speech signal (right).

Let $s, x : \mathbb{R} \to \mathbb{R}$, with support in $[0, \infty)$, be the functions associated with the clean and reverberant signals, respectively. Then, our reverberation model can be written as

$$x(t) = (h * s)(t), \tag{1}$$

where $h : \mathbb{R} \to \mathbb{R}$ is the room impulse response (RIR) signal, and "$*$" denotes convolution. This model is valid under the hypothesis of a linear, time-invariant system. In practice, this implies we are assuming the source and microphone positions to be static, and the signal energy to be low enough for the effect of the non-linear components to be insignificant.

When dealing with sound signals (particularly speech signals), it is often convenient to work with the associated spectrograms rather than the signals themselves. Thus, we make use of the short time Fourier transform (STFT), defined as

$$\mathbf{x}_k[t] \doteq \int_{-\infty}^{\infty} x(u)w(u-t)e^{-2\pi iuk}du, \quad t, k \in \mathbb{R}$$

where $w : \mathbb{R} \to \mathbb{R}_0^+$ is a given *window* function. Denoting the STFTs of h and s by $\mathbf{s}_k[t]$ and $\mathbf{h}_k[t]$, respectively, a discretized approximation of the STFT model associated to (1) is given ([4]) by

$$\mathbf{x}_k[t] \approx \tilde{\mathbf{x}}_k[t] \doteq \sum_{\tau=0}^{T_h-1} \mathbf{s}_k[t-\tau]\mathbf{h}_k[\tau], \tag{2}$$

where $t = 1, \ldots, T$, is a discretized time variable that corresponds to window locations, $k = 1, \ldots, K$, denotes the frequency subband and T_h is a parameter of the model associated to the expected maximum duration of the reverberation phenomenon. Later on, the values of t will be chosen in such a way that the union of the windows' supports contain the support of the observed signal, and the values of k in such a way that they cover the whole frequency spectrum, up to half the sampling frequency.

Now, let us write $\mathbf{h}_k[\tau] = |\mathbf{h}_k[\tau]|e^{j\phi_k[\tau]}$. It is well known that the phase angles $\phi_k[\tau]$ are highly sensitive with respect to mild variations on the reverberation conditions. To overcome the problems derived from this, we shall proceed (see [4]) to treat the $K \times T_h$ variables $\phi_k[\tau]$ as random variables *i.i.d.* with uniform distribution in $[-\pi, \pi)$. Denoting the complex conjugate as "*" and the Kronecker delta as δ_{ij}, the expected value of $|\tilde{\mathbf{x}}_k[t]|^2$ is given by

$$\begin{aligned}
E|\tilde{\mathbf{x}}_k[t]|^2 &= E\Big(\sum_{\tau,\nu} \mathbf{s}_k[t-\tau]\mathbf{s}_k^*[t-\nu]\mathbf{h}_k[\tau]\mathbf{h}_k^*[\nu] \Big) \\
&= \sum_{\tau,\nu} \mathbf{s}_k[t-\tau]\mathbf{s}_k^*[t-\nu]|\mathbf{h}_k[\tau]||\mathbf{h}_k[\nu]|Ee^{j(\phi_k[\tau]-\phi_k[\nu])} \\
&= \sum_{\tau,\nu} \mathbf{s}_k[t-\tau]\mathbf{s}_k^*[t-\nu]\,|\mathbf{h}_k[\tau]|\,|\mathbf{h}_k[\nu]|\,\delta_{\tau\nu} \\
&= \sum_{\tau} |\mathbf{s}_k[t-\tau]|^2\,|\mathbf{h}_k[\tau]|^2.
\end{aligned}$$

Note that the $[-\pi, \pi)$ interval choice for $\phi_k[\tau]$ is arbitrary, since this result holds for any $2\pi-$length interval. Finally, let us define $S_k[t] \doteq |\mathbf{s}_k[t]|^2$, $H_k[t] \doteq |\mathbf{h}_k[t]|^2$ and $X_k[t] \doteq E|\tilde{\mathbf{x}}_k[t]|^2$. Then, our model reads

$$X_k[t] = \sum_{\tau} S_k[t-\tau]H_k[\tau], \tag{3}$$

and the square magnitude of the observed spectrogram components can be written as

$$Y_k[t] = X_k[t] + \epsilon_k[t], \tag{4}$$

where $\epsilon_k[t]$ denotes the representation error. As shown in [4], this model is equivalent to a convolutive NMF ([11]) with diagonal basis. In the next section, we build a cost function in order to find an appropriate convolutive representation that allows us to isolate the components of $S_k[t]$.

2 Mixed Penalization

As a way of measuring the representation error, we will use the square of the Frobenius norm $||Y-X||_F^2$, where Y and X are the matrices whose (k,t) components are $Y_k[t]$ and $X_k[t]$, respectively.

Since we are dealing with a blind dereverberation problem, we have no information on the structure of the matrix H (with elements $H_k[t]$). Hence, we must impose some conditions on the representation (3) in order to ensure that S and H will provide a satisfactory representation for our dereverberation problem.

As it can be seen in Figure 1, for clean speech signals, the spectrogram is expected to have some sparse structure, which is not preserved under reverberant conditions. Sparsity can be regained by introducing a penalization term over the matrix S. In a similar fashion, certain regularity conditions over the matrix H can be imposed to improve its correspondence with a room impulse response (RIR) signal.

Based upon these ideas, we propose the following cost function:

$$J(H,S) \doteq \sum_{t,k} \left[(Y_k[t]-X_k[t])^2 + \lambda_{1,k}|H_k[t]|^{p_1} + \lambda_{2,k}|S_k[t]|^{p_2} \right],$$

where $\lambda_{1,k}, \lambda_{2,k} \geq 0$ are penalization parameters that quantify the weights of both penalizers relative to the fidelity term, whereas the exponents $p_1, p_2 \in (0,2)$ are tunning parameters. Small values of these parameters will promote sparsity, whereas values close to 2 will promote smoothness. Since there is a clear scale indeterminacy in the representation (3), we impose the (somewhat arbitrary) additional constraint $||S_k||_2 = ||Y_k||_2\ \forall k$, which means that the ℓ^2-norm (energy) shall remain equal for every frequency.

2.1 Regularization parameters

As mentioned before, the parameters $\lambda_{1,k}, \lambda_{2,k},\ k = 1, \ldots, K$, weight the penalizers against the fidelity term. In this sense, the optimal weights of these regularization parameters might vary as a function of the frequency subband, and hence their proposed dependency on k. Since searching blindly for $2K$ parameters is non-viable in practice, we quantify this dependency by defining $\lambda_{1,k} \doteq \lambda_1 \sum_{t=1}^{T} |Y_k[t]|^2$ and $\lambda_{2,k} \doteq \lambda_2 \sum_{t=1}^{T} |Y_k[t]|^2$. This means we only need to look for two paramenters (λ_1, λ_2) and then multiply them by the energy of the signal associated to each row of Y.

Next, we present an algorithm for approximating matrices H and S that minimize J.

3 Updating rules

We shall build an iterative algorithm following the idea in [4], which is based on the auxiliary function technique.

Let $\Omega \subset \mathbb{R}$ and $f : \Omega \to \mathbb{R}_0^+$. Then, $g : \Omega \times \Omega \to \mathbb{R}_0^+$ is called an *auxiliary function* for f if $\forall w, w' \in \Omega$, $g(w, w') \geq f(w)$ and $g(w, w) = f(w)$. With this definition, it can be shown ([7]) that for any $w^0 \in \Omega$, the sequence $\{f(w^j)\}_{j=0}^{\infty}$ is non-increasing under the update rule

$$w^j = \arg\min_{w} g(w, w^{j-1}),\ \ j = 1, \ldots, \infty. \tag{5}$$

We will use this approach to alternatively update the matrices H and S. Let us begin by fixing $H = H'$, where H' is an arbitrary $K \times T_h$ matrix. Then, if we let

$$X'_k[t] = \sum_{\tau} S'_k[\tau] H'_k[t-\tau],$$

it can be shown that the function g_s, defined as

$$\begin{aligned} g_s(S, S') \doteq &\sum_{k,t,\tau} \frac{S'_k[\tau] H'_k[t-\tau]}{X'_k[t]} \left(Y_k[t] - \frac{S_k[\tau]}{S'_k[\tau]} X'_k[t] \right)^2 \\ &+ \sum_{k,t} \lambda_{1,k} |H'_k[t]|^{p_1} \\ &+ \sum_{k,t} \lambda_{2,k} \Big(\frac{p_2}{2} S'_k[t]^{p_2-2} S_k[t]^2 + \Big(1 - \frac{p_2}{2}\Big) |S'_k[t]|^{p_2} \Big), \end{aligned}$$

is an auxiliary function for J with respect to S.

In an analogous way, fixing $S = S'$, an auxiliary function for J with respect to H is given by g_h, defined as

$$\begin{aligned} g_h(H, H') \doteq &\sum_{k,t,\tau} \frac{S'_k[t-\tau] H'_k[\tau]}{X'_k[t]} \left(Y_k[t] - \frac{H_k[\tau]}{H'_k[\tau]} X'_k[t] \right)^2 \\ &+ \sum_{k,t} \lambda_{1,k} \Big(\frac{p_1}{2} H'_k[t]^{p_1-2} H_k[t]^2 + \frac{2-p_1}{2} |H'_k[t]|^{p_1} \Big) \\ &+ \sum_{k,t} \lambda_{2,k} |S'_k[t]|^{p_2}. \end{aligned}$$

Now, since g_s is quadratic with respect to S and g_h is quadratic with respect to H, we can use the first order necessary conditions to find the minimizers complying with the update rule (5). This leads to the following updating rules:

$$S_k[\tau] = S'_k[\tau] \frac{\sum_t H'_k[t-\tau] Y_k[t]}{\sum_t H'_k[t-\tau] X'_k[t] + \frac{\lambda_{2,k}}{2} p_2 |S'_k[\tau]|^{p_2-1}},$$

$$H_k[\tau] = H'_k[\tau]\frac{\sum_t S'_k[t-\tau]Y_k[t]}{\sum_t S'_k[t-\tau]X'_k[t] + \frac{\lambda_{1,k}}{2}p_1|H'_k[\tau]|^{p_1-1}}.$$

In order to avoid the aforementioned scale indeterminacy, every updating step is to be followed by scaling S_k so that its ℓ^2 norm coincides with that of the observation Y_k. In principle, the algorithm is run until $\|S - S'\|_F^2$ decreases below an established threshold value, although it is worth noting that other stopping criteria might also be suitable.

4 Experimental results

For the experiments, we took 110 speech signals from the TIMIT database[1], recorded at 16 KHz, and we artificially made them reverberant using the software Room Impulse Response Generator by E.A.P. Habets[2], based on the model in [9]. Each signal was degraded under different reverberation conditions: three different room sizes, each with three different microphone positions and four different reverberation times.

In order to avoid preprocessing, the choice of the regularization parameters was made *a priori* by means of empirical rules, based upon signals from a different database. This is supported by the fact that the parameters were observed to be rather robust with respect to variations of the reverberation conditions, and hence they were chosen simply as $\lambda_1 = 1$ and $\lambda_2 = 10^{-4}$. The rest of the model parameters were chosen as specified in Table 1.

p_1	p_2	T_h	window size	win. overlapping	max. iter.
1.8	1	15	256 samples	128 samples	20

Table 1: Model parameter values

As previously discussed, the choice of $p_1 = 1.8$ is meant to promote smoothness over H, while the choice of $p_2 = 1$ aims to induce sparsity over S.

In order to evaluate the performance of our model, we made comparisons against two state of the art methods under the same conditions: the one proposed by Kameoka *et al* in [4], and the one proposed by Wisdom et. al. in [10] (with a window length of 2048), choosing all the parameters as suggested by the authors.

To measure performance, following [8], we made use of the frequency weighted segmental signal-to-noise ratio (fwsSNR) and the cepstral distance. The results for each performance measure are stated in Tables 2 and 3 and depicted in Figure 2, the different reverberation times: 300[ms], 450[ms], 600[ms] and 750[ms]. Notice that for the case of fwsSNR, higher values correspond to better performance, while for the cepstral distance, small values indicate higher quality.

Rev. time	Rev. Signal	Kameoka's	Wisdom's	Mixed pen.
300 [ms]	8.102(1.96)	7.950(1.73)	8.262(1.53)	**9.148**(1.71)
450 [ms]	4.815(1.42)	5.127(1.36)	5.771(1.28)	**6.458**(1.45)
600 [ms]	3.082(1.20)	3.358(1.19)	4.140(1.17)	**4.547**(1.31)
750 [ms]	1.998(1.11)	2.184(1.10)	3.013(1.12)	**3.239**(1.22)

Table 2: Mean and (standard deviation) of fwsSNR for each method and reverberation time.

In regard to the fwsSNR peformance measure, the values in Table 2 give account of a strong improvement of our proposed method with respect to the others. As for the cepstral distance, although our method outperforms the other two, an improvement with respect to the reverberant signal is

[1] https://catalog.ldc.upenn.edu/ldc93s1

[2] https://github.com/ehabets/RIR-Generator

Table 3: Mean and (standard deviation) of cepstral distance for each method and reverberation time.

Rev. time	Rev. Signal	Kameoka's	Wisdom's	Mixed pen.
300 [ms]	3.440(0.44)	4.057(0.45)	3.908(0.48)	**3.566**(0.44)
450 [ms]	4.264(0.44)	4.636(0.42)	4.511(0.41)	**4.124**(0.41)
600 [ms]	4.716(0.46)	5.006(0.42)	4.860(0.40)	**4.519**(0.41)
750 [ms]	5.011(0.48)	5.264(0.43)	5.089(0.40)	**4.807**(0.42)

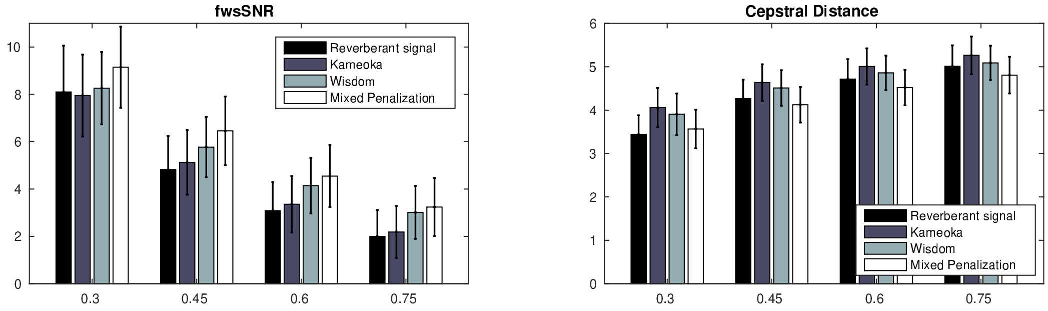


Figure 2: Mean and standard deviations of performance measures for different reverberation times.

observed only for reverberation times of 450[ms] or greater. A t-test with significance level $\alpha = 0.05$ was done using all the obtained results, showing statistical significance of the improvement on the performance of our method with respect to the reverberant signal and the other methods.

5 Conclusions

In this work we presented a model for signal dereverberation based on convolutive NMF with mixed penalization. An iterative updating algorithm was introduced and its performance was tested and compared with two state of the art methods. The results show that our mixed penalization method improves the quality of the restorations.

Although these preliminary results are promising, there is still room for improvement. For instance, other types of penalizing terms can be used, different ways to optimize the model parameters can be sought, etcetera.

Acknowledgments

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET through PIP 2014-2016 N° 11220130100216-CO, the Air Force Office of Scientific Research, AFOSR/SOARD, through Grant FA9550-14-1-0130 and by Universidad Nacional del Litoral, UNL, through CAID-UNL 2011 N°50120110100519 "Procesamiento de Señales Biomédicas." and CAI+D-UNL 2016, PIC 50420150100036LI "Problemas Inversos y Aplicaciones a Procesamiento de Señales e Imágenes".

References

[1] I. Tashev, *Sound Capture and Processing: Practical Approaches*, John Wiley & Sons, New Jersey, 2009.

[2] X. Huang, A.Acero and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, New Jersey, 2001.

[3] M. Delcroix, T. Yoshioka, A. Ogawam Y. Kubo, M. Fujimoto, I. Nobutaka, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, A. Namakura, *Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB Challenge*, Proceedings of Reverb Challenge 02.3 (2014).

[4] H. Kameoka, T. Nakatani, T Yoshioka, *Robust speech dereverberation based on non-negativity and sparse nature o speech spectrograms*, ICASSP (2009), pp. 45-48.

[5] S. Xizhong and M Guang, *Complex cepstrum based single channel speech dereverberation*, Proceedings of 4th International Conference on Computer Science & Education (2009), pp. 7-11.

[6] M. Moshirynia, F. Razzazi, A Haghbin, *A speech dereverberation method using adaptive sparse dictionary learning*, REVERB Workshop (2014), pp. 1-7.

[7] D. D. Lee, H. S. Seung, *Algorithms for non-negative matrix factorization*, NIPS (2000), pp. 556-562.

[8] Y. Hu and P. C. Loizou, *Evaluation of objective quality measures for speech enhancement*, IEEE Trans. Audio, Speech, Lang. Process. (2008), 16, pp. 229-238.

[9] J.B. Allen and D.A. Berkley, *Image method for efficiently simulating small-room acoustics*,Journal Acoustic Society of America. (1979), 65, pp. 943-950.

[10] S. Wisdom, T. Powers, L. Atlas and J. Pitton, *Enhancement of reverberant and noisy speech by extending its coherence*, Proceedings of Reverb Challenge (2014).

[11] P. Smaragdis, *Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs* , Fifth International Conference on Independent Component Analysis, LNCS 3195 (2004), pp 494-499.

A Bayesian approach to convolutive nonnegative matrix factorization for blind speech dereverberation

A Bayesian approach to convolutive nonnegative matrix factorization for blind speech dereverberation

Francisco J. Ibarrola *[1], Ruben D. Spies[2], and Leandro E. Di Persia[1]

[1]Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL, CONICET, Ciudad Universitaria UNL, (3000) Santa Fe, Argentina.
[2]Instituto de Matemática Aplicada del Litoral, IMAL, CONICET-UNL, Centro Científico Tecnológico CONICET Santa Fe, Colectora Ruta Nac. 168, km 472, Paraje "El Pozo", 3000, Santa Fe, Argentina and Departamento de Matemática, Facultad de Ingeniería Química, Universidad Nacional del Litoral, Santa Fe, Argentina.

Abstract

When a signal is recorded in an enclosed room, it typically gets affected by reverberation. This degradation represents a problem when dealing with audio signals, particularly in the field of speech signal processing, such as automatic speech recognition. Although there are some approaches to deal with this issue that are quite satisfactory under certain conditions, constructing a method that works well in a general context still poses a significant challenge. In this article, we propose a Bayesian approach based on convolutive nonnegative matrix factorization that uses prior distributions in order to impose certain characteristics over the time-frequency components of the restored signal and the reverberant components. An algorithm for implementing the method is described and tested. Comparisons of the results against those obtained with state-of-the-art methods are presented, showing significant improvement.

Keywords: signal processing, dereverberation, regularization.

1 Introduction

In recent years, many technological developments have attracted attention towards human-machine interaction. Since the most natural and easiest way of human communication is through speech, much research effort has been put into achieving the same natural interaction with machines. This effort has already generated many advances in a wide variety of fields such as automatic speech recognition ([1]), automatic translation systems ([2]) and control of remote devices through voice ([3]), to name only a few. A significant amount of work has been recently devoted to produce robustness in speech recognition ([4]), resulting in several advances in the areas of speech enhancement ([1], [5]), multiple sources separation ([6], [7]), and particularly in dereverberation techniques ([8]), which constitute the topic of this work.

When recorded in enclosed rooms, audio signals will most certainly be affected by reverberant components due to reflections of the sound waves in the walls, ceiling, floor or furniture. This can severely degrade the characteristics of the recorded signal ([9]), generating difficult problems for its processing, particularly when required for certain speech applications ([10]). The goal of any dereverberation technique is to remove or to attenuate the reverberant components in order to obtain a cleaner signal. The dereverberation problem is called "blind" when the available data consists only of the reverberant signal itself, and this is the problem we shall deal with in this work.

*fibarrola@sinc.unl.edu.ar

Depending on the problem, our observation might consist of a single or multi-channel signal, that is, we might have a signal recorded by one or more microphones. For the latter case, quite a few methods exist that work relatively well ([11], [12]).

For the single-channel case, we may distinguish between supervised and unsupervised approaches. The first kind refers to those that begin with a training stage that serves to learn some characteristics of the reververation conditions, while the second kind alludes to those methods that can be implemented directly over the reverberant signal. Some supervised methods ([13], [14], [15]) appear to perform somewhat better than unsupervised ones, but they pose the disadvantage of needing learning data corresponding to the specific room conditions, microphone and source locations, and a previous process that might take a significant amount of time.

In the context of unsupervised blind dereverberation, although some recently proposed methods ([12], [16]) work reasonably well, there is still much room for improvement. Our work is based on a convolutive non-negative matrix factorization (NMF) reverberation model, as proposed by Kameoka *et al* ([16]), along with a Bayesian approach for building a functional that takes into account *a priori* expected characteristics over the elements of the representation model. This functional can be thought of as the cost function of a mixed penalization model, such as in [17]. This kind of approach has been also recently used and successfully applied by several authors in many areas, mainly in signal and image processing applications ([18], [19], [20], [21], [22]). These techniques have shown to produce good results in terms of enhancing certain desirable characteristics on the solutions while precluding unwanted ones.

2 A Reverberation Model

Let $s, x : \mathbb{R} \to \mathbb{R}$, with support in $[0, \infty)$, be the functions associated to the clean and reverberant signals, respectively. As it is customary, we shall assume that the reverberation process is well represented by a Linear Time-Invariant (LTI) system. Thus, the reverberation model can be written as

$$x(t) = (h * s)(t), \tag{1}$$

where $h : \mathbb{R} \to \mathbb{R}$ is the room impulse response (RIR) signal, and "$*$" denotes convolution. This LTI hypothesis implies we are assuming the source and microphone positions to be static, and the energy of the signal to be low enough for the effect of the non-linear components to be relatively insignificant.

When dealing with sound signals (particularly speech signals), it is often convenient to work with the associated spectrograms rather than the signals themselves. Thus, we make use of the short time Fourier transform (STFT), defined as

$$\mathbf{x}_k(t) \doteq \int_{-\infty}^{\infty} x(u)w(u-t)e^{-2\pi iuk}du, \ \ t, k \in \mathbb{R},$$

where $w : \mathbb{R} \to \mathbb{R}_0^+$ is a compactly supported, even function such that $\|w\|_1 = 1$. This function is called *window*.

In practice, we work with discretized versions of the signals involved ($x[\cdot], h[\cdot], s[\cdot]$, and $w[\cdot]$). With this in mind, we shall define the discrete STFT as

$$\mathbf{x}_k[n] \doteq \sum_{m=-\infty}^{\infty} x[m]w[m-n]e^{-2\pi imk}, \ \ n, k \in \mathbb{N}.$$

Denoting the STFTs of s and h by $\mathbf{s}_k[n]$ and $\mathbf{h}_k[n]$, respectively, a discretized approximation of the STFT model associated to (1) is given by

$$\mathbf{x}_k[n] \approx \tilde{\mathbf{x}}_k[n] \doteq \sum_{\tau=0}^{N_h-1} \mathbf{s}_k[n-\tau]\mathbf{h}_k[\tau], \tag{2}$$

where $n = 1, \dots, N$, is a discretized time variable that corresponds to window location, $k = 1, \dots, K$, denotes the frequency subband and N_h is a parameter of the model associated to the expected maximum duration of the reverberation phenomenon. The model is built as in [23], being the approximation due to the use of band-to-band filters only. Later on, the values of n will be chosen in such a way that the union of the windows' supports contain the support of the observed signal, and the values of k in such a way that they cover the whole frequency spectrum, up to half the sampling frequency.

Now, let us write $\mathbf{h}_k[\tau] = |\mathbf{h}_k[\tau]|e^{j\phi_k[\tau]}$. It is well known ([24]) that the phase angles $\phi_k[\tau]$ are highly sensitive with respect to mild variations on the reverberation conditions. To overcome the problems derived from this, we shall proceed (see [16]) treating the $K \times N_h$ variables $\phi_k[\tau]$ as *i.i.d.* random variables with uniform distribution in $[-\pi, \pi)$. Denoting the complex conjugate by "*" and the Kronecker delta by δ_{ij}, the expected value of $|\tilde{\mathbf{x}}_k[t]|^2$ is given by

$$\begin{aligned}
E|\tilde{\mathbf{x}}_k[n]|^2 &= E\sum_{\tau,\tau'} \mathbf{s}_k[n-\tau]\mathbf{s}_k^*[n-\tau']\mathbf{h}_k[\tau]\mathbf{h}_k^*[\tau'] \\
&= E\sum_{\tau,\tau'} \mathbf{s}_k[n-\tau]\mathbf{s}_k^*[n-\tau']\,|\mathbf{h}_k[\tau]|\,e^{j\phi_k[\tau]}\,|\mathbf{h}_k[\tau']|\,e^{-j\phi_k[\tau']} \\
&= \sum_{\tau,\tau'} \mathbf{s}_k[n-\tau]\mathbf{s}_k^*[n-\tau']\,|\mathbf{h}_k[\tau]|\,|\mathbf{h}_k[\tau']|\,Ee^{j(\phi_k[\tau]-\phi_k[\tau'])} \\
&= \sum_{\tau,\tau'} \mathbf{s}_k[n-\tau]\mathbf{s}_k^*[n-\tau']\,|\mathbf{h}_k[\tau]|\,|\mathbf{h}_k[\tau']|\,\delta_{\tau\tau'} \\
&= \sum_{\tau} |\mathbf{s}_k[n-\tau]|^2\,|\mathbf{h}_k[\tau]|^2.
\end{aligned}$$

Note that the $[-\pi, \pi)$ interval choice for $\phi_k[\tau]$ is arbitrary, since this result holds for any $2\pi-$length interval. Finally, let us define $S_k[n] \doteq |\mathbf{s}_k[n]|^2$, $H_k[n] \doteq |\mathbf{h}_k[n]|^2$ and $X_k[n] \doteq E|\tilde{\mathbf{x}}_k[n]|^2$. Then, our model reads

$$X_k[n] = \sum_{\tau} S_k[n-\tau]H_k[\tau], \tag{3}$$

and the square magnitude of the observed spectrogram components can be written as

$$Y_k[n] = X_k[n] + \epsilon_k[n], \tag{4}$$

where $\epsilon_k[n]$ denotes the representation error. As shown in [16], this model is equivalent to a convolutive NMF ([25]) with diagonal basis. In the next section, we derive a cost function in order to find an appropriate convolutive representation that allows us to isolate the components $S_k[n]$.

3 A Bayesian approach

In the following, we will use a Bayesian approach to derive a cost function which we will then minimize in order to obtain our regularized solution. Let us begin by assuming, for every k, $\epsilon_k[n], S_k[n], H_k[n]$ are independent random variables, also independent with respect to k. Also, let us denote by $S, Y, X \in \mathbb{R}^{K\times N}$ and $H \in \mathbb{R}^{K\times N_h}$ the non-negative matrices whose (k,n)-th elements are $S_k[n], Y_k[n], X_k[n]$ and $H_k[n]$, respectively.

More often than not, some type of "patterns" can be observed in a speech spectrogram, mainly due to the harmonics of speech (see Figure 1). However, they seem to be strongly speaker and phoneme dependent, and although it would be interesting to try to model this correlation, this is not viable in a blind setting (since no a-priori information is available for estimating it). Besides, it is worth mentioning that the frequency independency assumption has shown to lead to quite good results.

As it is customary ([16]), for the representation error, we assume $\epsilon_k[n] \sim \mathcal{N}(0, \sigma_k^2)$, where $\sigma_k > 0$ is an unknown parameter, and the variables are non-correlated with respect to n. Hence, it follows from (4) that the conditional distribution of Y given S and H (i.e. the likelihood) is given by

$$\pi_{like}(Y|S,H) = \prod_{k=1}^{K}\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(Y_k[n]-X_k[n])^2}{\sigma_k^2}\right).$$

Note that, strictly speaking, in the above model for the representation error, the non-negativity constraint on the components of Y is not enforced. This is done mainly for simplicity reasons. It is rooted in the fact that this distribution provides a good model for the data Y; thus, the probability of one of its components be negative is very small, and enforcing non-negativity would unnecessarily complicate the model.

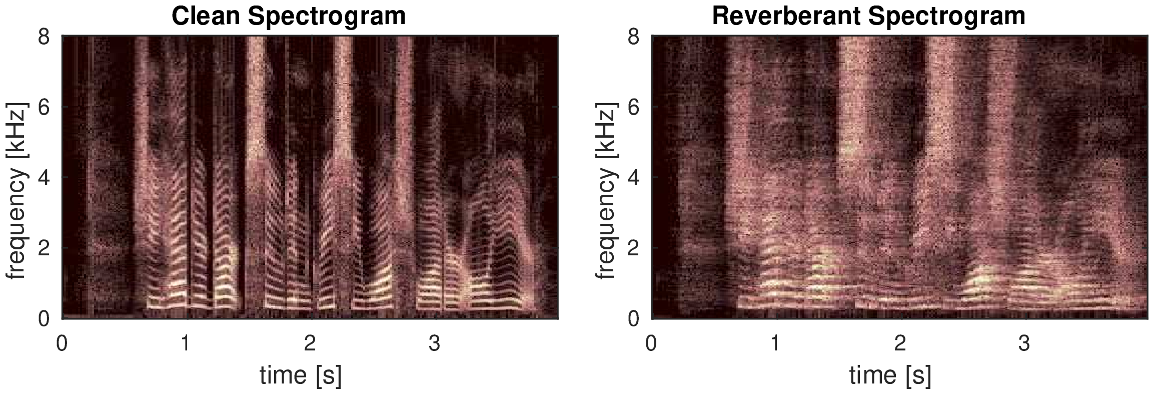


Figure 1: Spectrograms for a clean speech signal (left) and the corresponding reverberant speech signal (right). The clean signal, from the TIMIT database, was sampled at 16 [kHz], and corresponds to a female voice uttering the sentence 'She had your dark suite in greasy wash water all year.' The signal was artificially made reverberant by convolution with a room impulse response, with a reverberation time of 600 [ms], to produce the reverberant spectrogram. Both spectrograms were made using Hamming windows with 512 samples and an overlapping of 256.

Let us now turn our attention to S. Figure 1 depicts the log-spectrograms for a clean signal and its reverberant version. As it can be observed, while the spectrogram of the clean signal is somewhat sparse, the one corresponding to the reverberant signal presents a smoother or more diffuse structure. The presence of discontinuities in the spectrogram of the clean signal can be favored by assuming S follows a generalized non-negative Gaussian distribution ([26]). Thus,

$$\pi_{prior}(S) = \begin{cases} \prod_{k=1}^{K} \prod_{n=1}^{N} \frac{1}{\Gamma(1+1/p)b_k} \exp\left(-\frac{S_k[n]^p}{b_k^p}\right) & S_k[n] \geq 0, \\ 0 & S_k[n] < 0, \end{cases}$$

where $p \in (0,2)$ is a prescribed parameter and $b_k > 0$ is unknown.

In regards to H, although no general conditions are expected on its individual components, we do expect its first order time differences to exhibit a certain degree of regularity (see Figures 2 and 3). It can be observed that the log-spectrograms consist of a high-energy vertical band to the left, that corresponds to the linear impulse response, and some straight lines of less energy that correspond to the non-linear distortions produced by the increase on the rate at which the echoes reach the receiver ([27]). In fact, if windows are set close enough relative to the duration of the reverberation phenomenon, then consecutive time components of H will capture overlapped information, which along with the exponential decay characteristic of the RIR ([28]) accounts for a somewhat smooth structure. Therefore, we define the time differences matrix $V \in \mathbb{R}^{K\times(N_h-1)}$, with components $V_k[n] \doteq H_k[n] - H_k[n-1]$ $\forall n = 1, \ldots, N_h - 1$, $k = 1, \ldots, K$. The regularity of these variations is contemplated by assuming V follows a normal distribution with zero mean and variance η_k^2:

$$\pi_{prior}(V) = \prod_{k=1}^{K} \prod_{n=2}^{N_h} \frac{1}{\sqrt{2\pi}\eta_k} \exp\left(-\frac{V_k[n]^2}{\eta_k^2}\right).$$

Let $H_k \in \mathbb{R}^{N_h}$ be the transpose k^{th}-row of H, $L \in \mathbb{R}^{N_h-1\times N_h}$ be the matrix such that $LH_k = V_k$ and $\pi_{prior}(H)$ the prior induced from $\pi_{prior}(V)$ through this relation. Using Bayes' theorem, the *a*

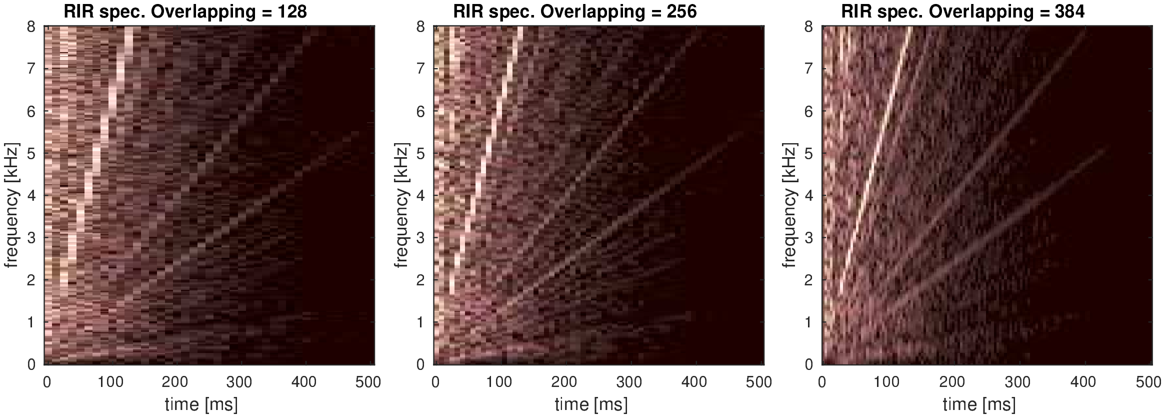


Figure 2: Log-spectrograms for an artificial 16 [kHz] RIR signal with reverberation time of 600 [ms]. The spectrograms were made using a hamming window length of 512 and different overlappings.

posteriori joint distribution of S and H conditioned to Y satisfies

$$\pi_{post}(S,H|Y) \propto \pi_{like}(Y|S,H)\pi_{prior}(S)\pi_{prior}(H). \tag{5}$$

Our goal is to find $\hat{S}$ and $\hat{H}$ that are representative of the *a posteriori* distribution (5). Although the immediate instinct might be to compute the expected value, there are quite a few other ways to proceed, with different degrees of reliability and complexity. In the light of the assumed distributions and the high dimensionality of the problem, the *maximum a posteriori* (MAP) estimator is a reasonable choice in this case. Note that maximizing (5) is tantamount to minimizing $-\log \pi_{post}(S,H|Y)$. If we denote by $S_k, Y_k, X_k \in \mathbb{R}^N$, $H_k \in \mathbb{R}^{N_h}$ and $V_k \in \mathbb{R}^{N_h-1}$ the (transposed) rows of S, Y, X, H and V, then

$$\begin{aligned} J(S,H) &\doteq -\log \pi_{post}(S,H|Y) \qquad (6)\\ &= \sum_{k=1}^{K}\left[\frac{1}{\sigma_k^2}||Y_k - X_k||_2^2 + \frac{1}{b_k^p}\sum_n S_k[n]^p + \frac{1}{\eta_k^2}||LH_k||_2^2\right] + C, \end{aligned}$$

where C is a constant independent of S and H. Our goal is to minimize J, subject to the non-negativity restrictions $S_k[n] \geq 0\, \forall k = 1,\ldots,K,\ n = 1,\ldots,N\ ,\ H_k[n] \geq 0\, \forall k = 1,\ldots,K,\ n = 1,\ldots,N_h$.

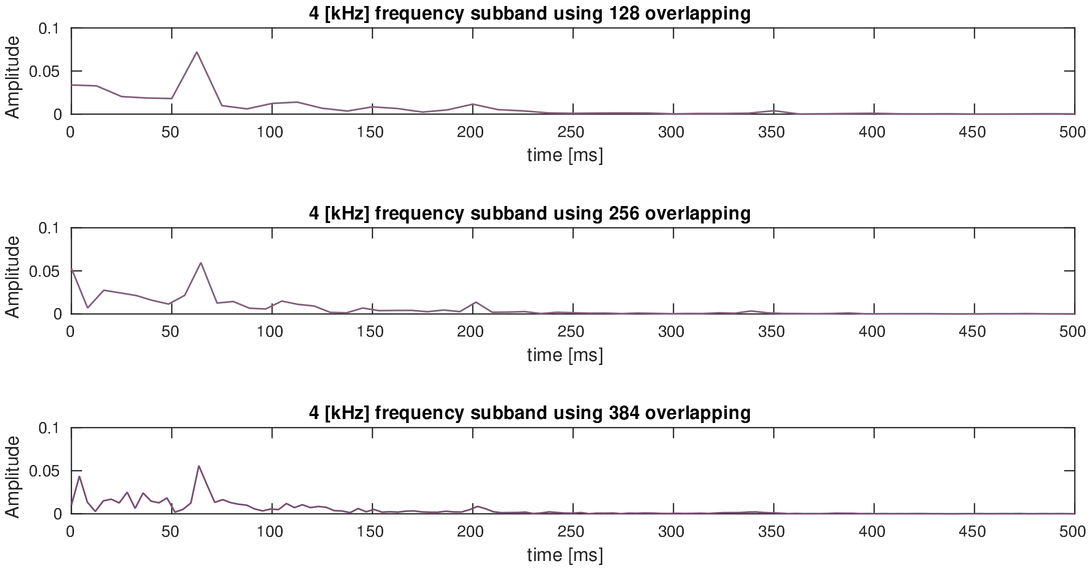
Although it is likely that different frequency sub-bands be affected differently by the RIR, with the reverberant spectrogram being the only available data for a blind approach, there will always be an arbitrary frequency dependent scaling ambiguity. In this way, it is impossible to exactly recover the original scaling of the source. Since given this fundamental indeterminacy, any frequency bin amplitude would be arbitrary in some sense, we have imposed the constraint $||S_k||_\infty = ||Y_k||_\infty\ \forall k$, which means that the maximum values shall remain equal for every frequency bin (this is similar to the minimum distortion principle ([29]) applied in frequency domain blind source separation). Additionally, we have experimentally found this constraint to be adequate.

3.1 Model parameters

Before proceeding to minimize equation (6), some comments on the model parameters $\{\sigma_k, b_k, \eta_k, p\}_{k=1,\ldots,K}$ are in order.

The value of the exponent $p \in (0,2)$ is related to the degree of sparsity of S. While small values of p will promote high sparsity, choosing $p \approx 2$ will yield low sparsity.

Notice that for any given $k \in \{1,\ldots,K\}$, the variance of the representation error is proportional to the energy (the square of the L^2-norm) of the corresponding frequency sub-band. That is, we


4 [kHz] frequency subband using 128 overlapping

4 [kHz] frequency subband using 256 overlapping

4 [kHz] frequency subband using 384 overlapping

Figure 3: Signals corresponding to the 4[kHz] frequency subband of RIR spectrograms $H_{129}[n],\ n = 1, \dots, N$, with window length 512 and different overlappings. The sampling frequency is of 16[kHz] and the reverberation time is 600 [ms]. The signals show certain regularity, which increases with the window overlapping.

choose $\sigma_k^2 \doteq \sigma_0^2 \|Y_k\|^2$, where σ_0 is a constant independent of k. In a similar fashion, we choose $b_k \doteq b_0 \|Y_k\|$. Finally, since we have no evidence of any relationship between the frequency sub-band and the variations of H, we choose $\eta_k \doteq \eta_0$, independent of the frequency bin. Furthermore, since the functional (6) can be minimized separately in each frequency bin, the selection of the parameters is simplified by first choosing p and then the ratios σ_0^2 / b_0^p and σ_0^2 / η_0^2.

4 Hypermodel approach

To better deal with uncertainty on some of the parameter values, the previous model can be extended to a hypermodel by considering those parameters as random variables. For instance, due to the aforementioned uncertainty on the variance of H, we shall assume that the standard deviations of H_k, $\eta_k > 0, k = 1, \dots, K$, are realizations of *i.i.d.* random variables with gamma distribution. That is,

$$\pi_{hyper}(\eta_k) \doteq \frac{\eta_k^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} \exp\left(-\frac{\eta_k}{\beta}\right),$$

where $\alpha > 1$ and $\beta > 0$ are shape and scale parameters, respectively. Using this hyperprior, the new functional (the negative logarithm of the *a-posteriori* distribution) turns out to be:

$$\begin{aligned} J_{hyp}(S, H, \eta) &\doteq -\log \pi_{post}(S, H, \eta | Y) \\ &= \sum_{k=1}^{K} \left[\frac{1}{\sigma_k^2} ||Y_k - X_k||_2^2 + \frac{1}{b_k^p} \sum_n S_k[n]^p + \frac{1}{\eta_k^2} ||L H_k||_2^2 \right] \\ &+ \sum_{k=1}^{K} \left[(N_h + 1 - \alpha) \log \eta_k + \frac{\eta_k}{\beta} \right] + C, \end{aligned} \tag{7}$$

where η denotes the vector whose components are $\eta_k, k = 1, \dots, K$ and C is a constant independent of S, H, and η.

In what follows, we focus on minimizing the functionals J and J_{hyp} defined by (6) and (7), respectively.

5 Iterative minimization algorithms

5.1 Minimizing J

We begin by introducing a method for minimizing J, defined in (6). Later on, we will show that by adding an extra step, the same method can be used for minimizing J_{hyp}.

5.1.1 Auxiliary functions

The algorithm is constructed based on an auxiliary function technique, following similar ideas as those in [16]. Minimization procedures based in this kind of techniques are also known as Majorization-Minimization algorithms ([30]).

Let $\Omega \subset \mathbb{R}$ and $f : \Omega \to \mathbb{R}_0^+$. Then, $g : \Omega \times \Omega \to \mathbb{R}_0^+$ is called an *auxiliary function* for f if

$$(i)\ g(w, w) = f(w) \ \text{ and } \ (ii)\ g(w, w') \geq f(w), \ \ \forall w, w' \in \Omega. \tag{8}$$

Let $w^0 \in \Omega$ be arbitrary and let

$$w^j \doteq \arg\min_w g(w, w^{j-1}). \tag{9}$$

With this definition, it can be shown ([31]) that the sequence $\{f(w^j)\}_j$ is non-increasing. We intend to use this property as a tool for alternatively updating the matrices H and S. Let us begin by fixing $H = H'$, where H' is an arbitrary $K \times N_h$ matrix. Then, an auxiliary function for $J(S, H')$ (as defined in (6)) with respect to S is given by

$$\begin{aligned} g_s(S, S') \doteq & \sum_{k,n,\tau} \frac{1}{\sigma_k^2} \frac{S'_k[\tau] H'_k[n-\tau]}{X'_k[n]} \left(Y_k[n] - \frac{S_k[\tau]}{S'_k[\tau]} X'_k[n] \right)^2 + \sum_k \frac{1}{\eta_k^2} ||L H'_k||_2^2 \\ & + \sum_{k,n} \frac{1}{b_k^p} \left(\frac{p}{2} S'_k[n]^{p-2} S_k[n]^2 + S'_k[n]^p - \frac{p}{2} S'_k[n]^p \right), \end{aligned} \tag{10}$$

where $X'_k[n] = \sum_\tau S'_k[n-\tau] H'_k[\tau]$. The proof can be found in A.

In an analogous way, it can be shown that if we let $S = S'$ be fixed, where S' is an arbitrary $K \times N$ matrix, then

$$\begin{aligned} g_h(H, H') \doteq & \sum_{k,n,\tau} \frac{1}{\sigma_k^2} \frac{S'_k[n-\tau] H'_k[\tau]}{X'_k[n]} \left(Y_k[n] - \frac{H_k[\tau]}{H'_k[\tau]} X'_k[n] \right)^2 \\ & + \sum_k \frac{1}{b_k^p} ||S'_k||_p^p + \sum_k \frac{1}{\eta_k^2} ||L H_k||_2^2 \end{aligned}$$

is an auxiliary function for $J(S', H)$ with respect to H.

Having defined auxiliary functions, we will use the updating rule derived from (9) to build an algorithm for iteratively updating matrices S and H in order to minimize J. Notice this requires minimizing g_s and g_h with respect to the updating variables, but since g_s is quadratic with respect to S and g_h is quadratic with respect to H, we can simply use the first order necessary conditions in both cases. From this point on, in the context of the iterative updating process, S' and H' will refer not to arbitrary nonnegative matrices, but to those estimations of S and H obtained in the immediately previous step.

5.1.2 Updating rule for S

Firstly, we shall derive an updating rule for $S_k[\tau]$. That is, we wish to minimize g_s w.r.t. S. The first order necessary condition on g_s yields

$$
\begin{aligned}
0 =& \frac{\partial g_s(S,S')}{\partial S_k[\tau]} \\
=& -2\sum_n \frac{1}{\sigma_k^2} H'_k[n-\tau]\left(Y_k[n] - \frac{S_k[\tau]}{S'_k[\tau]} X'_k[n]\right) + \frac{p}{b_k^p} S'_k[\tau]^{p-2} S_k[\tau] \\
=& -\sum_n H'_k[n-\tau] Y_k[n] + \frac{S_k[\tau]}{S'_k[\tau]} \sum_n H'_k[n-\tau] X'_k[n] + \frac{p\sigma_k^2}{2b_k^p} S'_k[\tau]^{p-2} S_k[\tau] \\
=& -S'_k[\tau] \sum_n H'_k[n-\tau] Y_k[n] + \left(\sum_n H'_k[n-\tau] X'_k[n] + \frac{p\sigma_k^2}{2b_k^p} S'_k[\tau]^{p-1}\right) S_k[\tau],
\end{aligned}
$$

which easily leads to the multiplicative updating rule

$$
S_k[\tau] = S'_k[\tau] \frac{\sum_n H'_k[n-\tau] Y_k[n]}{\sum_n H'_k[n-\tau] X'_k[n] \; + \; \frac{p\sigma_k^2}{2b_k^p} S'_k[\tau]^{p-1}}.
$$

In order to avoid the aforementioned scale indeterminacy, every updating step is to be followed by scaling S_k so that its ℓ^∞ norm coincides with that of the observation Y_k.

5.1.3 Updating rule for H

In order to state an updating rule for H, we begin by defining the diagonal matrices $A^k, B^k \in \mathbb{R}^{N_h \times N_h}$, whose diagonal elements are $A^k_{\tau,\tau} \doteq \sum_n S'_k[n-\tau] X'_k[n]$ and $B^k_{\tau,\tau} \doteq H'_k[\tau]$, and the vector $\zeta^k \in \mathbb{R}^{N_h}$ with components $\zeta^k_\tau = \sum_n S'_k[n-\tau] Y_k[n]$.

It can be shown (see B) that with these definitions, H can be updated by solving the linear system

$$
\left(A^k + \frac{\sigma_k^2}{\eta_k^2} B^k L^{\mathrm{T}} L\right) H_k = B^k \zeta^k. \tag{11}
$$

Let us notice that under the assumption that the diagonal elements of A^k and B^k are strictly positive, and since $L^{\mathrm{T}}L$ is positive-semidefinite, $(B^k)^{-1}A^k + \lambda_{h,k} L^{\mathrm{T}}L$ is positive-definite, and hence the linear system has a unique solution. Furthermore, this implies that the solution is non-negative. The assumption of $A^k_{\tau,\tau} > 0$ is adequate, since these elements correspond to the discrete convolution of S'_k and X'_k. Although the validity of the hypothesis over $B^k_{\tau,\tau}$ is not so clear, in practice, the matrix in system (11) has turned out to be non-singular. Nonetheless, H_k can be computed as the best approximate solution in the least-squares sense. Solving this $N_h \times N_h$ linear system entails no challenge, since N_h is usually chosen relatively small, depending on the window step and the reverberation time.

5.2 Minimizing J_{hyp}

It follows immediately from the fact that the additional terms on equation (7) with respect to equation (6) do not depend on S nor H, that the minimization steps derived for J are suitable for J_{hyp} as well. Thus, it only remains to minimize J_{hyp} with respect to η, which can be shown (see C) to be equivalent to solving the following equation:

$$
\eta_k^3 + (N_h + 1 - \alpha)\beta\, \eta_k^2 - 2\beta ||LH_k||_2^2 = 0,
$$

for every $k = 1, \ldots, K$. This can be done either explicitly by means of the general solution of the cubic equation, or by an appropriate iterative method.

5.3 Final considerations

All steps of the dereverberation process are stated in Algorithm 1. The updating step in line 22 only concerns functional J_{hyp}, and it must be skipped when minimizing J.

Algorithm 1 Bayesian dereverberation

```
Initializing
$S \leftarrow Y$
$H_k[n] \leftarrow \exp(-n) \quad \forall k = 1 \dots K,\ n = 1 \dots N$
MAIN LOOP
for $i = 1 \dots \text{maxiter}$
  $X_k[n] \leftarrow \sum_\tau S_k[n-\tau] H_k[\tau] \quad \forall k = 1 \dots K,\ n = 1 \dots N$

  for $k = 1 \dots K$
    for $\tau = 1 \dots N$
      $S_k[\tau] \leftarrow S_k[\tau] \dfrac{\sum_n H_k[n-\tau] Y_k[n]}{\sum_n H_k[n-\tau] X_k[n] \ + \ \frac{p\sigma_k^2}{2b_k^p} S_k[\tau]^{p-1}}.$
    end for
    $S_k \leftarrow S_k \dfrac{\|Y_k\|_\infty}{\|S_k\|_\infty}.$
  end for

  for $k = 1 \dots K$
    Build the diagonal matrices$A^k, B^k \in \mathbb{R}^{N_h \times N_h}$ :
      $A^k_{\tau,\tau} = \sum_n S_k[n-\tau] X_k[n],$
      $B^k_{\tau,\tau} = H_k[\tau].$
    Build the vector $\zeta^k$ :
      $\zeta^k_\tau = \sum_n S_k[n-\tau] Y_k[n]$
    Solve for $H_k$ :
      $(A^k + \dfrac{\sigma_k^2}{\eta_k^2} B^k L^{\mathrm{T}} L) H_k = B^k \zeta^k.$

    if Using the hypermodel ($J_{hyp}$)
      Solve for $\eta_k$ :   $\eta_k^3 + (N_h + 1 - \alpha)\beta\, \eta_k^2 - 2\beta ||L H_k||_2^2 = 0.$
    end if
  end for

  if $\|S - S'\|_F \leq \delta$
    return
  end if

end for
```

In the Initialization Step we define the clean spectrogram S equal to the observation, which is natural since in a way they both correspond to the same signal, and H_k as a vector with exponential time decay, which is an expected characteristic of a RIR. Note that with this initialization all the variables result non-negative. Under this condition, it is easy to see that all the updating rules maintain non-negativitiy, thus complying with the aforementioned restrictions $S_k[n] \geq 0\,\forall k = 1,\ldots,K$, $n = 1,\ldots,N$, and $H_k[n] \geq 0\,\forall k = 1,\ldots,K$, $n = 1,\ldots,N_h$..

Finally, we set the stopping criterion over the decay of the norm of two consecutive approximations of S. This has shown to work quite well, although other stopping criteria might be considered.

Results to illustrate the performance of the proposed algorithms are presented in the next section.

6 Experimental results

For the experimental results we used both simulated and recorded reverberant signals. While a large number of artificially reverberant signals were produced to get statistically significant results, recorded signals were used to corroborate the performance of the methods using real data.

6.1 Experiments with simulations

For the experiments, we took 110 speech signals from the TIMIT database ([32]), recorded at 16 kHz, and artificially made them reverberant by convolution with impulse responses generated with the software Room Impulse Response Generator[1], based on the model in [33]. Each signal was degraded under different reverberation conditions: three different room sizes, each with three different microphone positions and four different reverberation times, which gives us a total of 3960 signals for testing. Table 1 gives account of the room dimensions and source and microphone positions that were chosen.[2]

	Length	Width	Height
Room 1 dimensions	5.00 [m]	4.00 [m]	6.00 [m]
Room 2 dimensions	4.00 [m]	4.00 [m]	3.00 [m]
Room 3 dimensions	10.0 [m]	4.00 [m]	5.00 [m]
Source position	2.00 [m]	3.50 [m]	2.00 [m]
Microphone 1 position	2.00 [m]	1.50 [m]	1.00 [m]
Microphone 2 position	2.00 [m]	2.00 [m]	1.00 [m]
Microphone 3 position	2.00 [m]	2.00 [m]	2.00 [m]

Table 1: Simulated room settings

In order to avoid preprocessing, the choice of the probabilistic model parameters was made *a priori* by means of empirical rules, based upon signals from a different database. This is supported by the fact that the parameters were observed to be rather robust with respect to variations of the reverberation conditions, and hence they were chosen simply as $\sigma_k^2 = \|Y_k\|^2$, $\eta_k = 1$ and $b_k = \|Y_k\| \times 10^7$. For the case of minimizing functional J_{hyp}, we set $\alpha = 10^2$ and $\beta = 10^{-2}$, so the expected value for η_k is $\alpha\beta = 1$, for the comparison between the Bayesian model and Hypermodel to be fair. The rest of the model parameters were chosen as specified in Table 2.

p	N_h	win.	window size	win. overlap.	δ	max. iter.
1	15	Ham.	512 samples	256 samples	$\|Y\|_F \times 10^{-3}$	20

Table 2: Model parameter values

Let us point out that the choice of N_h was done as to allow H to capture early reverberation while precluding overlapped representations. In the first place, it is desirable for H to represent the RIR

[1] https://github.com/ehabets/RIR-Generator

[2] A web demo can be found in sinc.unl.edu.ar/web-demo/blindder/

along the full Early Decay Time (EDT), the time period in which the reverberation phenomenon alters the clean signal the most, so its effect can be nullified. On the other hand, if we were to choose N_h too large, it might lead certain similarities in the observation Y within a fixed frequency range to be represented as echoes from high energy components of S. It is worth mentioning, however, that the performance of our dereverberation method has shown no high sensitivity with respect to the choice of N_h.

In order to evaluate the performance of our models, using both functionals J and J_{hyp}, we made comparisons against three state-of-the-art methods that work under the same conditions. Two of the methods we used were those proposed by Kameoka *et al* in [16] and the mixed penalization method proposed in [17], which are not only recent but in a sense precursors to the method proposed in this article. Also, we included the method proposed by Wisdom *et al* in [12], with a window length of 2048, because of its great performance in the Reverb Challenge ([34]).

To measure performance, following [35], we made use of the frequency weighted segmental signal-to-noise ratio (fwsSNR) and cepstral distance. Furthermore, we also measured the speech-to-reverberation modulation energy ratio (SRMR, [36]), which has the advantage of being non-intrusive (it does not use the clean signal as an input). The results for each performance measure are stated in Table 3, and depicted in Figures 4- 6, classified in function of the reverberation times: 300[ms], 450[ms], 600[ms] and 750[ms]. Notice that for the cases of fwsSNR and SRMR, higher values correspond to better performance, while for the cepstral distance, small values indicate higher quality.

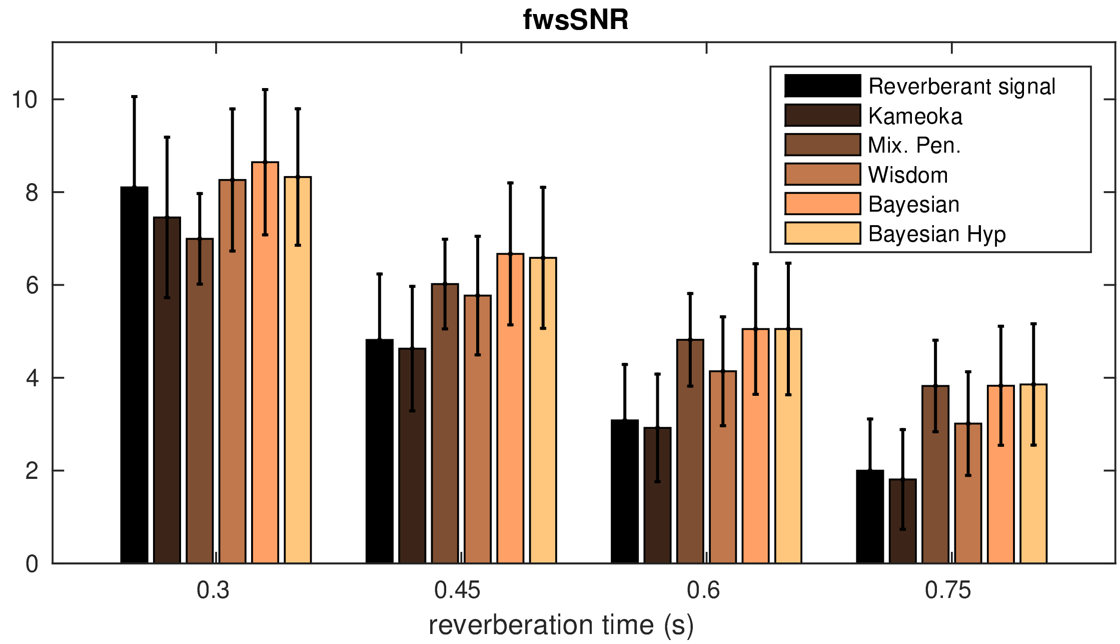


Figure 4: Mean and standard deviations of fwsSNR for different reverberation times.

Table 3 shows that the results obtained using the Bayesian methods with functionals J and J_{hyp} are significantly better ($p < 0.01$) than those produced by the other methods for all the considered performance measures. Also, Figures 4-6 clearly show that in all cases the improvement is more evident for larger reverberation times, specially for the fwsSNR and the Cepstral Distance. Furthermore, Figure 5 shows that no competing method is able to reduce the Cepstral Distance for a reverberation time of 300[ms]. This most likely occurs because the reverberation time is too short and therefore the introduced distortion, when doing dereverberation, cancels out the potential gains. Yet, for larger reverberation times, our method does produce a significant improvement as measured by the Cepstral

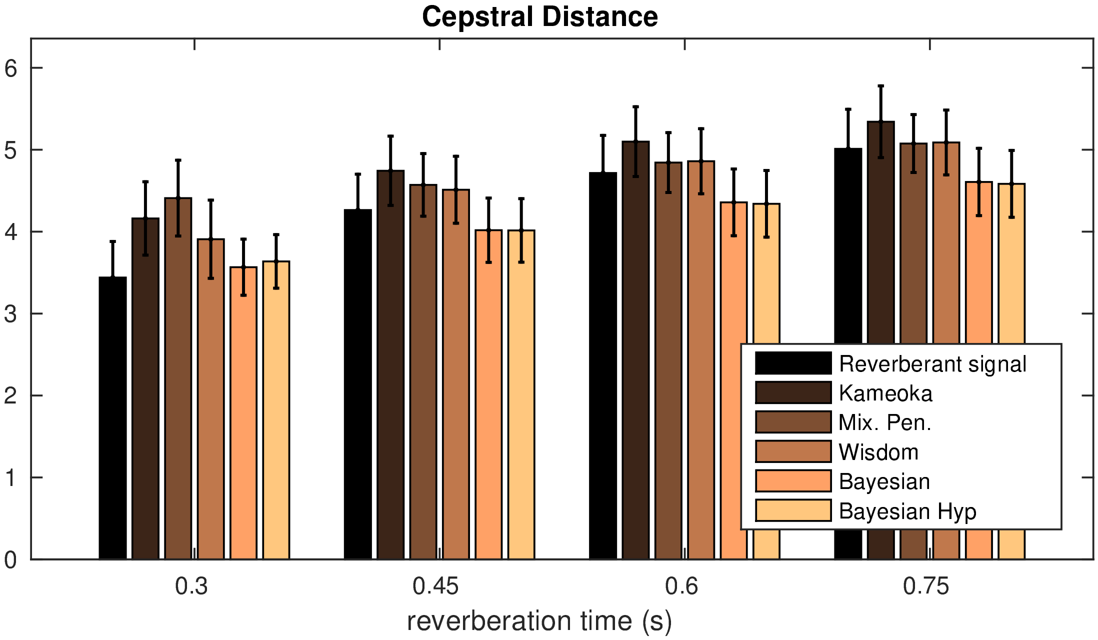


Figure 5: Mean and standard deviations of Cepstral Distance for different reverberation times.

Distance. It is timely to mention that all the differences between the performance of our methods and every competing one hold statistical significance ($p < 0.01$) for every reverberation time (as depicted in Figures 4-6), with the only exception of the SRMR with a 300[ms] reverberation time, where our methods produce no significant improvement with respect to Wisdom's.

6.2 Experiments with recorded signals

For this experiment we have used real recordings obtained in our own office rooms, with a sampling frequency of 16[kHz]. Two male and two female speakers were randomly selected from the TIMIT database, and 10 speech signals for each were played in two different rooms. The dimensions of the fully furnished rooms and microphone positions are specified in Table 4. The reverberation times, measured using sine sweeps ([37]), were found to be 460[ms] on the first room and 440[ms] on the second. It is timely to mention that for the recordings to be realistic, they were made during standard office hours, with people working in nearby offices (although no people were present in the recording room), and some of the computers and air conditioning were left on.

The model parameters were chosen equal to those used for the experiment with simulations, except for the variance of the distribution of S, that was changed to cope with the considerably high noise level. The new choice was simply $b_k = 10\|S_k\|/\sigma_n$, where σ_n is the standard deviation of the noise, estimated from the first 1000 samples (61[ms]) of the recordings. The parameters for the competing methods were properly adjusted to the noise level as well.

Results are depicted in Table 5. Once again, we see that the Bayesian methods outperform the other methods in terms of the fwsSNR and SRMR, although Wisdom's method performs slightly better (but not significantly, $p > 0.01$) in terms of Cepstral Distance.

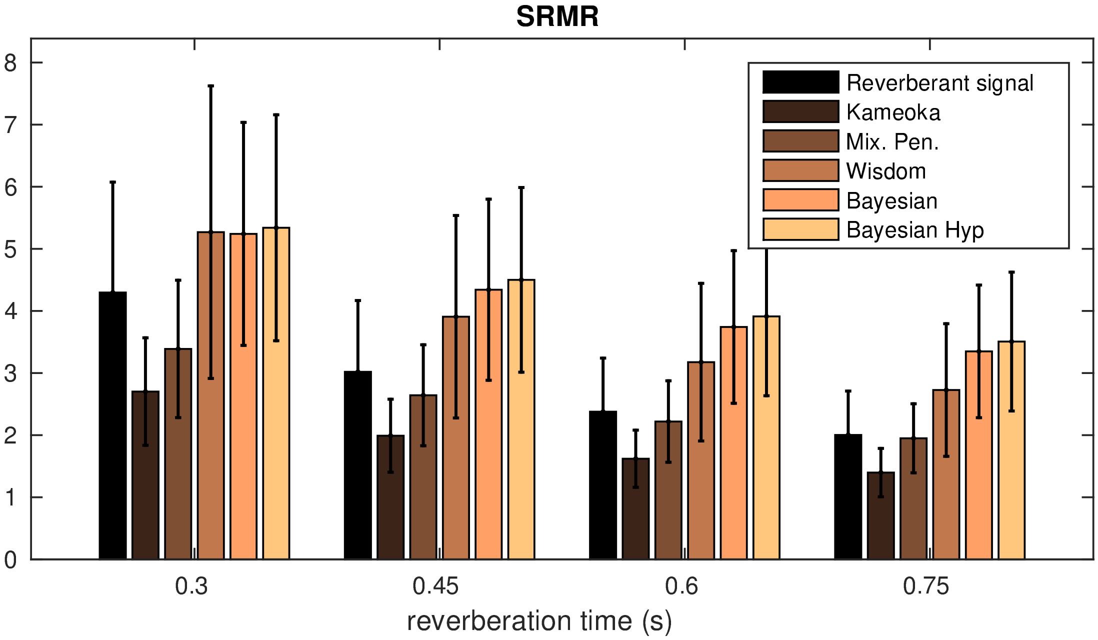


Figure 6: Mean and standard deviations of SRMR for different reverberation times.

Measure	fwsSNR	Cepstral Dist.	SRMR
Reverberant	5.411 (3.23)	5.521 (0.87)	2.755 (0.75)
Kameoka	6.041 (3.19)	5.125 (0.68)	2.126 (0.48)
Mixed Pen	7.089 (3.19)	5.735 (0.79)	2.45 (0.58)
Wisdom	6.241 (3.60)	**4.640** (0.51)	3.227 (0.77)
Bayesian	8.608 (2.83)	4.839 (0.47)	4.860 (1.13)
Hypermodel	**8.660** (2.92)	4.824 (0.41)	**4.878** (1.14)

Table 5: Mean and standard deviation (between parenthesis) of performance measures for each method. Best results are shown in boldface.

6.3 Computing performance

Finally, we also compared the computing performance of the aforementioned methods using the TIMIT database of the first experiment. The examples were run using MatLab in a PC with an Intel Core i7-2600k CPU @3.4GHz×8, with 8Gb of RAM. The CPU-times for each method are depicted in Table 6, where it can be seen that although not as fast as the Mixed Penalization method, it is twice as fast as the closest competing method in terms of restoration quality. Finally, it is appropriate to mention that the speed of our method could be further improved using parallel computing. This is due to the fact that in our algorithm (just as in Kameoka's) the minimization can be performed simultaneously in every frequency bin.

Method	Kameoka	Mixed Pen	Wisdom	Bayesian	Hyper.
CPU time	7.61[s]	4.15 [s]	11.14[s]	5.47[s]	5.58[s]

Table 6: Mean CPU time for dereverberation with each algorithm.

Measure	fwsSNR	Cepstral Dist.	SRMR
Reverberant	4.499 (2.73)	4.358 (0.75)	2.924 (1.48)
Kameoka	4.203 (2.52)	4.836 (0.62)	1.928 (0.78)
Mixed Pen	5.414 (1.55)	4.723 (0.47)	2.550 (0.98)
Wisdom	5.296 (2.35)	4.592 (0.61)	3.770 (1.91)
Bayesian	**6.048** (2.32)	**4.137** (0.55)	4.168 (1.58)
Hypermodel	5.954 (2.20)	4.144 (0.52)	**4.315** (1.60)

Table 3: Mean and standard deviation (between parenthesis) of performance measures for each method, using simulations. Best results are shown in boldface.

	Length	Width	Height
Room 1 dimensions	4.15 [m]	3.00 [m]	3.00 [m]
Source 1 position	3.60 [m]	1.50 [m]	1.50 [m]
Microphone 1 position	1.10 [m]	1.50 [m]	1.50 [m]
Room 2 dimensions	5.85 [m]	4.55 [m]	3.00 [m]
Source 2 position	1.10 [m]	1.50 [m]	1.50 [m]
Microphone 2 position	1.10 [m]	4.00 [m]	1.50 [m]

Table 4: Office rooms settings

7 Conclusions

In this work a new blind dereverberation method for speech signals based on a Bayesian approach over a convolutive NMF representation of the spectrograms was introduced and tested. This includes a basic Bayesian model as well as a model with hyperpriors.

Results show the new introduced method is faster and outperforms the others in terms of fwsSNR and SRMR, and, moreover, it is comparable to the best of those in terms of Cepstral Distance. A significant improvement in performance stands out for high reverberation times.

It is also worth mentioning that the proposed algorithm results fast enough to be considered for performing on-line dereverberation, endeavor that we plan to engage on in future work.

There is certainly much room for further improvement. Among others, the use of other prior distributions depending on *a-priori* information, the introduction of time variability, and exploring the use of other time-frequency representations analogous to STFT that could help to improve the obtained restorations.

Acknowledgements

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET through PIP 2014-2016 N° 11220130100216-CO, the Air Force Office of Scientific Research, AFOSR/SOARD, through Grant FA9550-14-1-0130, by Universidad Nacional del Litoral, UNL, through CAID-UNL 2011 N° 50120110100519 "Procesamiento de Señales Biomédicas" and CAI+D-UNL 2016, PIC 50420150100036LI "Problemas Inversos y Aplicaciones a Procesamiento de Señales e Imágenes".

A Proof of the fact that g_s is an auxiliary function for J

We want to prove that g_s, defined as in (10), is an auxiliary function for J, defined in (6). That is, we must show that g_s complies with both conditions stated in (8) .

The equality condition (i) is rather straightforward. In fact,

$$\begin{aligned}
g_s(S,S) =& \sum_{k,n,\tau} \frac{1}{\sigma_k^2} \frac{S_k[\tau]H'_k[n-\tau]}{\sum_\nu S_k[\nu]H'_k[n-\nu]} \left(Y_k[n] - \frac{S_k[\tau]}{S_k[\tau]} \sum_\nu S_k[\nu]H'_k[n-\nu]\right)^2 \\
&+ \sum_k \frac{1}{\eta_k^2} ||LH'_k||_2^2 + \sum_{k,n} \frac{1}{b_k^p} \left(\frac{p}{2} S_k[n]^{p-2} S_k[n]^2 + S_k[n]^p - \frac{p}{2} S_k[n]^p\right) \\
=& \sum_{k,n,\tau} \frac{1}{\sigma_k^2} \frac{S_k[\tau]H'_k[n-\tau]}{\sum_\nu S_k[\nu]H'_k[n-\nu]} \left(Y_k[n] - \sum_\nu S_k[\nu]H'_k[n-\nu]\right)^2 \\
&+ \sum_k \frac{1}{\eta_k^2} ||LH'_k||_2^2 + \sum_{k,n} \frac{1}{b_k^p} S_k[n]^p \\
=& \sum_{k,n} \frac{1}{\sigma_k^2} \left(Y_k[n] - \sum_\nu S_k[\nu]H'_k[n-\nu]\right)^2 + \sum_k \frac{1}{\eta_k^2} ||LH'_k||_2^2 + \sum_{k,n} \frac{1}{b_k^p} S_k[n]^p \\
=& J(S,H').
\end{aligned}$$

To prove condition (ii) in (8) we begin by defining

$$\begin{aligned}
P_{k,n} \doteq& \sum_\tau \frac{S'_k[\tau]H'_k[n-\tau]}{X'_k[n]} \left(Y_k[n] - \frac{S_k[\tau]}{S'_k[\tau]} X'_k[n]\right)^2, \\
R_{k,n} \doteq& (Y_k[n] - \sum_\tau S_k[\tau]H'_k[n-\tau])^2,
\end{aligned}$$

and $Q : \mathbb{R}^+ \to \mathbb{R}$ such that $Q(x) \doteq \frac{p}{2} x^{p-2} S_k[n]^2 + x^p - \frac{p}{2} x^p$. With these definitions, we can write

$$g_s(S,S') = \sum_k \left(\sum_n \left(\frac{1}{\sigma_k^2} P_{k,n} + \frac{1}{b_k^p} Q(S'_k[n]) \right) + \frac{1}{\eta_k^2} ||LH'_k||_2^2 \right),$$

and

$$J(S,H') = \sum_k \left(\sum_n \left(\frac{1}{\sigma_k^2} R_{k,n} + \frac{1}{b_k^p} S_k[n]^p \right) + \frac{1}{\eta_k^2} ||LH'_k||_2^2 \right).$$

Hence, to prove that $g_s(S,S') \geq J(S,H')$ $\forall S, S'$ it is sufficient to show that $P_{k,n} \geq R_{k,n}$ and $Q(S'_k[n]) \geq S_k[n]^p$ $\forall n = 1, \dots, N, k = 1, \dots, K$. In fact,

$$\begin{aligned}
P_{k,n} - R_{k,n} =& \sum_\tau \frac{S'_k[\tau]H'_k[n-\tau]}{X'_k[n]} \left(Y_k[n] - \frac{S_k[\tau]}{S'_k[\tau]} X'_k[n]\right)^2 \\
&- (Y_k[n] - \sum_\tau S_k[\tau]H'_k[n-\tau])^2 \\
=& \sum_\tau \frac{H'_k[n-\tau]S_k[\tau]^2 X'_k[n]}{S'_k[\tau]} - \left(\sum_\tau S_k[\tau]H'_k[n-\tau]\right)^2 \\
=& \sum_{\tau,\nu} \frac{H'_k[n-\tau]S_k[\tau]^2 H'_k[n-\nu]S'_k[\nu]}{S'_k[\tau]} - \sum_{\tau,\nu} S_k[\tau]H'_k[n-\tau]S_k[\nu]H'_k[n-\nu] \\
=& \sum_{\tau,\nu} \left(\frac{H'_k[n-\tau]S_k[\tau]^2 H'_k[n-\nu]S'_k[\nu]}{S'_k[\tau]} - S_k[\tau]H'_k[n-\tau]S_k[\nu]H'_k[n-\nu] \right) \\
=& \sum_{\tau \neq \nu} \left(\frac{H'_k[n-\tau]S_k[\tau]^2 H'_k[n-\nu]S'_k[\nu]}{S'_k[\tau]} - S_k[\tau]H'_k[n-\tau]S_k[\nu]H'_k[n-\nu] \right) \\
=& \sum_{\tau < \nu} H'_k[n-\tau]H'_k[n-\nu] \left(\frac{S_k[\tau]^2 S'_k[\nu]}{S'_k[\tau]} - 2S_k[\tau]S_k[\nu] + \frac{S_k[\nu]^2 S'_k[\tau]}{S'_k[\nu]} \right) \\
=& \sum_{\tau < \nu} \frac{H'_k[n-\tau]H'_k[n-\nu]}{S'_k[\nu]S'_k[\tau]} \left(S_k[\tau]S'_k[\nu] - S_k[\nu]S'_k[\tau]\right)^2 \geq 0.
\end{aligned}$$

To prove that $Q(S'_k[n]) \geq S_k[n]^p$, we begin by noting that $Q \in \mathcal{C}^\infty(\mathbb{R}^+)$. Then, the first order necessary condition for Q yields

$$0 = \frac{\partial Q}{\partial x} = \frac{p(p-2)}{2} x^{p-3} S_k[n]^2 + p x^{p-1} - \frac{p^2}{2} x^{p-1} = \frac{p(p-2)}{2} x^{p-1}(x^{-2} S_k[n]^2 - 1),$$

meaning the only point at which the derivative of Q equals zero is at $x = S_k[n]$. Furthermore, $\frac{\partial^2}{\partial x^2} Q(S_k[n]) = S_k[n]^{p-2}(2p - p^2) > 0 \; \forall p \in (0,2)$, meaning that $Q(S_k[n]) = S_k[n]^p$ is the global minimum of Q. This yields

$$\begin{aligned} g_s(S, S') &= \sum_k \left(\sum_n \left(\frac{1}{\sigma_k^2} P_{k,n} + \frac{1}{b_k^p} Q(S'_k[n]) \right) + \frac{1}{\eta_k^2} ||LH'_k||_2^2 \right) \\ &\geq \sum_k \left(\sum_n \left(\frac{1}{\sigma_k^2} R_{k,n} + \frac{1}{b_k^p} S_k[n]^p \right) + \frac{1}{\eta_k^2} ||LH'_k||_2^2 \right) = J(S, H'). \end{aligned}$$

■

B Derivation of updating rule for H

In order to derive the updating rule for H, we shall write g_h as a function of the transposed rows H_k. We begin by noting

$$\begin{aligned} g_h(H, H') &= \sum_{k,n,\tau} \frac{1}{\sigma_k^2} \frac{S'_k[n-\tau] H'_k[\tau]}{X'_k[n]} \left(Y_k[n] - \frac{H_k[\tau]}{H'_k[\tau]} X'_k[n] \right)^2 \\ &\quad + \sum_k \frac{1}{b_k^p} ||S'_k||_p^p + \sum_k \frac{1}{\eta_k^2} ||LH_k||_2^2 \\ &= \sum_{k,n,\tau} \frac{1}{\sigma_k^2} \frac{S'_k[n-\tau] H'_k[\tau] Y_k^2[n]}{X'_k[n]} - 2 \sum_{k,n,\tau} \frac{1}{\sigma_k^2} S'_k[n-\tau] Y_k[n] H_k[\tau] \\ &\quad + \sum_{k,n,\tau} \frac{1}{\sigma_k^2} \frac{S'_k[n-\tau] X'_k[n] H_k^2[\tau]}{H'_k[\tau]} \\ &\quad + \sum_k \frac{1}{b_k^p} ||S'_k||_p^p + \sum_k \frac{1}{\eta_k^2} ||LH_k||_2^2. \end{aligned}$$

Next, we recall the definition of the diagonal matrices $A^k, B^k \in \mathbb{R}^{N_h \times N_h}$, whose diagonal elements are $A^k_{\tau,\tau} \doteq \sum_n S'_k[n-\tau] X'_k[n]$ and $B^k_{\tau,\tau} \doteq H'_k[\tau]$, and the vector $\zeta^k \in \mathbb{R}^{N_h}$ with components $\zeta^k_\tau = \sum_n S'_k[n-\tau] Y_k[n]$. With these definitions, we can write

$$\begin{aligned} g_h(H, H') &= \sum_{k,n,\tau} \frac{1}{\sigma_k^2} \frac{S'_k[n-\tau] H'_k[\tau] Y_k^2[t]}{X'_k[n]} - 2 \sum_k \frac{1}{\sigma_k^2} H_k^{\mathrm{T}} \zeta^k \\ &\quad + \sum_k \frac{1}{\sigma_k^2} H_k^{\mathrm{T}} A^k (B^k)^{-1} H_k + \sum_k \frac{1}{b_k^p} ||S'_k||_p^p + \sum_k \frac{1}{\eta_k^2} H_k^{\mathrm{T}} L^{\mathrm{T}} L H_k. \end{aligned}$$

Now, the first order necessary condition for g_h with respect to H_k is given by

$$0 = \frac{\partial g_h(H, H')}{\partial H_k} = -\frac{2}{\sigma_k^2} \zeta^k + \frac{2}{\sigma_k^2} A^k (B^k)^{-1} H_k + \frac{2}{\eta_k^2} L^{\mathrm{T}} L H_k,$$

which readily leads to the linear system

$$\left(A^k + \frac{\sigma_k^2}{\eta_k^2} B^k L^{\mathrm{T}} L \right) H_k = B^k \zeta^k.$$

C Updating rule for η

In order to derive the updating rule for $\eta_k, k = 1, \ldots, K$, we begin by noting that $-\log \pi_{post}(S, H, \eta|Y) \in \mathcal{C}^1(0, \infty)$ with respect to η_k, and hence a local minimum must corresponds to a point with derivative equal to zero. Differentiating (7) with respect to η_k, we obtain

$$\frac{\partial}{\partial \eta_k} - \log \pi_{post}(S, H, \eta|Y) = -\frac{2}{\eta_k^3}||LH_k||_2^2 + \frac{N_h + 1 - \alpha}{\eta_k} + \frac{1}{\beta}.$$

The first order necessary condition over (7) is thus tantamount to

$$\eta_k^3 + (N_h + 1 - \alpha)\beta\, \eta_k^2 - 2\beta||LH_k||_2^2 = 0.$$

By Descartes' rule, this polynomial has exactly one positive root η_0. Since $\lim_{\eta_k \to \infty} (-\log \pi_{post}(S, H, \eta|Y)) = \infty$ and $\lim_{\eta_k \to 0^+} (-\log \pi_{post}(S, H, \eta|Y)) = \infty$, then η_0 is the global minimizer.

Bibliography

References

[1] M. Kim and H.-M. Park, "Efficient online target speech extraction using DOA-constrained independent component analysis of stereo data for robust speech recognition," *Signal Processing*, vol. 117, pp. 126–137, 2015.

[2] S. Yun, Y. J. Lee, and S. H. Kim, "Multilingual speech-to-speech translation system for mobile consumer devices," *IEEE Transactions on Consumer Electronics*, vol. 60, no. 3, pp. 508–516, 2014.

[3] R. Neßelrath, M. M. Moniri, and M. Feld, "Combining speech, gaze, and micro-gestures for the multimodal control of in-car functions," in *12th IEEE International Conference on Intelligent Environments (IE)*, 2016, pp. 190–193.

[4] L. Di Persia, D. Milone, H. L. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Processing*, vol. 88, no. 10, pp. 2578–2583, 2008.

[5] C. E. Martínez, J. Goddard, L. E. Di Persia, D. H. Milone, and H. L. Rufiner, "Denoising sound signals in a bioinspired non-negative spectro-temporal domain," *Digital Signal Processing*, vol. 38, pp. 22–31, 2015.

[6] L. Di Persia, D. Milone, and M. Yanagida, "Indeterminacy free frequency-domain blind separation of reverberant audio sources." *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 2, pp. 299–311, 2009.

[7] L. E. Di Persia and D. H. Milone, "Using multiple frequency bins for stabilization of FD-ICA algorithms," *Signal Processing*, vol. 119, pp. 162–168, 2016.

[8] A. Tsilfidis and J. Mourjopoulos, "Signal-dependent constraints for perceptually motivated suppression of late reverberation," *Signal Processing*, vol. 90, no. 3, pp. 959–965, 2010.

[9] I. J. Tashev, *Sound capture and processing: practical approaches.* John Wiley & Sons, 2009.

[10] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development.* Prentice hall PTR Upper Saddle River, 2001, vol. 95.

[11] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani *et al.*, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge," in *Proceedings of REVERB Challenge Workshop*, 2014.

[12] S. Wisdom, T. Powers, L. Atlas, and J. Pitton, "Enhancement of reverberant and noisy speech by extending its coherence," in *Proceedings of REVERB Challenge Workshop*, 2014.

[13] M. Moshirynia, F. Razzazi, and A. Haghbin, "A speech dereverberation method using adaptive sparse dictionary learning," in *Proceedings of REVERB Challenge Workshop*, 2014.

[14] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E.-S. Chng, and H. Li, "The NTU-ADSC systems for reverberation challenge 2014," in *Proceedings of REVERB Challenge Workshop*, 2014.

[15] K. Nathwani and R. M. Hegde, "Joint source separation and dereverberation using constrained spectral divergence optimization," *Signal Processing*, vol. 106, pp. 266–281, 2015.

[16] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 45–48.

[17] F. Ibarrola, L. Di Persia, and R. Spies, "Blind speech dereverberation using convolutive nonnegative matrix factorization with mixed penalization." in *Proceedings of VI Congreso de Matemática Aplicada, Computacional e Industrial*, 2017, pp. 404–407.

[18] F. Ibarrola, G. Mazzieri, R. Spies, and K. Temperini, "Anisotropic BV-L^2 regularization of linear inverse ill-posed problems," *Journal of Mathematical Analysis and Applications*, vol. 450, no. 1, pp. 427–443, 2017.

[19] D. Lazzaro, L. B. Montefusco, and S. Papi, "Blind cluster structured sparse signal recovery: A nonconvex approach," *Signal Processing*, vol. 109, pp. 212–225, 2015.

[20] F. Ibarrola and R. Spies, "A two-step mixed inpainting method with curvature-based anisotropy and spatial adaptivity," *Inverse Problems & Imaging*, vol. 11, no. 2, pp. 247–262, 2017.

[21] V. Peterson, H. L. Rufiner, and R. D. Spies, "Generalized sparse discriminant analysis for event-related potential classification," *Biomedical Signal Processing and Control*, vol. 35, pp. 70–78, 2017.

[22] G. Mazzieri, R. Spies, and K. Temperini, "Mixed spatially varying L^2-BV regularization of inverse ill-posed problems," *Journal of Inverse and Ill-posed Problems*, vol. 23, no. 6, pp. 571–585, 2015.

[23] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[24] B. Yegnanarayana, P. S. Murthy, C. Avendaño, and H. Hermansky, "Enhancement of reverberant speech using LP residual," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1. IEEE, 1998, pp. 405–408.

[25] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," *Proceedings of the 5th Conference on Independent Component Analysis and Blind Signal Separation*, pp. 494–499, 2004.

[26] C. Bouman and K. Sauer, "A generalized gaussian image model for edge-preserving map estimation," *IEEE Transactions on Image Processing*, vol. 2, no. 3, pp. 296–310, 1993.

[27] E. De Sena, N. Antonello, M. Moonen, and T. Van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 4, pp. 774–786, 2015.

[28] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien Jr, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[29] K. Matsuoka, "Minimal distortion principle for blind source separation," in *41st SICE Annual Conference*, vol. 4. IEEE, 2002, pp. 2138–2143.

[30] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[31] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2001, pp. 556–562.

[32] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[33] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[34] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[35] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[36] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[37] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Audio Engineering Society Convention 122*. Audio Engineering Society, 2007.

Switching divergences for spectral learning in blind speech dereverberation

Switching divergences for spectral learning in blind speech dereverberation

Francisco J. Ibarrola [*] Leandro E. Di Persia [*] Ruben D. Spies [†]

Abstract

When recorded in an enclosed room, a sound signal will most certainly get affected by reverberation. This not only undermines audio quality, but also poses a problem for many human-machine interaction technologies that use speech as their input. In this work, a new blind, two-stage dereverberation approach based in a generalized β-divergence as a fidelity term over a non-negative representation is proposed. The first stage consists of learning the spectral structure of the signal solely from the observed spectrogram, while the second stage is devoted to model reverberation. Both steps are taken by minimizing a cost function in which the aim is put either in constructing a dictionary or a good representation by changing the divergence involved. In addition, an approach for finding an optimal fidelity parameter for dictionary learning is proposed. An algorithm for implementing the proposed method is described and tested against state-of-the-art methods. Results show improvements for both artificial reverberation and real recordings.

Keywords: signal processing, dereverberation, penalization

1 Introduction

Over the last years, with the technological advances and massive adoption of portable electronic devices with high computational capacity, much effort has been devoted to improving audio signal quality. Given that speech signals are recorded in a wide variety of environments, the audio quality usually results degraded by noise, reverberation or the presence of other sources, often severely diminishing intelligibility. This problem does not arise solely when people communicate with each other, but also affects human-machine interaction capabilities of electronic devices. The need for improving audio

[*] Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), UNL, CONICET, FICH, Ciudad Universitaria, CC 217, Ruta Nac. 168, km 472.4, (3000) Santa Fe, Argentina. (fibarrola@sinc.unl.edu.ar).

[†] Instituto de Matemática Aplicada del Litoral, IMAL, UNL, CONICET, Centro Científico Tecnológico CONICET Santa Fe, Colectora Ruta Nac. 168, km 472, Paraje "El Pozo", (3000), Santa Fe, Argentina and Departamento de Matemática, Facultad de Ingeniería Química, Universidad Nacional del Litoral, Santa Fe, Argentina.

quality is therefore inherent to a number of hot topics in the field of signal processing, including automatic translation systems ([1]), emotion and affective state recognition ([2]), digital personal assistants ([3]), to name just a few that require the use of speech as inputs.

One of the main difficulties within this context comes from the fact that when recorded in enclosed rooms, audio signals are affected by reverberant components due to reflections of the sound waves in the walls, floor and ceiling. This can severely degrade the quality of the recorded signals (particularly when the microphones are far away from the sources, [4]), which in turn makes them unsuitable for direct use in certain speech applications ([5]). The goal of this work is to produce a dereverberation technique for removing or highly attenuating the reverberant components of a recorded signal in order to enhance its quality.

A speech dereverberation problem can be classified as "blind" whenever the available data consist only of the reverberant signal itself, or as "supervised" when information of the environment or the speakers is available. The problem can also be classified as single or multi-channel, depending on the number of microphones used for recording. In this work, we shall address the problem within a blind, single-channel setting, which is the most common in real-life problems, but also the most difficult, because of the scarce information.

Due to the characteristics of speech signals, most state-of-the-art methods deal with the dereverberation problem in a transformed domain, such as the one obtained by the Fan-Chirp Transform (see [6]) or the Short-Time Fourier Transform (STFT) ([7]). Some of these methods make use of non-negative matrix factorization (NMF) or its variants, such as convolutive NMF ([8]), along with Bayesian or penalization approaches.

Although some early NMF-based approaches, such as those introduced in [9] and [10], have shown to produce satisfactory results, they often neglect the relation between frequency components. In fact, an underlying hypothesis is that the frequency bins on the spectrogram of a clean signal are uncorrelated, which turns out to be an oversimplification when dealing with speech signals, given their harmonic structure. This spectral structure can be incorporated on a convolutive NMF model by using a *dictionary*-based approach, as proposed in [11]. This approach has shown good performance within a supervised framework, but not quite so within a blind setting. This has to do with the fact that there are too many variables that have to be learned simultaneously from the scarce available data. This often results in a good representation of the reverberant spectrogram with elements that do not allow for a good representation of the associated clean spectrogram.

In order to overcome the aforementioned issues, in this work we propose to approach the problem in a two stage setting. The first stage is meant to build a dictionary from the data that allows for a good representation of the clean signal, while the second one is devoted to use such a dictionary for getting an appropriate representation of the reverberant spectrogram. This is accomplished by introducing a general form of a cost function with mixed penalization over a dictionary-based convolutive reverberation model, that can then be tuned up to fit either the task on the first or the second stage of the method. The main novelty of this work is that the process of learning the spectral structure (*i.e.* the first stage) is not aimed to obtain an optimal representation of the reverberant signal.

The paper is organized as follows: in Section 2 we state the reverberation model combining a convolutive and a dictionary-based NMF representations. In Section 3 we provide a brief overview of the dereverberation algorithm and introduce the cost function to be minimized. The updating rules for iteratively approaching minimizers of such a cost function are derived in Section 4, along with the full algorithm. Parameter estimation processes and validation experiments are detailed in Section 5, as well as the obtained results. Conclusions are discussed in Section 6.

2 Reverberation Model

Let $s, x, h : \mathbb{R} \to \mathbb{R}$, supported in $[0, \infty)$, denote the functions associated to the clean and reverberant signals, and the room impulse response (RIR), respectively. As it is customary, we make the assumption that reverberation is well represented by a Linear Time-Invariant (LTI) system, which can be written as

$$x(t) = (h * s)(t), \tag{1}$$

where "$*$" denotes convolution. The use of this representation is underlaid by the hypotheses that the source and microphone positions are fixed, and the non-linear components are small enough to be neglected.

As we previously mentioned, when dealing with speech signals, it is often convenient to work with time-frequency representations rather than in the time domain. Thus, we shall make use of the Short Time Fourier Transform (STFT).

2.1 STFT-based reverberation model

The STFT of a function x can be defined as

$$\mathbf{x}_k(t) \doteq \int_{-\infty}^{\infty} x(u) w(u-t) e^{-2\pi i u k} du, \quad t, k \in \mathbb{R},$$

where $w : \mathbb{R} \to \mathbb{R}_0^+$ is a prescribed even and compactly supported function such that $\|w\|_1 = 1$, called *window*.

Naturally, in practice we work with discretized versions of the signals, denoted as $x[\cdot]$, $h[\cdot]$, $s[\cdot]$, and $w[\cdot]$. The corresponding discrete STFT can be defined as

$$\mathbf{x}_k[n] \doteq \sum_{m=-\infty}^{\infty} x[m] w[m-n] e^{-2\pi i m k},$$

where $n = 1, \ldots, N$, is a discrete time variable associated to the window locations, and $k = 1, \ldots, K$, denotes the frequency sub-band. Similarly, we denote by $\mathbf{s}_k[n]$ and

$\mathbf{h}_k[n]$ the STFTs of s and h, respectively. A discrete approximation of (1) in the STFT domain is given by

$$\mathbf{x}_k[n] \approx \tilde{\mathbf{x}}_k[n] \doteq \sum_{m=0}^{M-1} \mathbf{s}_k[n-m]\mathbf{h}_k[m], \quad n, k \in \mathbb{N}. \tag{2}$$

where M is a given model parameter determined by the reverberation time. The model is built as in [12], where the approximation in (2) holds due to the use of band-to-band only filters. The window locations are chosen so that the support of the observed signal is contained in the union of the supports of the windows, and K as to reach up to half the sampling frequency.

Since phase angles on the STFT components have been shown to be highly sensitive to mild variations on the associated signal ([13]), and within our blind setting we have no information about reverberation conditions, we proceed as in [9], by treating the phase angles $\phi_k[m]$ of $\mathbf{h}_k[m]$ as random variables. Let us assume them to be *i.i.d.* with uniform distribution in $[-\pi, \pi)$. Under this hypothesis, it can be shown ([7]) that the expected value of $|\tilde{\mathbf{x}}_k[t]|^2$ is given by

$$E|\tilde{\mathbf{x}}_k[n]|^2 = \sum_m |\mathbf{s}_k[n-m]|^2 \, |\mathbf{h}_k[m]|^2.$$

Note that the choice of $[-\pi, \pi)$ is arbitrary, since the equality holds for any $2\pi-$length interval. Finally, by defining $S_{k,n} \doteq |\mathbf{s}_k[n]|^2$, $H_{k,n} \doteq |\mathbf{h}_k[n]|^2$ and $X_{k,n} \doteq E|\tilde{\mathbf{x}}_k[n]|^2$, the convolutive NMF model reads

$$X_{k,n} = \sum_{m=0}^{M'} S_{k,n-m} H_{k,m}, \tag{3}$$

for $k = 1, \ldots, K,\ n = 1, \ldots, N$. Here, $M' \doteq \min\{M-1, n-1\}$, so we can treat X, S and H as nonnegative matrices with elements $X_{k,n}$, $S_{k,n}$ and $H_{k,n}$, respectively.

Since we intend to introduce a spectral modeling of the clean signal, we shall make use of an NMF approach over the clean spectrogram S.

2.2 NMF model

Let us assume that there exist $W \in \mathbb{R}_{0,+}^{K \times J},\ U \in \mathbb{R}_{0,+}^{J \times N}$, $(J < \min\{K, N\})$ that provide a "good" NMF representation for $S \in \mathbb{R}_{0,+}^{K \times N}$. That is,

$$S \cong WU.$$

The accuracy of this approximation can be defined in terms of the Euclidean distance or some divergence measure (details on this will be discussed later on). In order to keep the notation simple, we shall assume the latter approximation to hold exactly and replace S in (3) by WU, which results in the model

$$X_{k,n} = \sum_{m=0}^{M'} \sum_{j=1}^{J} W_{k,j} U_{j,n-m} H_{k,m}. \tag{4}$$

Two remarks are in order: firstly, note that the approximation error in the assumption $S = WU$ will be taken into account by the representation error of X with respect to the data, and hence the latter assumption poses no problem. Secondly, we note that the model (4) has a scale indeterminacy, in the sense that for any $\alpha > 0$, the matrices $\tilde{W} = \alpha W$, $\tilde{H} = \alpha H$, and $\tilde{U} = \alpha^{-2} U$ would give the same representation X. Hence, in order to avoid numerical issues, we add the constraints $\|W_j\|_1 = \|H_k^T\|_\infty = 1$, where $W_j,\ j = 1, \ldots, J$, are the columns of W and $H_k,\ k = 1, \ldots, K$ are the rows of H. This means that the spectrogram S is represented by a normalized dictionary and that reverberation preserves the signal's maximal energy.

3 Algorithm and cost functions

In this section, a fidelity term and penalizers for building an appropriate cost function f will be defined. This cost function will then be minimized in order to obtain the desired matrices $\hat{W}$, $\hat{U}$ and $\hat{H}$, as follows:

Algorithm overview

1. Set the parameters of $f = f(Y, X)$ so as to prioritize spectral learning and minimize f with respect to its arguments in order to find an appropriate dictionary $\hat{W}$.

2. Reset the parameters of f in order to emphasize accuracy in the representation. Then minimize f with respect to U and H subject to $W = \hat{W}$, to obtain $\hat{U}$ and $\hat{H}$.

3. Approximate the clean spectrogram S using $\hat{W}$ and $\hat{U}$.

3.1 Fidelity term

Given a reverberant (and possibly noisy) spectrogram Y, we intend to find matrices W, U and H that, while complying with certain desired characteristics, provide a representation X, as in (4), that accurately approximates Y.

Many ways of measuring the fidelity of that approximation have been proposed: the Euclidean distance ([9]), the Kullback-Leibler divergence ([11]), and the Itakura-Saito divergence ([14]) being the most commonly used. Assume we have a known clean spectrogram S that we want to represent using an NMF factorization WU. Different choices of the fidelity measure will lead to dictionary atoms (column vectors of W) with different characteristics. This is illustrated in Fig. 1, where it can be observed that a particular fidelity measure may emphasize the appearance of atoms with high values in the low frequencies (thus enabling a good approximation in the higher energy zones while neglecting the low-energy ones), while another fidelity measure may result in a more evenly represented frequency range. Even though it might appear that the latter would be better, the relation between the observed differences in the dictionaries and

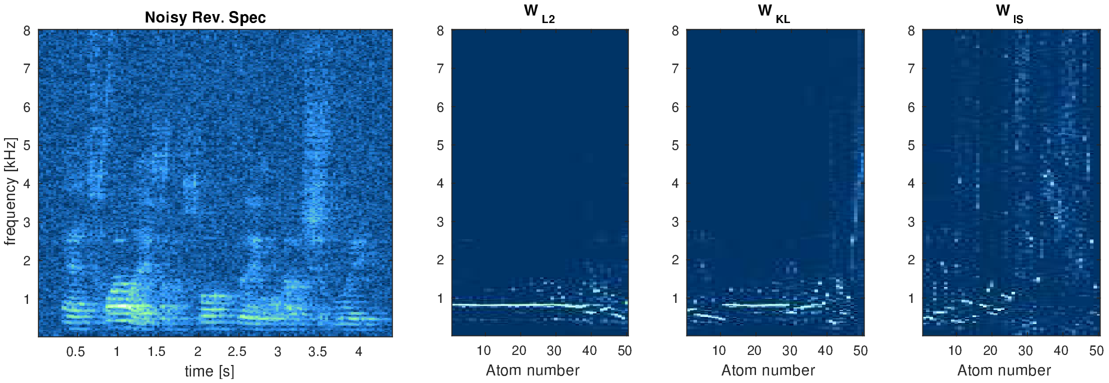


Figure 1: Left: The spectrogram of a clean signal, sampled at 16[kHz], using a 512 samples window with overlapping of 256. W_{L2}: dictionary obtained using Frobenius norm. W_{KL}: dictionary obtained using Kullback-Leibler divergence. W_{IS}: dictionary obtained using Itakura-Saito divergence. All the dictionary atoms were ordered by correlation in order to help visualization.

their capability for representing a high-quality audio signal is unclear at this point. In fact, later on it will be shown that the optimal choice (in some sense) of the divergence measure for building a dictionary does not correspond to any of the ones depicted here.

In order to find an "optimal" dictionary W, we begin by recalling a generalized divergence, as introduced in [15]. For $X, Y \in \mathbb{R}_{0,+}^{K\times N}$ and $\beta \in \mathbb{R}_+ \backslash \{1\}$, the β-divergence of X from Y is defined as

$$D_\beta(Y||X) \doteq \sum_{k,n} \left(Y_{k,n} \frac{Y_{k,n}^{\beta-1} - X_{k,n}^{\beta-1}}{\beta(\beta-1)} + X_{k,n}^{\beta-1} \frac{X_{k,n} - Y_{k,n}}{\beta} \right).$$

This β-divergence generalizes all three aforementioned fidelity measures. In fact, it can be seen that $D_2(\cdot||\cdot)$ corresponds to (half) the squared Frobenius norm of $Y - X$, whereas $D_\beta(\cdot||\cdot)$ approaches the Kullback-Leibler divergence as $\beta \to 1$ and the Itakura-Saito divergence as $\beta \to 0$. An appropriate way of choosing the parameter β will be discussed later on. We now proceed to introduce the penalization terms which shall embed the desired characteristics on the components that constitute the model.

3.2 Penalizers

Clearly, there are many ways of building the matrices W, U and H leading to a representation with small divergence with respect to the observation. One way of narrowing down the possible choices is by introducing penalizing terms into our cost function for promoting certain desired features over its minimizers. In a quite general context, this leads to a cost function of the form

$$f(W, U, H) \doteq D_\beta(Y||X) + P_u(U) + P_h(H),$$

where $P_u : \mathbb{R}_{0,+}^{J\times N} \rightarrow \mathbb{R}_{0,+}$, and $P_h : \mathbb{R}_{0,+}^{K\times M} \rightarrow \mathbb{R}_{0,+}$ are penalizing functions, each one imposing a cost over the appearance of certain features on U and H, respectively.

As it can be observed, while the spectrogram of the clean signal depicted in Fig. 2 presents a somewhat sparse structure, the one corresponding to the reverberant signal presents a smoother, more diffuse structure. It is well known that sparsity over the coefficient matrix can be induced by using the ℓ^1 norm (see [11]). In our case, we shall hinder the smoothness observed in the reverberant spectrogram from appearing in the restored spectrogram by using a penalizer over the activation coefficients matrix U of the form

$$P_u(U) \doteq \sum_{j,n} \lambda_n^{(u)} U_{j,n},$$

where $\lambda_n^{(u)} \geq 0$, $n = 1, \ldots, N$, are penalization parameters. Note that since the elements of U are non-negative, this penalizer corresponds to a weighted ℓ^1 norm of U. In order to allow for better compliance with the inherent silences of the recorded signals, we let the penalizer depend on the time index n (more on this subject in Section 5.2.3).

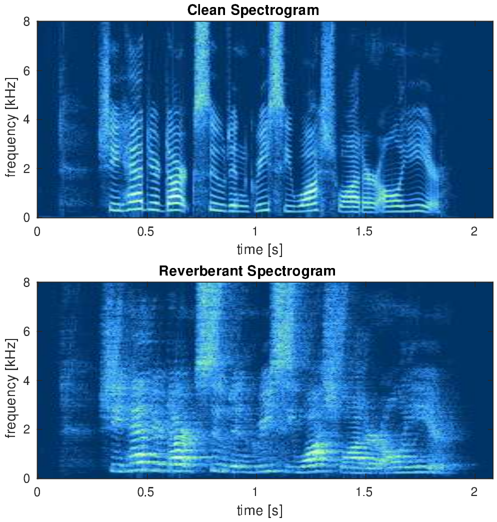


Figure 2: Top: spectrogram of a clean signal, sampled at 16[kHz], using a 512 samples window with overlapping of 256. Bottom: the spectrogram of a reverberant (600[ms]) version of the same signal.

In order to define a penalizer over H, we turn our attention to Fig. 3, that shows a recorded RIR in a room with a reverberation time of 600[ms]. The log-spectrogram exhibits a high-energy vertical band on the left, corresponding to the first echoes to reach the receiver, that slowly fades to the right, as deemed by a linear impulse response. From this characteristic, and the fact that the overlapping of windows results in consecutive time components of H capturing common information, it is reasonable to expect the components of H to exhibit a smooth decay over time ([16]). This structure can be

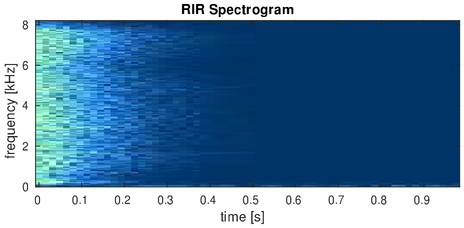


Figure 3: Log-spectrogram for a recorded 16 [kHz] RIR signal with reverberation time of 600 [ms]. The spectrogram was made using a Hanning window length of 512 samples and overlapping of 256, from a recording made in a fully furnished office room

promoted (see [7]) by introducing a penalizer of the form

$$P_h(H) \doteq \sum_k \lambda_k^{(h)} \|LH_k^T\|_2^2,$$

where $\lambda_k^{(h)} \geq 0,\ H_k \in \mathbb{R}_{0,+}^M,\ k = 1, \ldots, K$ are the rows of H, and $L \in \mathbb{R}^{(M-1)\times M}$ is a *finite difference matrix*, so that $[LH_k^T]_m = H_{k,m+1} - H_{k,m}$.

With all of the above, the cost function is defined as follows:

$$f(W, U, H) \doteq D_\beta(Y||X) + \sum_{j,n} \lambda_n^{(u)} U_{j,n} + \sum_k \lambda_k^{(h)} \|LH_k^T\|_2^2. \tag{5}$$

In the next section we state a two-stage optimization process in order to minimize f, first with respect to W, and then with respect to both U and H. In-line with the core idea stated before, by appropriately tunning its parameters, the cost function (5) can be used for building a good dictionary in a first stage, and for seeking a good representation of the data in a second step.

4 Optimization

The optimization process that shall yield the restored spectrogram $\hat{S}$ is divided in two main steps: firstly, given the observed reverberant spectrogram $Y \in \mathbb{R}_{0,+}^{K\times N}$, a suitable dictionary $\hat{W} \in \mathbb{R}_{0,+}^{K\times J}$ that be able to provide a good representation of the target clean spectrogram S is built. Once this is accomplished, the algorithm proceeds to find $\hat{U} \in \mathbb{R}_{0,+}^{J\times N}$ and $\hat{H} \in \mathbb{R}_{0,+}^{K\times M}$ minimizing f given $\hat{W}$.

In order to minimize the cost function, we shall begin by recalling the concept of auxiliary function.

4.1 Auxiliary function

Definition 4.1. *Let $\Omega \subset \mathbb{R}^P$ and $f : \Omega \to \mathbb{R}_0^+$. Then, $g : \Omega \times \Omega \to \mathbb{R}_0^+$ is called an* auxiliary function *for f if $g(\omega, \omega) = f(\omega)$ and $g(\omega, \omega') \geq f(\omega),\ \ \forall \omega, \omega' \in \Omega$.*

Lemma 4.2. *If we let f and g be as in the definition above, $\omega^0 \in \Omega$ be arbitrary and*

$$\omega^t \doteq \arg\min_{\omega} g(\omega, \omega^{t-1}),\ t \in \mathbb{N}$$

then it can be shown ([17]) that the sequence $\{f(\omega^t)\}_{t\geq 1}$ is non-increasing.

The idea is to build an auxiliary function g for f with respect to each of its three arguments individually, and then use them iteratively for minimizing f.

We will proceed in a similar fashion than in [18]. Firstly, let us notice that $\forall Y \in \mathbb{R}_{0,+}^{K\times N}$, $D_\beta(Y||\cdot) \in \mathcal{C}^\infty(\mathbb{R}_+^{K\times N})$, and

$$\frac{\partial^2 D_\beta(Y||X)}{\partial X_{k,n}^2} = (\beta-1)X_{k,n}^{\beta-2} + (2-\beta)X_{k,n}^{\beta-3}Y_{k,n}. \tag{6}$$

By defining

$$\check{D}_\beta(Y||X) \doteq \sum_{k,n}\left(\frac{\chi_{\beta>1}(\beta)}{\beta}X_{k,n}^\beta - \frac{\chi_{\beta\leq 2}(\beta)}{\beta-1}Y_{k,n}X_{k,n}^{\beta-1} + \frac{1}{\beta(\beta-1)}Y_{k,n}^\beta\right),$$

and

$$\hat{D}_\beta(Y||X) \doteq \sum_{k,n}\left(\frac{\chi_{\beta<1}(\beta)}{\beta}X_{k,n}^\beta - \frac{\chi_{\beta>2}(\beta)}{\beta-1}Y_{k,n}X_{k,n}^{\beta-1}\right),$$

we have $D_\beta = \check{D}_\beta + \hat{D}_\beta$, where $\check{D}_\beta$ is convex and $\hat{D}_\beta$ is concave (both w.r.t. X). In the following, we will make use of this decomposition in order to build auxiliary functions for updating each one of the components of X.

4.2 Building $\hat{W}$

As mentioned before, the parameters required for building a proper dictionary $\hat{W}$ are not necessarily the same as those leading to an optimal representation. Thus, we begin by fixing $H_{k,n} = 1$ if $n = 1$ and $H_{k,n} = 0, \forall n = 2, \ldots, M,\ \ k = 1\ldots, K$. This means that we are precluding H from modeling reverberation, and henceforth it does not make sense to promote temporal sparsity over U, and so we set $\lambda_n^{(u)} = 0,\ \forall n = 1, \ldots, N$, only for the first stage.

Now, provided we have found adequate parameters (what we address in Section 5.2.3), the problem of finding an appropriate dictionary reduces to minimizing (5) with respect to W and U subject to H and $\lambda_n^{(u)}$ be set as above. To do this, we begin by stating an auxiliary function for f w.r.t. W as follows:

$$g_w(W, W') \doteq \sum_{k,n,j,m} \frac{W'_{k,j}U_{j,n-m}H_{k,m}}{X'_{k,n}}\check{D}_\beta\left(Y_{k,n}\middle|\middle| X'_{k,n}\frac{W_{k,j}}{W'_{k,j}}\right) + \sum_{k,n}\hat{D}_\beta(Y_{k,n}||X'_{k,n})$$
$$+ \sum_{k,n,j,m}\frac{\partial \hat{D}_\beta(Y_{k,n}||X'_{k,n})}{\partial X_{k,n}}(W_{k,j} - W'_{k,j})U_{j,n-m}H_{k,m}.$$

A proof that g_w complies with the conditions in Definition 4.1 can be found in Appendix A. Since $g_w(W, W')$ is convex with respect to W, it can be minimized by equating its gradient to zero, what leads to

$$0 = \left(\frac{W_{k,j}}{W'_{k,j}}\right)^{\alpha_1} \sum_{n,m} X'^{\beta-1}_{k,n} U_{j,m} H_{k,n-m} - \left(\frac{W_{k,j}}{W'_{k,j}}\right)^{\alpha_2} \sum_{n,m} X'^{\beta-2}_{k,n} Y_{k,n} U_{j,m} H_{k,n-m},$$

where $\alpha_1 = (\beta - 1)\chi_{\beta>1}(\beta)$, and $\alpha_2 = (\beta - 2)\chi_{\beta\leq 2}(\beta)$. This automatically leads to the updating equation

$$W^{(t)}_{k,j} = W^{(t-1)}_{k,j} \frac{\left[\left(\sum\limits_{m,n} \left(X^{(t-1)}_{k,n}\right)^{\beta-2} Y_{k,n} U_{j,m} H_{k,n-m}\right)^{\eta}\right]_{\epsilon}}{\left(\sum\limits_{m,n} \left(X^{(t-1)}_{k,n}\right)^{\beta-1} U_{j,m} H_{k,n-m}\right)^{\eta}}, \tag{7}$$

where $\eta \doteq \frac{1}{\alpha_1 - \alpha_2}$. Here, the supra index t denotes the iteration number and $[\cdot]_\epsilon$ denotes the operation $\max\{\cdot\,, \epsilon\}$, with ϵ being a small constant ($\sim 10^{-10}$). This is used to avoid the elements of W from dropping to 0 (or below), as once an element is null, it cannot regain positive values by a multiplicative updating procedure (see [19]). For simplicity of notation, we have avoided the use of superscripts in all the variables that do not depend directly on W.

An analogous procedure (see Appendix B) leads to the following updating rule for U:

$$U^{(t)}_{j,m} = U^{(t-1)}_{j,m} \frac{\left[\left(\sum\limits_{k,n} \left(X^{(t-1)}_{k,n}\right)^{\beta-2} Y_{k,n} W_{k,j} H_{k,n-j} - \lambda^{(u)}_m\right)^{\eta}\right]_{\epsilon}}{\left(\sum\limits_{k,n} \left(X^{(t-1)}_{k,n}\right)^{\beta-1} W_{k,j} H_{k,n-j}\right)^{\eta}}. \tag{8}$$

The dictionary $\hat{W} = \arg\min_W f(W, U, H)$ can thus be obtained by alternatively updating W and U using (7) and (8), respectively, until convergence.

Once $\hat{W}$ is obtained, we proceed to find $\hat{U}$ and $\hat{H}$ that be able to effectively model reverberation.

4.3 Building $\hat{U}$ and $\hat{H}$

Unlike in the first step, now we do want to impose a sparse structure over U, and so $\lambda^{(u)}_n$ should no longer be null for every $n = 1, \ldots, N$. Furthermore, it should be pointed out that the value of β in this stage is not necessarily the same as in the previous one (and in fact they will be chosen differently in practice).

The updating rule for U is exactly the same as stated in (8). For updating H, we begin by defining, for every $k = 1, \ldots, K$, the diagonal matrix $A^{(k)} \in \mathbb{R}^{M\times M}_{0,+}$ with $A^{(k)}_{m,m} = \sum_{j,n} W_{k,j} U_{j,n-m} \left(X^{(t-1)}_{k,n}\right)^{\alpha_1} / H^{(t-1)}_{k,m}$ and the vector $b^{(k)} \in \mathbb{R}^M_{0,+}$ as $b^{(k)} =$

$\sum_{j,n} W_{k,j}U_{j,n-m}Y_{k,n}\left(X_{k,n}^{(t-1)}\right)^{\alpha_2}$. Then, under the same approximation used for arriving at (8), we can update H by solving for $H_k^{(t)}$, $k = 1, \ldots, K$, the linear system

$$\left(A^{(k)} + 2\lambda_k^{(h)} L^T L\right) H_k^{(t)} = b^{(k)}. \tag{9}$$

The justification for this can be found in Appendix C.

4.4 Additional considerations

Our approximate solution could be defined simply as $\hat{S} = \hat{W}\hat{U}$, but although this clearly leaves out reverberation (which is captured by $\hat{H}$), this low-rank approximation still entails some error. In order to avoid this, we estimate the clean spectrogram by multiplying the data elements $Y_{k,n}$ by a time-varying gain function $G_{k,n} \doteq \frac{\sum_j \hat{W}_{k,j}\hat{U}_{j,n}}{\sum_{j,m} \hat{W}_{k,j}\hat{U}_{j,n-m},\hat{H}_{k,m}}$, as suggested in [11]. Some results corroborating the importance of this step can be found in Section 5.2.3.

All steps of our dereverberation method are summarized in Algorithm 1.[1]

Next, we proceed to show some experimental results.

5 Experiments and results

In this section we present a series of experiments, firstly for parameter search and then for validating our method. All signals used in the experiments were taken from the TIMIT database ([20]), sampled at 16[kHz]. For the artificial RIR signals we made use of the software Room Impulse Response Generator[2]. The signals used for parameter estimation and those used for validation tests were uttered by different speakers, and the simulated RIRs correspond to different rooms.

In order to measure the quality of the restored signals, we used the well known frequency weighted segmental signal-to-noise ratio (fwsSNR) and the cepstral distance ([21]). Additionally, we computed the values of the speech-to-reverberation modulation energy ratio (SRMR, [22]). However, since the SRMR is non intrusive, its values must be used with caution for comparison purposes, keeping in mind that the resemblance of a restoration with the corresponding clean signal is not taken into account.

5.1 Parameter estimation

5.1.1 Estimation method for β_1

We begin by addressing the main parameter estimation problem for Stage 1 of Algorithm 1. Namely, finding an optimal value of β for building a dictionary whose atoms

[1] To try online: http://sinc.unl.edu.ar/web-demo/beta-dereverberation/

[2] https://github.com/ehabets/RIR-Generator

Algorithm 1 Variable β-divergence dereverberation

```
Preliminaries
  Given a speech signal $y$, build $Y_{k,n} = |\mathrm{STFT}(y)_{k,n}|^2$.
Stage 1
  Set $\beta = \beta_1$ and $\lambda_n^{(u)} = 0,\ \forall n$.
  Let $H_{k,n} = 1$ if $n = 1$ and $H_{k,n} = 0, \forall n \geq 2, \forall k$.
  Initialize $W^{(0)}$ and $U^{(0)}$ randomly.
  Let $t = 0$,
  while $\|W^{(t)} - W^{(t-1)}\|_F^2 > \delta$
    $t \leftarrow t + 1$
    Update $W^{(t)}$ as stated in (7).
    Update $U^{(t)}$ as stated in (8).
  end while
  Let $\hat{W} = W^{(t)}$
Stage 2
  Set $\beta = \beta_2$ and reset $\lambda_n^{(u)}, \lambda_k^{(h)}\ \forall n, k$.
  Let $H_{k,n}^{(0)} = \exp(1 - n),\ \forall n, k$.
  Initialize $U^{(0)}$ as the last approximation in Stage 1.
  Let $t = 0$,
  while $\|S^{(t)} - S^{(t-1)}\|_F^2 > \delta$
    $t \leftarrow t + 1$
    Update $U^{(t)}$ as stated in (8).
    Update $H^{(t)}$ as stated in (9).
  end while
  Let $\hat{U} = U^{(t)}$
  Let $\hat{H} = H^{(t)}$
Reconstruction
  Let $G_{k,n} \doteq \sum_j \hat{W}_{k,j}\hat{U}_{j,n} / \left(\sum_{j,m} \hat{W}_{k,j}\hat{U}_{j,n-m}, \hat{H}_{k,m}\right)$.
  Let $\hat{S}_{k,n} = G_{k,n} Y_{k,n}$.
  Define $Z \in \mathbb{C}^{K \times N}$ by $Z_{k,n} = \sqrt{\hat{S}_{k,n}}\,\arg(Y_{k,n})$.
  Define the restored signal in the time domain as
  $\hat{s} \doteq \mathrm{ISTFT}(Z)$.
```

(columns) be able to provide a good representation of the clean spectrogram. In order to evaluate whether a given parameter β_1 is good for dictionary building, we take a reverberant spectrogram Y, build a dictionary $W^{(\beta_1)}$ by minimizing $D_{\beta_1}(Y||WU)$, and then proceed to check how well can $W^{(\beta_1)}$ represent the corresponding clean spectrogram S. To do this, given β^*, we minimize $D_{\beta^*}(S||W^{(\beta_1)}U)$ with respect to U. It is important to point out that in this second step, β^* is not necessarily the same as β_1, and hence the two steps above are performed for every pair (β_1, β^*) in order to find the optimal.

Summarizing, given a reverberant spectrogram Y and each admissible pair (β_1, β^*), we take the following steps:

1. Build a dictionary $W^{(\beta_1)} = \arg\min_{W,U} D_{\beta_1}(Y||WU)$.

2. Use $W^{(\beta_1)}$ to find a representation $\hat{S} = W^{(\beta_1)}\hat{U}$ for the associated clean spectrogram S, where $\hat{U} = \arg\min_U D_{\beta^*}(S||W^{(\beta_1)}U)$.

3. Test the accuracy of the representation $\hat{S}$ by computing the cepstral distance with respect to S.

It is timely to point out that although this parameter estimation method can be costly, depending on the resolution of the grid over which (β_1, β^*) is defined, it has shown to be robust with respect to the reverberation conditions. This means it is not necessary to perform this experiment for every specific RIR, and the proposed method is intended to be used by simply setting the parameters as found optimal here.

5.1.2 Settings and results

To perform this experiment, we five random clean signals were taken from the TIMIT database and made reverberant by means of a discrete convolution with three different RIRs (450[ms], 600[ms], and 750[ms]). The speakers and the room dimensions are different than those later used for the validation experiments.

Fig. 4 depicts the resulting mean cepstral distance as a function of the parameters β_1 and β^*. The minimizer is reached at (0.75,1.45), showing that $\beta_1 = 0.75$ is the best parameter choice for Stage 1 of Algorithm 1. Note that this does not necessarily mean that $\beta_2 = 1.45$ is the best choice for the second stage of Algorithm 1, since here we are minimizing $D_\beta(S||\hat{S})$ whereas the second step of the dereverberation method requires minimizing $f(W, U, H)$ w.r.t. U and H (see Eq. 5).

It should be pointed out that functional (5) is a generalization of a Bayesian approach (similar to the one in [7]) if U and $\nabla_t H$ are treated as random variables with exponential and normal *a-priori* distributions, respectively. In fact, by choosing $\beta = 2$, the minimizer of (5) corresponds to a *maximum-a-posteriori* (MAP) estimator, given proper choices of the penalization parameters. Therefore, we have chosen $\beta = 2$ for Stage 2 of Algorithm 1, which in fact was observed to lead to better results than $\beta = 1.45$. This choice was corroborated by minimizing the cepstral distance over a

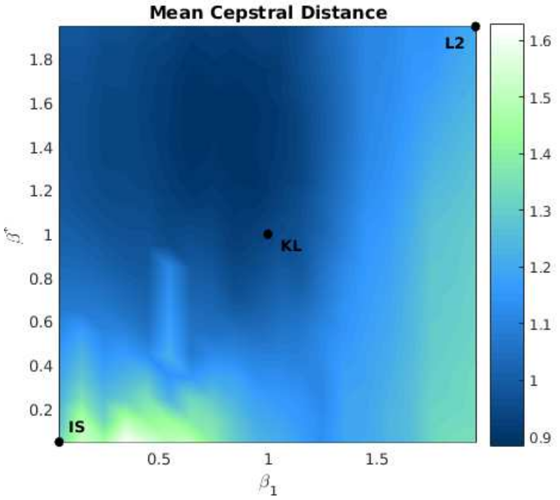


Figure 4: Mean cepstral distance values obtained from a representation of clean signals using a β^* divergence, with a dictionary built from reverberant versions using β_1 . Smaller values correspond to better results The marks IS, KL and L2 indicate the results using the Itakura-Saito divergence, Kullback-Leibler divergence and Frobenius norm, respectively.

uniform grid of values of β. Detailed explanations on the relation between Bayesian and penalization approaches can be found in [23], along with other possible Bayesian interpretations.

A few relevant conclusions can be derived by observing Fig 4. First, that the values of (β_1, β^*) leading to the smallest cepstral distances are away from the diagonal, thus corroborating our original conjecture that using different parameter values for the learning and representation steps could lead to improved results. Furthermore, note that better results are obtained for values of (β_1, β^*) in the top left area. This most probably reflects the fact that small values of β_1 lead to dictionaries which take all the frequency range into account, whereas high values of β^* promote fidelity on the high-energy zones of the represented spectrogram.

It is reasonable to expect the optimal value of β_1 to depend on the reverberation conditions. In fact, the optimal values for β_1 were found to be $\beta_1 = 0.75$ for reverberation times of 450[ms] and 600[ms], and $\beta_1 = 0.85$ for 750[ms]. This suggests that the choice of β_1 could be further tuned up within a supervised setting, but the method is robust enough to cope with moderate variations on the reverberation conditions.

5.1.3 Other parameters

The choice of parameters was made taking information from the reverberant signal spectrogram into account. As customary ([11]), given that the norms of the columns of W were set equal to 1 and the elements of U are meant to reconstruct the clean spectrogram, $\lambda_n^{(u)}$ was chosen proportional to the mean value of Y. On the other

win. size	win. overl.	J	M	β_1	β_2
512	256	64	20	0.75	2

$\lambda_n^{(u)}$	$\lambda_k^{(h)}$	δ
$mean(Y) \times 10^{-3}$	$0.3\|Y_k\|^2$	$\|Y\| \times 10^{-3}$

Table 1: Model parameters

hand, given that reverberation can depend on the frequency range, $\lambda_k^{(h)}$ was chosen proportional to $\|Y_k\|^2$, $k = 1, \dots, K$. After these considerations, the method showed to be robust enough with respect to the regularization parameters for the minimization of the cepstral distance over a coarse logarithmic grid to provide a good estimation. Two randomly chosen signals (from a male and a female speaker not in the validation set) were artificially made reverberant (600[ms] RIR) and then used for computing the values depicted in Table 1. Nonetheless, it is worth mentioning that finding an optimal way of choosing regularization parameters remains an open problem.

The number of columns of the matrices W and H (J and M, respectively) were empirically chosen. The number of dictionary atoms has to be large enough to provide a good representation of the clean spectrogram but not so large to allow for a trivial representation. In the extreme case, if $J >> N$, then W could contain atoms solely devoted to model reverberant components. In regard to M, it simply has to be large enough to capture most of the RIR, though an extremely large value of M along with a $\lambda_k^{(h)}$ not large enough may result in little oscillations on H leading to a degraded representation. It should be pointed out that no issues were observed within mild variations in these parameters.

5.2 Validation

We have chosen two different settings for the validation experiments. The first one using simulations in order have a large number of trials available, and the second one using real recordings to guarantee the method is applicable in real-life conditions.

The divergence parameters were fixed as $\beta_1 = 0.75$ and $\beta_2 = 2$. The rest of the parameters used for all the experiments are detailed in Table 1.

In order to evaluate the performance of our method, comparisons against three state-of-the-art methods applicable under the same conditions were made. The first one was proposed by Wisdom *et al* in [6], and showed an excellent performance in the Reverb Challenge ([24]). The second one, by Mohammadiha *et al* ([11]) is an early attempt to incorporate a dictionary-based approach to a convolutive NMF approach. The third one ([7]) uses a Bayesian approach, and it has shown to perform quite well, yet posing the frequency decorrelation issue mentioned in the introduction.

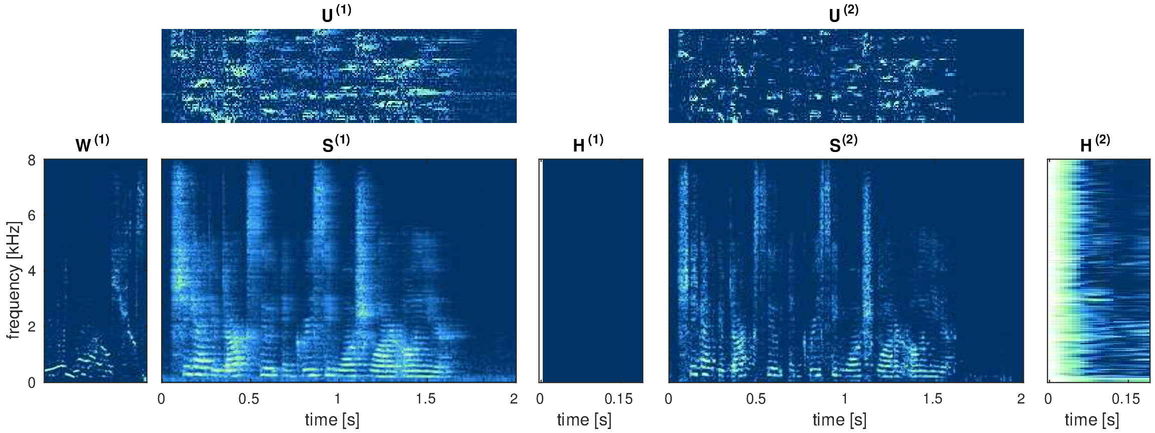


Figure 5: Representation elements obtained with the proposed method. $W^{(1)}$, $U^{(1)}$, $H^{(1)}$, and $S^{(1)}$=$W^{(1)}U^{(1)}$ are the matrices at the end of Stage 1, and $U^{(2)}$, $H^{(2)}$, and $S^{(2)} = W^{(1)}U^{(2)}$ are the matrices at the end of the dereverberation process. All the elements are in log scale, in amplitude.

5.2.1 Performance illustration

Before beginning with the actual experiments we show how the method works by plotting the result obtained for just one signal. The signal corresponds to a female speaker pronouncing the sentence "She had your dark suit in greasy wash water all year", from the TIMIT database, recorded in an office room (Room 1, in Table 4) in real-life conditions, as specified in Section 5.2.3. All representation elements are depicted in Fig. 5. It can be seen that at the end of Stage 1, a dictionary $W^{(1)}$ is built while reverberation is captured in the coefficient matrix $U^{(1)}$. In the second stage, reverberation is mostly represented by $H^{(2)}$, thus allowing the coefficients in $U^{(2)}$ to provide a good representation $S^{(2)}$ of the clean spectrogram S.

5.2.2 Simulated experiments

For the simulations, 110 speech signals from the TIMIT database were taken, and made reverberant by convolution with artificial impulse responses. The artificial RIRs were generated varying the microphone positions and room dimensions, as specified in Table 2. The reverberation time was set at either 450[ms], 600[ms] or 750[ms], resulting in 27 different reverberation conditions, and hence a total of 2970 reverberant signals for testing.

Table 3 and Fig. 6 show the results obtained with each performance measure and each one of the methods. A t-test was performed to test the hypothesis that the underlying distributions of the results have -or do not have- the same means. This has shown our proposed method (labeled "Beta") outperforms ($p < 0.01$) the other three in terms of fwsSNR and cepstral distance, but not the Bayesian ([7]) in terms of SRMR. However, taking into account that SRMR quantifies the extent to which a signal "seems"

	Length	Width	Height
Room 1 dimensions	5.00 [m]	4.00 [m]	6.00 [m]
Room 2 dimensions	4.00 [m]	4.00 [m]	3.00 [m]
Room 3 dimensions	10.0 [m]	4.00 [m]	5.00 [m]
Source position	2.00 [m]	3.50 [m]	2.00 [m]
Microphone 1 position	2.00 [m]	1.50 [m]	1.00 [m]
Microphone 2 position	2.00 [m]	2.00 [m]	1.00 [m]
Microphone 3 position	2.00 [m]	2.00 [m]	2.00 [m]

Table 2: Simulated room settings

Measure	fwsSNR	Cepstral Dist.	SRMR
Reverberant	5.377 (1.70)	5.308 (0.61)	2.470 (1.01)
Wisdom	5.593 (1.67)	5.279 (0.60)	2.898 (1.14)
Mohammadiha	5.939 (1.56)	5.117 (0.61)	3.183 (1.28)
Bayesian	7.604 (1.60)	4.614 (0.52)	**4.423** (1.48)
Beta	**8.153** (1.51)	**4.573** (0.48)	3.751 (1.21)

Table 3: Mean and standard deviation (between parenthesis) of performance measures for each method, using simulations. Best results are shown in boldface.

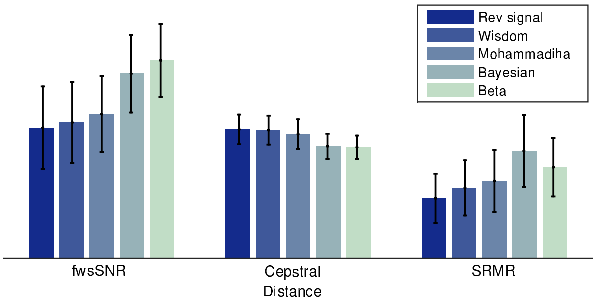


Figure 6: Mean and standard deviation of performance measures for each method, using simulations.

reverberant, but not how much such a restoration resembles the corresponding clean signal, it should only be considered as a complement to the other two measures.

5.2.3 Experiments using recordings

In order to test whether our method works in real-life situations, we made recordings in two of our own office rooms, during standard office hours and with air conditioners and computers left on. The offices' dimensions are shown in Table 4, along with the speaker and microphone positions. The reverberation times of the rooms turned out to be of 460[ms] in Room 1 and of 440[ms] in Room 2, as measured using sine sweeps ([25]). Four speakers (two male and two female) were randomly selected from the TIMIT database,

	Length	Width	Height
Room 1 dimensions	4.15 [m]	3.00 [m]	3.00 [m]
Source 1 position	3.60 [m]	1.50 [m]	1.50 [m]
Microphone 1 position	1.10 [m]	1.50 [m]	1.50 [m]
Room 2 dimensions	5.85 [m]	4.55 [m]	3.00 [m]
Source 2 position	1.10 [m]	1.50 [m]	1.50 [m]
Microphone 2 position	1.10 [m]	4.00 [m]	1.50 [m]

Table 4: Office rooms settings

Measure	fwsSNR	Cepstral Dist.	SRMR
Reverberant	3.613 (1.52)	4.994 (0.56)	2.756 (0.75)
Wisdom	4.917 (1.37)	4.577 (0.43)	3.222 (0.77)
Mohammadiha	4.431 (1.48)	5.172 (0.78)	3.627 (1.00)
Bayesian	6.254 (1.33)	4.769 (0.60)	**4.809** (1.10)
Beta	**6.678** (1.18)	**4.524** (0.53)	4.036 (0.84)

Table 5: Mean and standard deviation (between parenthesis) of performance measures for each method. Best results are shown in boldface.

and 10 speech signals from each were recorded in each room, with a sampling frequency of 16[kHz].

As it is customary, the clean speech sources had their low-frequency components filtered out. Hence, we pre-processed our reverberant recordings using a 5000 tap FIR high-pass filter with cut-off frequency of 30[Hz] to mitigate the low frequency noise. For the comparisons to be fair, all the methods were tested after this pre-processing was made.

In order to better cope with the noise, the penalization parameters for U were reset to $\lambda_n^{(u)} = \frac{mean(Y)}{\|U_n^1\|_1} \times 10^{-1}$, where U_n^1 is the n-th column of U as estimated at the end of Stage 1 of Algorithm 1. Given that Stage 1 is somewhat equivalent to a non-regularized NMF factorization of a noisy, reverberant signal, ambient noise during speech silences will tend to be represented by linear combinations of the atoms with small coefficients. Hence, we can use this information at the end of Stage 1 to augment penalization in these areas simply dividing the penalization parameter by $\|U_n^1\|_1$.

Results are depicted in Table 5 and illustrated in Figure 7. Once again, we see that our proposed method outperforms the others in terms of the fwsSNR, but loses to the Bayesian in terms of SRMR. As for the cepstral distance, the improvement between our proposed method and Wisdom's is the only one not reaching statistical significance ($p > 0.01$).

In order to illustrate the relevance of the gain operation at the end of Algorithm 1, the experiment was also run omitting this last step. The obtained values were fwsSNR = 5.952, Cepstral Distance = 4.852 and SRMR = 3.920, which compared to the results observed in Table 5 highlight the importance of this operation.

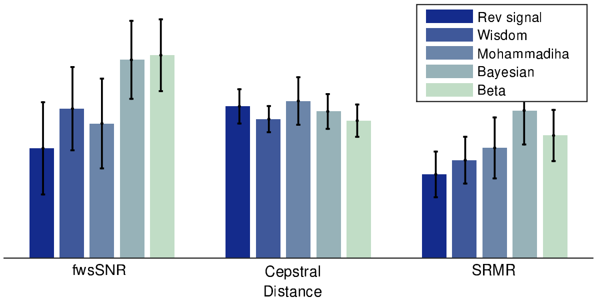


Figure 7: Mean and standard deviation of performance measures for each method, using recordings.

6 Conclusions

In this work, a new blind, single channel dereverberation method in the time-frequency domain that makes use of variable β-divergence as a cost function was presented and tested. The method comprises two stages: one for learning the spectral structure into a dictionary, and a second one for using such a dictionary to build an accurate representation by means of a convolutive NMF model. The corresponding algorithm for implementing the method was introduced and tested. Additionally, a method for finding an optimal learning divergence was introduced.

Results show that the proposed method improves restoration quality with respect to state-of-the-art methods, as measured by the fwsSNR and cepstral distance. Improvement in regard to SRMR is only partial, but given that it is a non-intrusive measure, that is not too much of a drawback.

There is certainly much room for improvement. For instance, exploring the use of penalization terms at the learning stage and other ways of enhancing the quality of the dictionary, as well as generating atoms for specifically modeling (and then removing) noise and incorporating specific initialization methods. All this is subject of future study.

Finally, although our method is constructed for a blind setting, it is worth noting that it can be easily adapted to be supervised by modifying the learning stage, provided speaker information is available.

Appendices

A Auxiliary function for W

We shall build an auxiliary function for (5) w.r.t. W. To do so, let $W' \in \mathbb{R}_+^{K\times J}$, and let us denote $X'_{k,n} = \sum_{j,m} W'_{k,j} U_{j,n-m} H_{k,m}$. Then,

$$\begin{aligned}\check{D}_\beta(Y_{k,n}||X_{k,n}) &= \check{D}_\beta\left(Y_{k,n}\middle|\middle|\sum_{j,m} W_{k,j}U_{j,n-m}H_{k,m}\right) \qquad (10)\\ &= \check{D}_\beta\left(Y_{k,n}\middle|\middle|\frac{\sum_{j,m} W_{k,j}U_{j,n-m}H_{k,m}X'_{k,n}\frac{W'_{k,j}}{W'_{k,j}}}{X'_{k,n}}\right)\\ &= \check{D}_\beta\left(Y_{k,n}\middle|\middle|\frac{\sum_{j,m} W'_{k,j}U_{j,n-m}H_{k,m}X'_{k,n}\frac{W_{k,j}}{W'_{k,j}}}{\sum_{j,m} W'_{k,j}U_{j,n-m}H_{k,m}}\right)\\ &\le \sum_{j,m}\frac{W'_{k,j}U_{j,n-m}H_{k,m}}{X'_{k,n}}\check{D}_\beta\left(Y_{k,n}\middle|\middle|X'_{k,n}\frac{W_{k,j}}{W'_{k,j}}\right),\end{aligned}$$

where the last step is due to Jensen's inequality.

In regard to $\hat{D}_\beta$, since it is concave w.r.t. X, it follows that

$$\hat{D}_\beta(Y_{k,n}||X_{k,n}) \le \hat{D}_\beta(Y_{k,n}||X'_{k,n}) + \frac{\partial \hat{D}_\beta(Y_{k,n}||X'_{k,n})}{\partial X_{k,n}}\sum_{j,m}(W_{k,j}-W'_{k,j})U_{j,n-m}H_{k,m}. \qquad (11)$$

Given U and H fixed, let us define $g_w : \mathbb{R}_+^{K\times J}\times\mathbb{R}_+^{K\times J}\to\mathbb{R}$ by

$$\begin{aligned}g_w(W,W') \doteq &\sum_{k,n,j,m}\frac{W'_{k,j}U_{j,n-m}H_{k,m}}{X'_{k,n}}\check{D}_\beta\left(Y_{k,n}\middle|\middle|X'_{k,n}\frac{W_{k,j}}{W'_{k,j}}\right)+\sum_{k,n}\hat{D}_\beta(Y_{k,n}||X'_{k,n}).\\ &+\sum_{k,n,j,m}\frac{\partial \hat{D}_\beta(Y_{k,n}||X'_{k,n})}{\partial X_{k,n}}(W_{k,j}-W'_{k,j})U_{j,n-m}H_{k,m}.\end{aligned}$$

Then, it follows from (10) and (11) that g_w is an auxiliary function for f w.r.t. W. Note that the equality condition in Definition 4.1 also holds.

B Updating rule for U

An analogous procedure as that in Appendix A shows that an auxiliary function for f with respect to U is given by

$$g_u(U,U') \doteq \sum_{k,n,j,m}\frac{W_{k,j}U'_{j,m}H_{k,n-m}}{X'_{k,n}}\check{D}_\beta\left(Y_{k,n}\middle|\middle|X'_{k,n}\frac{U_{j,m}}{U'_{j,m}}\right)$$

$$\sum_{k,n,j,m} \frac{\partial \hat{D}_\beta(Y_{k,n}||X'_{k,n})}{\partial X_{k,n}} W_{k,j}(U_{j,m} - U'_{j,m})H_{k,n-m} + \sum_{k,n} \hat{D}_\beta(Y_{k,n}||X'_{k,n}) + \sum_{j,n} \lambda_n^{(u)} U_{j,n}.$$

Here again, since $g_u(U, \cdot)$ is convex, it can be minimized by equating its gradient to zero, which is tantamount to solving

$$U_{j,m} = U'_{j,m} \left(\frac{\sum\limits_{k,n} X'^{\beta-2}_{k,n} Y_{k,n} W_{k,j} H_{k,n-m} - \lambda_m^{(u)} \left(\frac{U'_{j,m}}{U_{j,m}} \right)^{\alpha_2}}{\sum\limits_{k,n} X'^{\beta-1}_{k,n} W_{k,j} H_{k,n-m}} \right)^{\eta}.$$

Let us notice that this is an implicit equation for $U_{j,m}$ for $\beta < 2$ (and $\lambda_j^{(u)} \neq 0$). Nonetheless, a simpler updating rule can be derived. Firstly, note that since g_u is an auxiliary function for f w.r.t. U, Lemma 4.2 guarantees that $f(U^{(t)})$ approaches a limit as t tends to infinity. Although this does not ensure the convergence of $U^{(t)}$ as $t \to \infty$, this was if fact observed to be the case in all the performed experiments. It is timely to point out, however, that under certain additional hypotheses on the auxiliary function g (e.g. strongly convex on the first variable, uniformly with respect to the second one) it can be rigorously proved that $U^{(t)}$ indeed converges as t tends to infinity.

Under this convergence assumption, the quotient $U_{j,m}^{(t)}/U_{j,m}^{(t-1)}$ approaches 1, and hence the approximation $U_{j,m}^{(t)}/U_{j,m}^{(t-1)} \approx 1$ yields the multiplicative updating rule in (8). It is timely to note that this updating equation can be shown to be equivalent to that derived using a gradient descent approach, as explained in [26].

C Updating rule for H

It can be shown, by a similar procedure as that in Appendix A, that

$$g_h(H, H') \doteq \sum_{k,n,j,m} \frac{W_{k,j} U_{j,n-m} H'_{k,m}}{X'_{k,n}} \check{D}_\beta \left(Y_{k,n} \middle|\middle| X'_{k,n} \frac{H_{k,m}}{H'_{k,m}} \right) + \sum_{k,n,j,m} \frac{\partial \hat{D}_\beta(Y_{k,n}||X'_{k,n})}{\partial X_{k,n}} (H_{k,m} - H'_{k,m}) W_{k,j} U_{j,n-m} + \sum_{k,n} \hat{D}_\beta(Y_{k,n}||X'_{k,n}) + \sum_k \lambda_k^{(h)} \|L H_k^T\|^2,$$

constitutes an auxiliary function for f w.r.t. H. By equating its gradient (with respect to $H_{k,m}$) to zero, we obtain, for every $k = 1, \ldots, K, m = 1, \ldots, M$,

$$\begin{aligned} 0 = &\sum_{j,n} W_{k,j} U_{j,n-m} \left(X'_{k,n}\right)^{\alpha_1} \left(\frac{H_{k,m}}{H'_{k,m}}\right)^{\alpha_1} - \sum_{j,n} W_{k,j} U_{j,n-m} Y_{k,n} \left(X'_{k,n}\right)^{\alpha_2} \left(\frac{H_{k,m}}{H'_{k,m}}\right)^{\alpha_2} \\ &- 2\lambda_k^{(h)} [L^T L H_k^T]_m. \end{aligned}$$

During the experiments, we have observed that using a multiplicative updating rule analogous to those used for $W^{(t)}$ and $U^{(t)}$ might incur in undesired oscillations in the elements of $H^{(t)}$. This is most likely due to the alternating signs in the rows of $L^T L$. In order to overcome this potential drawback, we define, for every $k = 1, \ldots, K$, the diagonal matrix $A^{(k)} \in \mathbb{R}_{0,+}^{M \times M}$ with $A_{m,m}^{(k)} = \sum_{j,n} W_{k,j} U_{j,n-m} \left(X_{k,n}^{(t-1)}\right)^{\alpha_1} / H_{k,m}^{(t-1)}$ and the vector $b^{(k)} \in \mathbb{R}_{0,+}^{M}$ as $b^{(k)} = \sum_{j,n} W_{k,j} U_{j,n-m} Y_{k,n} \left(X_{k,n}^{(t-1)}\right)^{\alpha_2}$. Then, under the same approximation used in Appendix B, we can update H by solving for $H_k^{(t)}$, $k = 1, \ldots, K$, the linear system

$$\left(A^{(k)} + 2\lambda_k^{(h)} L^T L\right) H_k^{(t)} = b^{(k)}. \tag{12}$$

It can be shown that the matrix $A^{(k)} + 2\lambda_k^{(h)} L^T L$ is strictly positive definite (unless $A^{(k)}$ is null), and hence the linear system (9) has a unique solution, whose elements are non-negative.

Acknowledgments

This research was funded by ANPCyT under projects PICT 2014-2627 and PICT 2015-0977, by UNL under projects CAI+D 50420150100036LI, CAI+D 50020150100059LI, CAI+D 50020150100055LI and CAI+D 50020150100082LI.

References

[1] S. Yun, Y. J. Lee, and S. H. Kim, “Multilingual speech-to-speech translation system for mobile consumer devices,” *IEEE Transactions on Consumer Electronics*, vol. 60, no. 3, pp. 508–516, 2014.

[2] L. D. Vignolo, S. R. M. Prasanna, S. Dandapat, H. L. Rufiner, and D. H. Milone, “Feature optimisation for stress recognition in speech,” *Pattern Recognition Letters*, vol. 84, pp. 1–7, 2016.

[3] R. Sarikaya, P. A. Crook, A. Marin, M. Jeong, J.-P. Robichaud, A. Celikyilmaz, Y.-B. Kim, A. Rochette, O. Z. Khan, X. Liu *et al.*, “An overview of end-to-end language understanding and dialog management for personal digital assistants,”

in *Spoken Language Technology Workshop (SLT), 2016 IEEE.* IEEE, 2016, pp. 391–397.

[4] I. J. Tashev, *Sound capture and processing: practical approaches.* John Wiley & Sons, 2009.

[5] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development.* Prentice hall PTR Upper Saddle River, 2001, vol. 95.

[6] S. Wisdom, T. Powers, L. Atlas, and J. Pitton, "Enhancement of reverberant and noisy speech by extending its coherence," in *Proceedings of REVERB Challenge Workshop*, 2014, pp. 1–8.

[7] F. Ibarrola, L. Di Persia, and R. Spies, "A bayesian approach to convolutive non-negative matrix factorization for blind speech dereverberation," *Signal Processing*, vol. 151, pp. 89–98, 2018.

[8] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," *Proceedings of the 5th Conference on Independent Component Analysis and Blind Signal Separation*, pp. 494–499, 2004.

[9] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 45–48.

[10] F. Ibarrola, L. Di Persia, and R. Spies, "On the use of convolutive nonnegative matrix factorization with mixed penalization for blind speech dereverberation," in *2017 XLIII Latin American Computer Conference (CLEI).* IEEE, 2017, pp. 1–4.

[11] N. Mohammadiha, P. Smaragdis, and S. Doclo, "Joint acoustic and spectral modeling for speech dereverberation using non-negative representations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2015, pp. 4410–4414.

[12] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[13] B. Yegnanarayana, P. S. Murthy, C. Avendaño, and H. Hermansky, "Enhancement of reverberant speech using LP residual," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1. IEEE, 1998, pp. 405–408.

[14] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[15] R. Kompass, “A generalized divergence measure for nonnegative matrix factorization,” *Neural computation*, vol. 19, no. 3, pp. 780–791, 2007.

[16] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O‘Brien Jr, C. R. Lansing, and A. S. Feng, “Blind estimation of reverberation time,” *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[17] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Processing Systems*, 2001, pp. 556–562.

[18] C. Févotte and J. Idier, “Algorithms for nonnegative matrix factorization with the β-divergence,” *Neural computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[19] S. Choi, A. Cichocki, H.-M. Park, and S.-Y. Lee, “Blind source separation and independent component analysis: A review,” *Neural Information Processing-Letters and Reviews*, vol. 6, no. 1, pp. 1–57, 2005.

[20] V. Zue, S. Seneff, and J. Glass, “Speech database development at MIT: TIMIT and beyond,” *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[21] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[22] T. H. Falk, C. Zheng, and W.-Y. Chan, “A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[23] R. Gribonval and M. Nikolova, “On Bayesian estimation and proximity operators,” *arXiv preprint arXiv:1807.04021*, 2018.

[24] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, “A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research,” *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[25] A. Farina, “Advancements in impulse response measurements by sine sweeps,” in *Audio Engineering Society Convention 122*. Audio Engineering Society, 2007.

[26] A. Cichocki, R. Zdunek, and S. Amari, “Csiszar’s divergences for non-negative matrix factorization: Family of new algorithms,” in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2006, pp. 32–39.

Penalized nonnegative representations for speech separation

Penalized nonnegative representations for speech separation

Francisco J. Ibarrola *[1], Ruben D. Spies[2], and Leandro E. Di Persia[1]

[1]Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL, CONICET, Ciudad Universitaria UNL, (3000) Santa Fe, Argentina.

[2]Instituto de Matemática Aplicada del Litoral, IMAL, CONICET-UNL, Centro Científico Tecnológico CONICET Santa Fe, Colectora Ruta Nac. 168, km 472, Paraje "El Pozo", 3000, Santa Fe, Argentina and Departamento de Matemática, Facultad de Ingeniería Química, Universidad Nacional del Litoral, Santa Fe, Argentina.

Abstract

In this work we address the problem of supervised audio source separation within a reverberant environment. We make use of a nonnegative representation in order to model the mixture along with reverberation. This kind of models often pose the problem that the number of variables to learn is large with respect to the data, which is to say there are many possible choices of the elements that result in the same approximation of the mixture. We use a probabilistic approach in order to derive a penalized cost function that aims to overcome this issue by inducing a certain structure over the representation elements. Preliminary results account for a considerable improvement in restoration quality with the introduction of penalizers.

Keywords: signal processing, bayesian model, reverberation
2000 AMS Subject Classification: 92C55 - 68T10

1 Introduction

One of the main problems arising in audio signal processing is that of source separation. That is, given a recording of sound coming from two or more sources, we want to isolate the signals produced by each one.

In this work we shall use a nonnegative matrix factorization (NMF) model to address the problem of speech separation. Some early approaches have made use of training data in order to build a dictionary that encompass the spectral characteristics of each speaker and separate the sources by isolating the activation components ([6]). This works reasonably well for purely additive mixtures, but in reality we are usually faced with reverberation and noisy environments. A more recent approach ([5]) made use of a mixture of NMF and convolutive NMF (CNMF) in order to address this issue. This kind of model entails learning many coefficients from limited data, and thus poses a practical problem: an observed power spectrogram can be accurately represented by a set of parameters that is not representative of the phenomenon we intend to model.

In this work we make use of a Bayesian approach to define proper penalization terms over a cost function. Then, we build an algorithm for minimizing such function which results in effectively learning a mixed NMF-CNMF representation by inducing certain structure over its elements.

2 A reverberant mixture model

Let us consider a setting with I speakers and R microphones. In order to model the reverberant mixture, we begin by defining the continuous, compactly supported functions $s_i, h_{r,i} : \mathbb{R} \to \mathbb{R},\ i =$

[*] fibarrola@sinc.unl.edu.ar

$1, \ldots, I,\ r = 1, \ldots R$, where s_i is the signal from the i-th source, and $h_{i,r}$ is the impulse response signal from the i-th source to the j-th microphone. Then, under the hypothesis that the phenomenon can be accurately represented by a linear time invariant (LTI) system, we can define

$$x_r(t) \doteq \sum_{i=1}^{I} (h_{r,i} * s_i)(t), \quad r = 1, \ldots, R, \tag{1}$$

where x_r is an approximation to the recording y_r, obtained from the r-th microphone.

Since speech signals present large oscillations, we switch to the time frequency domain by means of the Short Time Fourier Transform (STFT). That is, we define $Y_{k,n,r} \doteq |\hat{y}_{r;k}(n)|^2$, $X_{k,n,r} \doteq |\hat{x}_{r;k}(n)|^2$, $H_{k,n,r,i} \doteq |\hat{h}_{r,i;k}(n)|^2$ and $S_{k,n,i} \doteq |\hat{s}_{i;k}(n)|^2$, where $\hat{\cdot}_k(n)$ denotes the STFT at frequency k and time n. With these definitions, Equation 1 leads to

$$Y_{k,n,r} = X_{k,n,r} + \epsilon_{k,n,r} = \sum_{i=1}^{I} \sum_{m=1}^{M} H_{k,m,r,i} S_{k,n-m+1,i} + \epsilon_{k,n,r}, \tag{2}$$

where $\epsilon \in \mathbb{R}^{K \times N \times R}$ is a tensor modeling both the representation error and noise. Details on how this model can be built from (1) can be found in [2].

Finally, let us assume that each source signal can be well represented by using NMF. This means that $\exists W \in \mathbb{R}_{0,+}^{K \times J \times I}, U \in \mathbb{R}_{0,+}^{J \times N \times I}$ such that $S_{k,n,i} \approx \sum_j W_{k,j,i} U_{j,n,i}$. Making a small abuse of notation, model (2) now reads

$$Y_{k,n,r} = X_{k,n,r} + \epsilon_{k,n,r} = \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{j=1}^{J} H_{k,m,r,i} W_{k,j,i} U_{j,n-m+1,i} + \epsilon_{k,n,r}. \tag{3}$$

We now need a way to find representation elements in (3) that allow for a good separation.

3 Cost function

Given that we do not know the model components, we shall treat them as realizations of random variables, and from there, build a cost function whose minimizer will provide a good representation.

Firstly, given that no information is available on the error, we shall assume $\epsilon_{k,n,r}$ to be a realization of a zero-mean normal distribution. This corresponds to $Y_{k,n,r}$ having the following distribution, conditioned on $X_{k,n,r}$:

$$\pi_{like}(Y_{k,n,r}|X_{k,n,r}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_{k,n,r} - X_{k,n,r})^2}{2\sigma^2}\right).$$

On the other hand, an NMF representation of a clean spectrogram is expected to exhibit a sparse use of the atoms (columns of W) to build a representation, unlike the smooth structure expected for a reverberant one. Sparsity can be favored by assuming the elements of U to be realizations of an exponential distribution. This corresponds to the following Probability Density Function (PDF):

$$\pi_{prior}(U_{j,n,i}) = \frac{\lambda_u}{2} \exp(-\frac{\lambda_u}{2} U_{j,n,i}).$$

Finally, we would expect the components of the impulse response tensor H to exhibit a smooth decay over time. This can be induced by assuming the time gradient of the associated random tensor has a normal distribution. If we let $\underline{H}_{k,r,i}$ be the row vector with components $H_{k,m,r,i}, m = 1, \ldots, M$, this leads to

$$\pi_{prior}(\underline{H}_{k,r,i}) = \frac{1}{\det(2\pi(L\Sigma^{-1}L^T)^{-1})} \exp\left(-\frac{1}{2}\underline{H}_{k,r,i} L\Sigma^{-1}L^T \underline{H}_{k,r,i}^T\right),$$

where L is a finite difference matrix associated to the gradient, and $\Sigma \doteq \frac{1}{\lambda_h} I_{(M-1\times M-1)}$, for some $\lambda_h > 0$.

In order to get a representation whose elements are representative of the aforementioned PDFs, we can find the *maximum-a-posteriori* (MAP) estimator, which amounts to minimizing

$$-\log \pi_{post}(X|Y) = -\log[\pi_{like}(Y|X)\pi_{prior}(U)\pi_{prior}(H)].$$

Under the assumption that the underlying random variables are uncorrelated, this is equivalent to minimizing

$$f(U,H) \doteq \sum_{k,n,r} (Y_{k,n,r} - X_{k,n,r})^2 + \sigma^2\lambda_u \sum_{j,n,i} U_{k,n,i} + \sigma^2\lambda_h \sum_{k,r,i} \|\underline{H}_{k,r,i} L\|^2. \tag{4}$$

Next, we introduce a procedure for minimizing this cost function.

4 Optimization

In order to minimize the cost function f, defined in (4), we resort to a minimization-majorization method. This essentially consists of building a new function in a larger space that is easier to minimize than f, and use it to iteratively approach a minimizer of f.

Let $\Omega \subset \mathbb{R}^P$ and $f : \Omega \to \mathbb{R}_0^+$. Then, $g : \Omega \times \Omega \to \mathbb{R}_0^+$ is called an *auxiliary function* for f if $g(\omega, \omega) = f(\omega)$ and $g(\omega, \omega') \geq f(\omega), \ \forall \omega, \omega' \in \Omega$.

Then, given an arbitrary $\omega^{(0)} \in \Omega$ and $\omega^t \doteq \arg\min_\omega g(\omega, \omega^{t-1}), \ t \in \mathbb{N}$, it can be shown ([4]) that the sequence $\{f(\omega^t)\}_{t\geq 1}$ is non-increasing.

With this in mind, we can define an auxiliary function for f with respect to U as

$$g_u(U, U') \doteq \sum_{k,n,j,m,r} \frac{W_{k,j,i} U'_{j,m,i} H_{k,n-m+1,i,r}}{X'_{k,n,i}} \left(Y_{k,n,r} - X'_{k,n,r} \frac{U_{j,m,i}}{U'_{j,m,i}} \right)^2 + \sigma^2\lambda_u \sum_{j,n,i} U_{j,n,i},$$

where $U'_{j,m,i}$ is arbitrary, and $X'_{k,n,r} \doteq \sum_{i,m,j} H_{k,m,r,i} W_{k,j,i} U'_{j,n-m+1,i}$. The proof that this is indeed an auxiliary function for f w.r.t. U is omitted due to space limitations, but the reader is referred to [3] for an analogous, detailed proof.

Now, given that g_u is quadratic w.r.t. U, it can be minimized simply by meeting its first order necessary condition, which readily leads to the updating rule

$$U^{(t)}_{j,m,i} = U^{(t-1)}_{j,m,i} \frac{\left[\sum\limits_{k,n,r} Y_{k,n,r} W_{k,j,i} H_{k,n-m+1,r,i} - \sigma^2\lambda_u \right]_\epsilon}{\sum\limits_{k,n,r} X^{(t-1)}_{k,n,r} W_{k,j,i} H_{k,n-m+1,r,i}}, \tag{5}$$

where the operation $[\cdot]_\epsilon \doteq \max\{\cdot, \epsilon\}$ is meant to preclude the elements of U from dropping to zero (or below), given that within a multiplicative rule, if an element drops to zero it cannot regain positive values.

An analogous procedure leads to an iterative updating rule for H. By defining, for every k, r and i, the diagonal matrix $A^{(k)} \in \mathbb{R}^{M\times M}_{0,+}$ with $A^{(k,r,i)}_{m,m} = \sum_{j,n} W_{k,j} U_{j,n-m} X^{(t-1)}_{k,n}$ and the vector $b^{(k,r,i)} \in \mathbb{R}^M_{0,+}$ as $b^{(k,r,i)}_m = \sum_{j,n} W_{k,j,i} U_{j,n-m+1,i} Y_{k,n,r} H^{(t-1)}_{k,m,r,i}$. Then, H can be updated by solving for $\underline{H}^{(t)}_{k,r,i}$, the linear system

$$\left(A^{(k)} + \sigma^2\lambda_u L^T L \right) \left(\underline{H}^{(t)}_{k,r,i} \right)^T = b^{(k)}. \tag{6}$$

It can be shown that the matrix $A^{(k)} + \sigma^2\lambda_u L^T L$ is strictly positive definite (unless $A^{(k)}$ is null), and hence the linear system (6) has a unique solution, whose elements are non-negative.

All of the above shows that the cost function f can be minimized by iteratively and sequentially updating U and H according to (5) and (6).

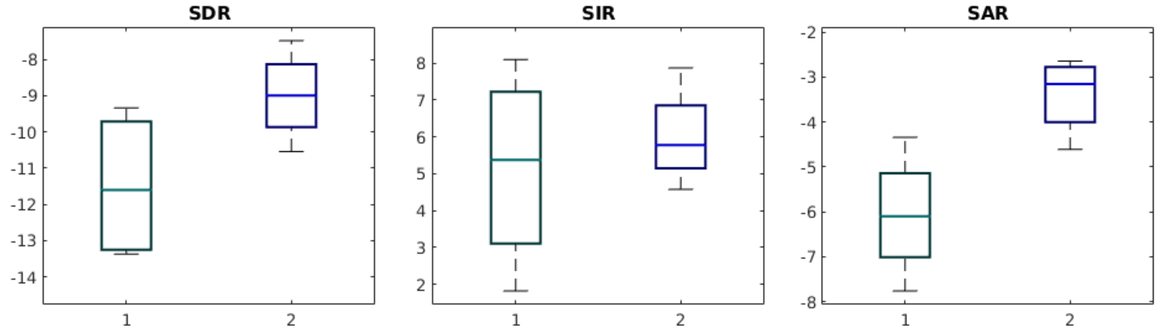


Figure 1: Separation measures results. Those obtained without penalization are on the left, and those using penalization on the right on every plot.

5 Experiments

In order to analyze the improvement (if any) accomplished by the introduction of the penalizers into the model, we tested our method against the one described in [5]. To do so, we randomly chose two male and two female speakers from the TIMIT database ([7]). We then chose a signal from each and built an artificial mixture by convolution with impulse responses generated with the software Room Impulse Response Generator[1], based on the model in [1]. The reverberation time was set to 450[ms].

The dictionaries for each speaker were built from seven signals, different from those used for the mixture. We do not delve into details on this matter due to space limitations, but the reader is again referred to [3] for details.

In order to evaluate the results, we used three standard separation measures: the Signal-to-Distortion Ratio (SDR), the Signal-to-Interference Ratio (SIR) and the Signal-to-Amplitude Ratio (SAR). Figure 1 depicts the obtained results, which clearly suggest an improvement when using our penalization approach.

6 Discussion

A penalization model based on a Bayesian approach over a mixed NMF model was introduced and tested. Although the results are preliminary, they clearly suggest a quality increment over the standard NMF-CNMF approach.

There is certainly much room for improvement. For one thing, exploring the use of probability density functions that take the correlation between the variables into account. Also, the model parameters can be set to depend on the speaker or frequency band and a way to optimally choose them is yet to be found. Finally, it is worth mentioning that the model can be easily extended to use a generalized β-divergence as a fidelity measure.

Acknowledgements

This research was funded by ANPCyT under projects PICT 2014-2627 and PICT 2015-0977, by UNL under projects CAI+D 50420150100036LI, CAI+D 50020150100059LI, CAI+D 50020150100055LI and CAI+D 50020150100082LI.

[1] https://github.com/ehabets/RIR-Generator

References

[1] J. B. Allen and D. A. Berkley, *Image method for efficiently simulating small-room acoustics*, The Journal of the Acoustical Society of America, 65 (1979), pp. 943–950.

[2] F. Ibarrola, L. Di Persia, and R. Spies, *A bayesian approach to convolutive nonnegative matrix factorization for blind speech dereverberation*, Signal Processing, 151 (2018), pp. 89–98.

[3] ——, *Switching divergences for spectral learning in blind speech dereverberation*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, (in press, 2019).

[4] D. D. Lee and H. S. Seung, *Algorithms for non-negative matrix factorization*, in Advances in Neural Information Processing Systems, 2001, pp. 556–562.

[5] N. Murata, H. Kameoka, K. Kinoshita, S. Araki, T. Nakatani, S. Koyama, and H. Saruwatari, *Reverberation-robust underdetermined source separation with non-negative tensor double deconvolution*, in Signal Processing Conference (EUSIPCO), 2016 24th European, IEEE, 2016, pp. 1648–1652.

[6] M. N. Schmidt and R. K. Olsson, *Single-channel speech separation using sparse non-negative matrix factorization*, in Ninth International Conference on Spoken Language Processing, 2006.

[7] V. Zue, S. Seneff, and J. Glass, *Speech database development at MIT: TIMIT and beyond*, Speech Communication, 9 (1990), pp. 351–356.

Doctorado en Ingeniería
Mención en Inteligencia Computacional, Señales y Sistemas

Título de la obra:

Modelos de factorización en matrices no negativas para procesamiento de audio

Autor: Francisco Javier Ibarrola

Lugar: Santa Fe, Argentina

Palabras Claves:

dereverberación, separación de fuentes sonoras,
aprendizaje maquinal, procesamiento de señales.