

Representaciones Ralas de Música Polifónica

Sparse Representations of Polyphonic Music

M. Plumbley, S. Abdallah, T. Blumensath y M. Davies
Revisor: M. Rodríguez

I. INTRODUCCIÓN

En el presente trabajo se encara el problema del análisis de música polifónica, en particular como individualizar o identificar las notas que pudieran estar presentes en una pieza musical a lo largo del tiempo. Este problema ha sido estudiado en el pasado empleando diferentes técnicas. Se han propuesto métodos de correlación para detectar notas en música monofónica como así también una variedad de métodos para el caso más complejo de música polifónica.

En el enfoque propuesto se plantea el uso de representaciones ralas como una forma natural de analizar música polifónica, donde un conjunto de átomos pertenecientes a un diccionario puede representar todas las notas presentes en la grabación. Se pretende que mediante el uso de una representación rala se logre una codificación apropiada de la pieza musical al activar un número reducido de átomos que representen las notas tocadas en determinado momento.

Con la finalidad de lograr estas representaciones los autores proponen dos alternativas, un enfoque temporal y otro espectral.

A. Enfoque en el dominio temporal

Dado el modelo generativo usual:

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{e}$$

$$x_i = \sum_k a_{ik}s_k + e_i$$

donde \mathbf{x} es un bloque de tamaño I de la señal temporal, \mathbf{A} es una matriz de tamaño $I \times K$ que contiene el conjunto de átomos que representan las características de la señal, \mathbf{s} es un vector de tamaño $K \times 1$ con los coeficientes de la representación, y \mathbf{e} es un ruido aditivo de tipo gaussiano.

Se propone una generalización del mismo de forma que se incluyan todos los defasajes posibles de los átomos en el diccionario, lo cual resulta equivalente a suponer que las señales \mathbf{x} se generan mediante un modelo convolutivo:

$$\mathbf{x} = \sum_k \mathbf{a}_k * \mathbf{s}_k + \mathbf{e}$$

Aquí \mathbf{x} es el resultado de haber convolucionado un conjunto de átomos $\{\mathbf{a}_k\}$ con un conjunto de vectores de coeficientes $\{\mathbf{s}_k\}$ y luego haber realizado la superposición de todas estas contribuciones. Inicialmente el diccionario a concebir es un arreglo de 3 dimensiones que contiene todos los átomos con sus posibles defasajes, sin embargo reestructurando este arreglo se puede obtener un diccionario que puede expresarse como una matriz rectangular para la cual se pueden aplicar técnicas convencionales de estimación de los coeficientes. La principal limitante que surge es el tamaño del espacio de búsqueda sobre el que se

deben estimar los coeficientes \mathbf{s} . Para salvar este problema se propone un método heurístico que a través de correlaciones entre los átomos y la pieza musical logra reducir dicho espacio a un subconjunto \mathcal{S} . Esto consecuentemente conduce a soluciones subóptimas.

B. Enfoque en el dominio Espectral

Este enfoque se basa en la suposición de que los espectros de potencia generados por las notas son aproximadamente aditivos suponiendo relaciones de fase aleatorias. La señal \mathbf{x} en este caso se toma como el espectro de potencia de corta duración. Dado que el espectro de potencia toma valores no negativos los autores consideran que es inapropiado incluir un ruido gaussiano auditivo en el modelo, en cambio proponen como novedad introducir un ruido multiplicativo para el cual realizan una estimación bayesiana de la varianza.

En los experimentos los diccionarios se entrenaron con una grabación de una pieza musical en la que intervino sólo un piano. Emplearon un piano con controles MIDI para contar con una versión MIDI exacta de la pieza musical. Ambos métodos condujeron a representaciones ralas que reflejan características similares a las de la versión MIDI. Las representaciones obtenidas no cambian de morfología al introducir desplazamientos en la señal de entrada y además son más eficientes ya que deja de ser necesario codificar la fase en términos de todos los elementos del diccionario. La solución en el dominio temporal resulta ser más costosa computacionalmente pero logra mayor rareza en las activaciones de los átomos, semejantes a las de los patrones de activación neuronal, por otro lado es posible la posterior reconstrucción. El método espectral es más rápido y representa mejor los armónicos de alta frecuencia, pero descarta la información de fase. Los resultados obtenidos muestran que estos métodos pueden proveer el enfoque adecuado para la transcripción o codificación de audio musical.

II. ANÁLISIS CRÍTICO GENERAL

A. Estructura general del artículo

La estructura general del artículo se encuentra bien organizada y brinda una lectura fluida y entendible en su mayor parte, salvo la sección de métodos en cual se comenten errores que opacan la claridad de las explicaciones. A continuación se comentan algunas de las secciones.

A.1. Resumen: Cita el objetivo del trabajo realizado, los métodos empleados, resultados y conclusiones. Puede entenderse por sí sólo y está en concordancia con el resto del texto.

A.2. *Introducción:* Se plantea brevemente el problema en estudio y se dan las motivaciones que conducen a los métodos propuestos. Se mencionan algunas técnicas espectrales alternativas que tienen la misma finalidad. No se menciona si dichas técnicas proveen una solución adecuada o razones por las cuales las mismas pudieran ser no satisfactorias.

A.3. *Métodos:* Si bien no entorpecen la lectura, los desarrollos matemáticos son poco claros y limitados. Cometen un error de signo en la expresión (4) de esa sección que no es acarreado en los desarrollos posteriores. Fallan en explicar con claridad las modificaciones introducidas en el modelo temporal. No justifican el uso de un modelo invariante. Cuando explican como está constituido el diccionario, no queda en claro su estructura. Cometen errores al hacer un cambio de índices ($jk \rightarrow p$) y al dar el tamaño final de tensor $Mx(JK)$ que en realidad tiene un tamaño $Ix(JK)$. A pesar de que los desarrollos abarcan muchos de los elementos necesarios al explicar sus métodos, aparecen en general muy confusos dificultando la posibilidad de que un tercero pudiera repetir con éxito la misma metodología.

A.4. *Experimentos-Resultados:* Los procedimientos empleados están claramente explicados. Se analizan los resultados haciendo comparaciones entre los métodos propuestos y un tercer método que por su similitud es comparable a los propuestos. Emplean gráficas para hacer un análisis visual e introducen una medida específica que les permite hacer comparaciones cuantitativas de entre las técnicas espectrales. Los resultados obtenidos son los que razonablemente se esperaría para los métodos propuestos. Se mencionan también algunas limitaciones de los mismos.

A.5. *Conclusiones:* Esta sección muestra ser concisa y adecuada a los fines de reflejar los puntos centrales de su trabajo. De acuerdo a los resultados, establecen que han logrado obtener representaciones ralas eficientes que sobrepasan el desempeño de otro método del estado del arte. Se reconocen desventajas de sus métodos tales como la pérdida de armónicos en alta frecuencia para el caso temporal así también como las limitaciones computacionales que esta técnica acarrea. Se proponen trabajos futuros como la combinación de ambas técnicas para brindar un nuevo enfoque en la transcripción musical.

B. Análisis del estado del arte

Mediante una extensión del modelo generativo usual pretenden lograr representaciones invariantes ante traslaciones en la señal temporal. Sin embargo este mismo planteo ya ha sido abordado por Lewicki en el contexto general de señales variables en el tiempo [1,2].

En la búsqueda del diccionario, acotan el problema de optimización MAP a un subdominio del espacio de soluciones. Para ello emplean correlaciones entre la señal de audio y los elementos del diccionario, descartando aquellos átomos que en característica o fase no expresen bien el contenido de la señal. Esta alternativa también fue propuesta por Lewicki.

Frente al enfoque temporal los autores no realizan una comparación directa con otros métodos que pudieran existir. Por otro lado, para el enfoque espectral se hacen comparaciones con técnicas relacionadas del estado del arte como lo son *Análisis de Componentes Independientes* (ICA), *Análisis de Subespacios Independientes* (ISA), *Factorización Matricial No-Negativa* (NMF), y *Representaciones Ralas*

No-Negativas (NNSC). Las diferencias que se marcan entre estos métodos y el propuesto es la inclusión del ruido multiplicativo.

El método propuesto en el dominio espectral que introduce ruido multiplicativo, es el mismo que los autores publican en [3], éste ya había sido publicado en una conferencia internacional antes del envío a corrección de la presente publicación. Por lo mencionado anteriormente este trabajo se trata más bien de una extensión de las ideas de Lewicki aplicadas al análisis de música polifónica en el dominio temporal, contrastadas frente al método propio en el dominio espectral.

C. Bibliografía presentada

La bibliografía abarca un número interesante de publicaciones, muchas de ellas de los últimos 5 años previos a la publicación del artículo. Las referencias más importantes centrales a las temáticas del artículo son [3-9], sin embargo se han omitido las referencias de Lewicki [1,2] que tratan el problema con el mismo enfoque temporal.

III. DESARROLLOS EN EL DOMINIO TEMPORAL

En esta sección se darán en detalle los desarrollos matemáticos para el método en el dominio temporal.

A. Representaciones ralas invariantes ante traslaciones

El modelo de mezcla empleado habitualmente es:

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{e} \quad (1)$$

donde \mathbf{x} es un bloque de tamaño I de la señal temporal, \mathbf{A} es una matriz $I \times K$ que contiene el conjunto de átomos que representan las características de la señal, \mathbf{s} es un vector de tamaño $K \times 1$ con los coeficientes de la representación, y \mathbf{e} es un ruido gaussiano aditivo.

Al tomar bloques de la señal se tiene la limitación de que las componentes características de la señal están alineadas arbitrariamente a los bloques. Esto resulta inconveniente al trabajar con señales que presentan variaciones temporales. En primer lugar la información de fase que contienen aquellas estructuras de la señal deben ser codificadas en términos de todo el conjunto de átomos. Esto no conduce a una representación eficiente en la cual se activen sólo algunos elementos. Este problema puede superarse al usar algún ventaneado pero sigue siendo deseable una representación que sea invariante ante traslaciones [2]. Por otro lado la representación no es única ya que cambia su morfología cuando se hacen defasajes en la señal.

El objetivo es entonces desarrollar un modelo que represente eficientemente el contenido de información temporal de la pieza musical. Esto se logra mediante un modelo generativo invariante ante traslaciones de tipo convolutivo en el cual un conjunto de átomos puede desplazarse a lo largo del tiempo.

Seleccionando I muestras consecutivas de una señal $x[t]$, podemos generar un vector \mathbf{x} cuyos elementos esten dados por $x_i = x[t + i - 1]$ con $1 \leq i \leq I$. Este vector se puede aproximar mediante versiones escaladas por coeficientes \mathbf{s} y desplazadas de un conjunto de K átomos $\{\mathbf{a}_k\}$ es decir:

$$\mathbf{x} = \sum_k \mathbf{a}_k * \mathbf{s}_k + \mathbf{e} \quad (2)$$

En particular, cada muestra i de la señal puede ser expresada alternativamente como:

$$x_i = \sum_{j=1}^J \sum_{k=1}^K a_{ijk} s_{jk} + e_i$$

donde

$$a_{ijk} = \begin{cases} a_{lk} & \text{con } l = L + i - j \text{ si } 1 \leq l \leq L; \\ 0 & \text{de otra forma.} \end{cases}$$

siendo L el soporte de los átomos, j el defasaje relativo de los mismos, mientras que J corresponde a la totalidad de defasajes considerados. En esta notación los elementos a_{ijk} están invertidos al avanzar sobre j y el producto $a_{ijk} s_{jk}$ se trata como un producto elemento a elemento, por lo que se recupera el proceso convolutivo de la expresión (2).

El diccionario $\mathbf{A}^\dagger = [a_{ijk}]$ obtenido es un tensor de orden 3 en el sentido matemático de un arreglo multidimensional. Básicamente está compuesto por un único diccionario como en (1) junto con J desplazamientos del mismo sobre la dimensión j . Por este motivo se puede reestructurar la manera en que se escribe al combinar el par de índices jk en un solo índice p . Se obtiene entonces $a_{ijk} \rightarrow b_{ip}$ y $s_{jk} \rightarrow s_p$, con lo cual podemos escribir nuevamente x_i mediante el modelo generativo lineal:

$$x_i = \sum b_{ip} s_p + e_i \quad (3)$$

Gracias a estas modificaciones pueden aplicarse los métodos ya conocidos para los problemas de aprendizaje e inferencia aplicados a representaciones ralas.

B. Inferencia de los coeficientes

Tradicionalmente los coeficientes \mathbf{s} se estiman mediante una maximización a posteriori (MAP).

$$\hat{\mathbf{s}} = \arg \max_{\mathbf{s}} [\log p(\mathbf{s}|\mathbf{A}, \mathbf{x})] \quad (4)$$

En ella se seleccionan aquellos valores \hat{s}_p que mejor explican los datos, es decir que logren reducir el término de error \mathbf{e} en (3) a un mínimo. La función logaritmo se ha introducido sólo con fines prácticos.

La probabilidad $p(\mathbf{s}|\mathbf{A}, \mathbf{x})$ puede desglosarse mediante el teorema de Bayes en:

$$p(\mathbf{s}|\mathbf{A}, \mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{A}, \mathbf{s})p(\mathbf{s})}{p(\mathbf{A}|\mathbf{x})} \quad (5)$$

$p(\mathbf{s})$ es la probabilidad *a priori* de los coeficientes, en el caso de representaciones ralas estos se consideran independientes y con distribución laplaciana:

$$p(\mathbf{s}) = p(s_1, s_2, \dots, s_K) = \prod_{k=1}^K p(s_k)$$

con

$$p(s_k) \propto \exp(-|s_k|^r) \\ r \rightarrow 0$$

luego,

$$p(\mathbf{s}|\mathbf{A}, \mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{A}, \mathbf{s}) \prod_{k=1}^K p(s_k)}{p(\mathbf{A}|\mathbf{x})} \quad (6)$$

El término $p(\mathbf{x}|\mathbf{A}, \mathbf{s})$ se obtiene en función del error \mathbf{e} como se deduce en los siguientes pasos.

La expresión para ruido aditivo gaussiano está dada por [10]:

$$p(\mathbf{e}|\mu, \Sigma) = \frac{1}{(2\pi)^{I/2} |\Sigma|^{1/2}} \times \\ \exp \left\{ -\frac{1}{2} (\mathbf{e} - \mu)^T \Sigma^{-1} (\mathbf{e} - \mu) \right\} \quad (7)$$

siendo μ un vector de medias y Σ la matriz de covarianza $\langle \mathbf{e}\mathbf{e}^T \rangle$. Escribiendo (7) en términos de la inversa de la covarianza $\Lambda = \Sigma^{-1}$ obtenemos:

$$p(\mathbf{e}|\mu, \Sigma) = \sqrt{\frac{|\Lambda|}{(2\pi)^I}} \exp \left\{ -\frac{1}{2} (\mathbf{e} - \mu)^T \Lambda (\mathbf{e} - \mu) \right\}$$

Suponiendo media cero y ruido gaussiano esférico $\Lambda = \frac{1}{\sigma^2} \mathbf{I}$:

$$p(\mathbf{e}) = \frac{1}{(2\pi)^{I/2} \sigma} \exp \left\{ -\frac{1}{2\sigma^2} \|\mathbf{e}\|_2^2 \right\}$$

Dado que $\mathbf{e} = \mathbf{x} - \mathbf{A}\mathbf{s}$ podemos escribir la probabilidad condicional para \mathbf{x} como:

$$p(\mathbf{x}|\mathbf{A}, \mathbf{s}) = \frac{1}{(2\pi)^{I/2} \sigma} \exp \left\{ -\frac{1}{2\sigma^2} \|\mathbf{e}\|_2^2 \right\} \quad (8)$$

Finalmente llevando (8) a (6), y aplicando logaritmos obtenemos la expresión para la maximización a posteriori:

$$\hat{\mathbf{s}} = \arg \max_{\mathbf{s}} \left[-\frac{1}{2\sigma^2} \|\mathbf{e}\|_2^2 + \sum_k p(s_k) \right] \quad (9)$$

La implementación MAP se realiza a través del algoritmo de *maximización de la esperanza* (EM) [11]. Sin embargo dado el tamaño del diccionario $\{a_{ijk}\}$ es necesario que previamente se realice una reducción del espacio de búsqueda, seleccionando un subconjunto de los coeficientes \mathbf{s} cuyos átomos correspondientes tengan alta correlación con la señal \mathbf{x} . Adicionalmente se excluyen aquellos elementos del diccionario que se encuentran en una vecindad del átomo que mayor correlación mostró. Aún cuando este algoritmo se trate de una solución subóptima, brinda un buen desempeño en la práctica.

IV. DESARROLLOS EN EL DOMINIO ESPECTRAL

A. Modelo Generativo

Alternativamente al método temporal, se propone la búsqueda de representaciones ralas desde el dominio espectral. Este método está motivado por la noción de que los espectros de potencia para las distintas notas son aproximadamente aditivos. Como las magnitudes que se manejan son no-negativas se considera inapropiado incluir un ruido gaussiano aditivo, se propone en cambio un ruido de tipo multiplicativo.

Dado un bloque de la señal temporal, \mathbf{x} es el espectro de potencia de corta duración de la misma, tomado como las magnitudes de las partes real e imaginaria al cuadrado:

$$\mathbf{x} = \mathcal{R}e[\mathcal{F}(x(t))]^2 + \mathcal{I}m[\mathcal{F}(x(t))]^2 = \mathbf{x}_{\mathcal{R}e}^2 + \mathbf{x}_{\mathcal{I}m}^2$$

Las muestras i de las partes real e imaginaria se consideran pertenecientes a una distribución normal con media cero y varianza v_i desconocida:

$$\{x_{\mathcal{R}e(i)}\} \sim \mathcal{N}(0, v_i) \quad \{x_{\mathcal{I}m(i)}\} \sim \mathcal{N}(0, v_i)$$

Debido a que las muestras del espectro son sumas de cuadrado de variables aleatorias con distribución normal, éstas tienen una distribución gama o chi-cuadrado escalada:

$$x_i \sim \Gamma(v_i) = \frac{1}{2}v_i\chi_2^2$$

En general una distribución chi-cuadrado escalada tiene la forma:

$$\chi = \frac{1}{d} \sum_j Z_j^2 \sim \Gamma\left(\frac{d}{2}, \frac{2}{d}v\right) = \frac{1}{d}v\chi_2^d$$

con $Z_j \sim \mathcal{N}(0, v_i)$, siendo en nuestro caso $d = 2$ por las partes real e imaginaria.

En este contexto para lograr una representación rala del espectro, se considera que las varianzas son generadas a partir de una mezcla lineal de coeficientes s_k :

$$\mathbf{v} = \mathbf{A}\mathbf{s}$$

\mathbf{A} es un diccionario sobrecompleto cuyas columnas son los átomos del espectro. Las varianzas generadas actúan como ruidos multiplicativos que dan lugar a las muestras del espectro:

$$x_i \sim v_i\Gamma\left(\frac{d}{2}, \frac{2}{d}\right)$$

B. Inferencia

En el proceso de inferencia se desea hallar una estimación MAP de \mathbf{s} que maximice la probabilidad de los datos \mathbf{x} :

$$\begin{aligned} \hat{\mathbf{s}} &= \arg \max_{\mathbf{s}} [\log p(\mathbf{s}|\mathbf{A}, \mathbf{x})] \\ &= \arg \max_{\mathbf{s}} [\log p(\mathbf{x}|\mathbf{A}, \mathbf{s}) + \log p(\mathbf{s}) + \text{cte}] \\ &= \arg \max_{\mathbf{s}} [\log p(\mathbf{x}|\mathbf{v}) + \log p(\mathbf{s})] \\ &= \arg \max_{\mathbf{s}} f(\mathbf{s}) \end{aligned}$$

Suponiendo independiencia para las muestras x_i del espectro, y para los coeficientes s_k :

$$\begin{aligned} \log p(\mathbf{x}|\mathbf{v}) &= \sum_i \log p(x_i|v_i) \\ \log p(\mathbf{s}) &= \sum_k \log p(s_k) \end{aligned}$$

La distribución a priori adoptada es una gaussiana generalizada $p(s_k) = \frac{1}{2} \exp(-\frac{1}{\alpha}|s_k|^\alpha)$, con $\alpha = 0, 2$.

Para una varianza v_i , las realizaciones x_i tienen probabilidad:

$$p(x_i|v_i) = \frac{1}{x_i\Gamma(d/2)} \left(\frac{d}{2} \frac{x_i}{v_i}\right)^{d/2} \exp\left(-\frac{d}{2} \frac{x_i}{v_i}\right)$$

tomando logaritmos:

$$\log p(x_i|v_i) = \frac{d}{2} \left(\log \frac{d}{2} + \log \frac{x_i}{v_i} \right) - \frac{d}{2} \frac{x_i}{v_i} - \log [x_i\Gamma(d/2)]$$

El argumento $f(\mathbf{s})$ puede reescribirse haciendo uso de la notación de Einstein como:

$$\begin{aligned} f(\mathbf{s}) &= \sum_i \log p(x_i|[\mathbf{A}\mathbf{s}]_i) + \sum_k \log p(s_k) \\ &= \sum_i \left(\frac{d}{2} \log \frac{x_i}{a_{ik}s_k} - \frac{d}{2} \frac{x_i}{a_{ik}s_k} \right) + \sum_k \log p(s_k) + \text{cte} \end{aligned}$$

Para hallar el óptimo $\hat{\mathbf{s}}$ buscamos un gradiente ascendente para $f(\mathbf{s})$:

$$\begin{aligned} \frac{df(\mathbf{s})}{ds_k} &= \sum_i \left(\frac{d}{2} \frac{x_i}{v_i s_k} - \frac{d}{2s_k} \right) - \phi(s_k) \\ &= \frac{d}{2} \sum_i \frac{a_{ik}}{v_i} \left(\frac{x_i}{v_i} - 1 \right) - \phi(s_k) \end{aligned}$$

siendo $\phi(s_k) = -(d/ds_k) \log p(s_k) \geq 0$.

Debido a problemas de convergencia con las reglas de actualización de tipo aditivas, en este trabajo se propone emplear una regla multiplicativa como las empleadas en NMF [12] para la actualización de los coeficientes. Ya que el gradiente puede expresarse como $(d/ds_k)f(\mathbf{s}) = (A - B)$ siendo tanto A como B no negativos, resulta posible tal tipo de actualización. En la siguiente tabla se muestran las equivalencias entre los dos tipos de actualizaciones, η es un parámetro de aprendizaje.

TABLA I
EQUIVALENCIAS ENTRE LAS REGLAS ADITIVA Y MULTIPLICATIVA.

Regla aditiva	Regla multiplicativa
$s^{k+1} = s^k + \eta(A - B)$	$s^{k+1} = s^k \left(\frac{A}{B}\right)^\eta$
si $(A - B) > 0 \rightarrow s^{k+1} > s^k$	$\left(\frac{A}{B}\right) > 1 \rightarrow s^{k+1} > s^k$
si $(A - B) < 0 \rightarrow s^{k+1} < s^k$	$\left(\frac{A}{B}\right) < 1 \rightarrow s^{k+1} < s^k$
si $(A - B) = 0 \rightarrow s^{k+1} = s^k$	$\left(\frac{A}{B}\right) = 1 \rightarrow s^{k+1} = s^k$

Sin ahondar en cuestiones relacionadas a la convergencia y estabilidad de las reglas, se aprecia que la regla multiplicativa es una posible alternativa para el ajuste de los coeficientes.

De lo anterior la actualización de los coeficientes s_k en el problema de inferencia se calcula como:

$$\begin{aligned} A &= \frac{d}{2} \sum_i (a_{ik}/v_i)(x_i)(v_i) \\ B &= \frac{d}{2} \sum_i (a_{ik})/(v_i) + \phi(s_k) \\ s^{k+1} &= s^k \frac{\sum_i (a_{ik}/v_i)(x_i)(v_i)}{(2/d)\phi(s_k) + \sum_i (a_{ik}/v_i)} \end{aligned}$$

V. ANÁLISIS CRÍTICO DE RESULTADOS

A. Métodos experimentales

En base a los resultados experimentales para los métodos en el dominio temporal (SISC), espectral (NNSC), y NMF,

los autores realizan una evaluación de los mismos al considerar aspectos tales como los átomos aprendidos por cada método, la rareza de las representaciones logradas, cómo realizan la detección de las notas, así también como las limitaciones prácticas de los mismos. Los métodos SISC y NNSC muestran muchas de las características deseables, siendo adecuados para la transcripción o codificación de audio musical.

Por consideraciones prácticas se empleó en los experimentos una pieza musical en la que sólo intervino un piano, acotando de esta forma la cantidad de átomos necesarios para representar las distintas notas musicales. Es preciso mencionar que el entrenamiento se realizó empleando una grabación de audio que se obtuvo a partir de la versión MIDI de la pieza musical, la cual no es una grabación real de un pianista. Por otro lado se usó una única representación en el aprendizaje, esto puede considerarse aceptable a los fines de obtener resultados preliminares ya que se está en ausencia de ruido y hay un conjunto reducido de notas en un piano.

B. Cómo se presentan y discuten las tablas y figuras

Los resultados de las experiencias se analizaron en su mayor parte mediante el apoyo de gráficas que facilitaron la comparación de los tres métodos, no se hizo uso de tablas que resuman datos cuantitativos sino que donde fuera necesario se los incluyó en el texto. Cada una de las gráficas ayuda a la comprensión de las discusiones dadas en el texto, y contienen un epígrafe que es entendible en su contexto. Hubiese sido conveniente señalar sobre la figura 6 del artículo algunos de los puntos que se mencionan en el texto.

C. Trabajos futuros propuestos

Se proponen como trabajos a futuro el realizar transcripciones de grabaciones ejecutadas por un pianista real e incluir nuevos instrumentos musicales tales como instrumentos de cuerda, lo cual supone aprender un modelo generativo más complejo. También consideran la posibilidad de combinar lo mejor de ambas técnicas en un único método, y la aplicación de la técnica temporal como un método de reconstrucción directa.

REFERENCIAS

- [1] M. Lewicki y T. Sejnowski, "Coding time-varying signals using sparse, shift-invariant representations," en *Advances Neural Inf. Process. Syst.*, vol. 11, pp. 730–736, 1999.
- [2] R. Rao, B. Olshausen, y M. Lewicki, *Probabilistic Models of the Brain: Perception and Neural Function*. MIT Press, 2002.
- [3] M. Plumbley, S. Abdallah, T. Blumensath, y M. Davies, "Sparse representations of polyphonic music," en *Proceedings of the Fifth International Conference on Music Information Retrieval, Barcelona, Spain*, pp. 318–325, 2004.
- [4] K. Kreutz-Delgado, J. Murray, B. Rao, y T. S. K. Engan, T. Lee, "Dictionary learning algorithms for sparse representations," en *Neural Comput.*, vol. 15, pp. 349–396, 2003.
- [5] T. Blumensath y M. Davies, "Unsupervised learning of sparse and shift-invariant decompositions of polyphonic music," en *Proceedings of the IEEE International Conference Acoustics, Speech, and Signal Process.*, vol. 5, pp. 497–500, 2004.
- [6] —, "On shift-invariant sparse coding," en *Independent Component Analysis and Blind Signal Separation: Proceedings of the Fifth International Conference, ICA 2004, Lectures Notes in Computer Science*, vol. 3195, pp. 1205–1212, 2004.
- [7] M. Plumbley, S. Abdallah, J. Bello, M. Davies, G. Monti, y M. Sandler, "Automatic music transcription and audio source separation," en *Cybernetics and Systems*, vol. 6, pp. 603–627, 2002.

- [8] P. Smaragdis y J. Brown, "Non-negative matrix factorization for polyphonic music transcription," en *Proceedings of the 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York*, pp. 177–180, 2003.
- [9] P. Hoyer, "Non-negative sparse coding," en *Neural Networks for Signal Processing XII (Proceedings of the IEEE Workshop on Neural Networks for Signal Processing)*, pp. 557–565, 2002.
- [10] C. Bishop, *Pattern Recognition and Machine Learning*, 2006.
- [11] T. Blumensath, *PhD Thesis, Bayesian Modelling of Music: Algorithmic Advances and Experimental Studies of Shift-Invariant Sparse Coding*, 2006.
- [12] D. Lee y H. Seung, "Algorithms for non-negative matrix factorization," en *Advances in Neural Information Processing Systems. MIT Press*, vol. 13, pp. 556–562, 2001.