Journal of Signal and Information Processing, **2013, 4, 186-196**
http://dx.doi.org/10.4236/jsip.2013.42026 Published Online May 2013 (http://www.scirp.org/journal/jsip)

Single Channel Source Separation Using Filterbank and 2D Sparse Matrix Factorization

Xiangying Lu[1], Bin Gao[1], Li Chin Khor[1], Wai Lok Woo[1], Satnam Dlay[1], Wingkuen Ling[2], Cheng S. Chin[3]

[1]School of Electrical and Electronic Engineering, Newcastle University, England, UK; [2]Faculty of Information Engineering, Guangdong University of Technology, Guangzhou, China; [3]School of Marine Science and Technology, Newcastle University, England, UK.
Email: bin.gao@ncl.ac.uk, w.l.woo@ncl.ac.uk

Received December 20th, 2012; revised January 23rd, 2013; accepted January 31st, 2013

ABSTRACT

We present a novel approach to solve the problem of single channel source separation (SCSS) based on filterbank technique and sparse non-negative matrix two dimensional deconvolution (SNMF2D). The proposed approach does not require training information of the sources and therefore, it is highly suited for practicality of SCSS. The major problem of most existing SCSS algorithms lies in their inability to resolve the mixing ambiguity in the single channel observation. Our proposed approach tackles this difficult problem by using filterbank which decomposes the mixed signal into sub-band domain. This will result the mixture in sub-band domain to be more separable. By incorporating SNMF2D algorithm, the spectral-temporal structure of the sources can be obtained more accurately. Real time test has been conducted and it is shown that the proposed method gives high quality source separation performance.

Keywords: Blind Source Separation; Non-Negative Matrix Factorization; Filterbank Analysis

1. Introduction

Blind source separation has gained a great deal of attention in signal processing applications and these consist of medical signal analysis, telecommunications and speech recognition. There is an essential topic known as single channel source separation (SCSS) [1,2] which has not yet been enhanced enough to make its way out of laboratories. In recent years, new advances have been achieved in SCSS and this can be categorized into two main branches: supervised SCSS methods (e.g. model-based SCSS [3-6] techniques) and unsupervised SCSS methods (e.g. non-negative matrix factorization (NMF) [7] and computational auditory scene analysis (CASA) [8]). In this paper, we proposed a novel unsupervised SCSS method based on non-negative matrix factorization approach.

NMF methods have been widely exploited in the field of SCSS, and especially being used in separating audio mixtures, e.g. extracting drums from polyphonic music [9] and automatic transcription of polyphonic music [10]. Families of parameterized NMF cost functions such as the Beta divergence [11] and Csiszar's divergences [12] have been presented for the separation of audio signals [13] and in general case, the least square distance [14] and the Kullback-Leibler (KL) divergence [15] are two main cost functions which have been widely employed in NMF. However, the problem where conventional NMF methods fail in SCSS is when two notes are played simultaneously in which case they will be modeled as one component [7]. To overcome this limitation, the sparse non-negative matrix two dimensional deconvolution (SNMF2D) [16] is derived to track the spectral frequencies of the sources that change over time.

This paper presents a novel method based on SNMF2D and filterbank technique. The audio mixture is generated by sources composed of two different origins audio signals but received synchronously by one microphone. Intuitively, the proposed separation strategy utilizes filterbank to make the observed mixed signal analyzed in sub-band domain. The impetus behind this is that the degree of mixing of the sources in the sub-band domain is now less ambiguous and thus, the dominating source in the sub-band mixture can be easily detected. Therefore, the spectral and temporal patterns (*i.e*. the spectral bases and temporal codes, respectively) associated in each sub-band can be extracted more accurately by using

SNMF2D. Once the sub-sources are obtained the *k*-means based clustering method is used to group these sub-sources into clusters where each cluster consist a set of sub-sources [17,18] which will be subsequently used for recovering the original sources. The proposed method concentrates on the idea of performance source separation in the sub-band domain and avoids directly estimateing the sources using the mixture signal which contains too many mixing ambiguities between sources. In this way, we show that the proposed method can make a superior separation performance.

The paper is organized as follows: In Section 2, the proposed source separation framework is fully developed. In Section 3, experimental results are presented. The impact factors and a series of performance comparison are discussed in Section 4. Finally, Section 5 concludes the paper.

2. Proposed Model

The single channel audio mixture $x(t)$ is given as:

$$x(t) = s_1(t) + s_2(t) \tag{1}$$

where $t = 1, 2, \cdots, T$ denotes time index. The goal of SCSS is to estimate the two sources $s_1(t)$ and $s_2(t)$ when only the observation signal $x(t)$ is available.

The core procedure of the proposed method is shown in **Figure 1**. It consists of two main techniques-filterbank and SNMF2D. The benefits filterbank bring to SCSS are 1) the degree of mixing ambiguity from the original sources is reduced in that particular sub-band signal; and 2) the complexity of the spectral and temporal patterns associated with each sub-band will be simpler and sparser as compared with that of the mixed signal. The specific steps of the proposed method are summarized as follows:

Step 1: Transform the mixture from time domain into sub-band domain using filterbank, and then down-sampling the signal for reducing the aliasing problem. Hence, instead of processing the mixed signal directly, the sub-band signals are utilized as the new set of observations.

Step 2: Convert the sub-band mixed signals into time-frequency (TF) domain by using STFT (Short-Time Fourier Transform) and then construct log-frequency magnitude spectrogram, utilize SNMF2D to decompose the sub-band mixing TF mixtures into source related spectral and temporal patterns. The separated time domain sub-sources can be reconstructed by using the inverse STFT.

Step 3: Use the *k*-means clustering method to group the sub-sources into different clusters where each cluster consist a set of sub-sources correspond to one recovered source.

Step 4: Recover the time domain sources in the synthesis stage.

1) *Pre-processing stage*: Filterbank includes low-pass, band-pass, and high-pass filters which are served to isolate different frequency components in a signal. The perfect filterbank will be designed so that the output source is the same as the input source with no distortion through a time shift and amplitude scaling. Here, the down-sampling is served to reduce the aliasing [19-21] problem. In the sub-band analysis, the formulation of filterbank is given as follow:

$$h_0(n) = 4F_c \sin c(2F_c\gamma)\, w(n), \frac{-(N-1)}{2} \le \gamma \le \frac{(N-1)}{2} \tag{2}$$

$$h_k(n) = h_0(n)\cos\left[\left(k-\frac{1}{2}\right)\frac{n\pi}{K}\right], k = 1, 2, \cdots, K \tag{3}$$

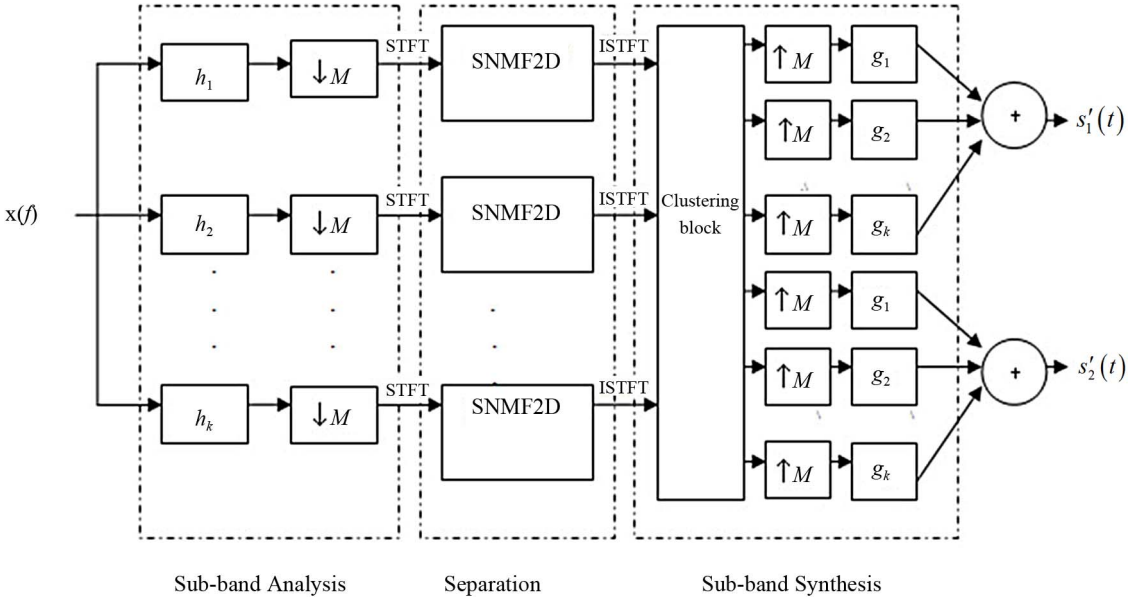


Figure 1. Core procedure of the proposed method.

where the finite number of sub-bands is K, N is the length of window, the cut-off frequency is defined as $F_c = \frac{1}{4K}$ and $w(n)$ is the Hamming window given by

$$w(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{N}\right), n = 0,1,\cdots,N \quad (4)$$

In this paper, the observations after filterbank processing can be effectively down-sampled by an integer decimation factor D_r (down-sampling rate) in each sub-band. The down-sampled observation $x(k,p)$ in the k^{th} $(k = 1,2,\cdots,K)$ sub-band is generated by using Equation (5), where $p = lD_r$ denotes the time index at the reduced sampling rate for some integer l, and $D < K$ for avoiding [21] any aliasing distortion.

$$x(k,p) = \sum_{n=1}^{N} h_k(n)x(p-n) \quad (5)$$

2) *Separation stage*: Once the new set of observations $x(k,p)$ has been generated, the sub-band mixed signals are transformed to the time-frequency (TF) domain using STFT. We then group the spectrogram bins into 175 logarithmically spaced frequency bins in the range of 50 Hz to 8 kHz with 24 bins per octave, which corresponds to twice the resolution of the equal tempered musical scale to construct log-frequency [7] magnitude spectrogram. Within the context of SCSS, the TF representation of the mixture in (1) is given by

$$|\boldsymbol{X}|^2 \approx |\boldsymbol{S}_1|^2 + |\boldsymbol{S}_2|^2 \quad (6)$$

where $\boldsymbol{X} = \left[X(i,j)\right]_{j=1,2,\cdots,T_s}^{i=1,2,\cdots,F}$ and $\boldsymbol{S}_d = \left[S_d(i,j)\right]_{j=1,2,\cdots,T_s}^{i=1,2,\cdots,F}$ are two-dimensional matrices (row and column vector represents the time slots and frequencies, respectively). In this paper, we term the sources at each sub-band as the sub-sources. To estimate these sub-sources, we project all the sub-band mixed signals from (5) into the TF domain, in which can be denoted as:

$$\left|\boldsymbol{C}_k^x\right|^2 \approx \left|\boldsymbol{C}_k^{s_1}\right|^2 + \left|\boldsymbol{C}_k^{s_2}\right|^2 \quad (7)$$

$\left|\boldsymbol{C}_k^x\right|^2$ and $\left|\boldsymbol{C}_k^{s_d}\right|^2$ denotes TF representations of the k^{th} sub-band mixture $x(k,p)$ and d^{th} source, respectively. The matrices we are interested to determine are $\left|\boldsymbol{C}_k^{s_1}\right|^2$ and $\left|\boldsymbol{C}_k^{s_2}\right|^2$ which can be estimated using any non-negative matrix factorization algorithm. In our approach, we favours the SNMF2D algorithm where the desirable matrices can be estimated as

$$\left|\tilde{\boldsymbol{C}}_k^{s_1}\right|^2 = \sum_{\tau}\sum_{\phi} \boldsymbol{W}_{k,1}^{\tau} \overset{\downarrow\phi}{\boldsymbol{H}}_{k,1}^{\phi} \overset{\rightarrow\tau}{} \quad \text{and} \quad \left|\tilde{\boldsymbol{C}}_k^{s_2}\right|^2 = \sum_{\tau}\sum_{\phi} \boldsymbol{W}_{k,2}^{\tau} \overset{\downarrow\phi}{\boldsymbol{H}}_{k,2}^{\phi} \overset{\rightarrow\tau}{}.$$

Here, $\boldsymbol{W}_{k,d}^{\tau}$ denotes the d^{th} column of $\boldsymbol{W}_k^{\tau}$ that corresponds to the d^{th} row of $\boldsymbol{H}_{k,d}^{\phi}$. In the case of two sources, we have $d = \{1,2\}$. The reasons SNMF2D [19,20] are favoured over other conventional NMF methods are noted as follows: 1) The NMF does not model notes but rather unique events only. Thus, if two notes are always played simultaneously they will be modelled as one component; 2) The structure of a factor in $\boldsymbol{H}$ can be input into signature of the same factor in $\boldsymbol{W}$ and vice versa. Thus, this leads to ambiguity that can be resolved by forcing the structure on $\boldsymbol{W}$ through imposing sparseness on $\boldsymbol{H}$. The two basic cost functions for optimizing $\boldsymbol{W}_k^{\tau}$ and $\boldsymbol{H}_{k,d}^{\phi}$ are given by the Least Squares (LS) distance and Kullback-Leibler divergence (KLd) where λ is a sparseness parameter and $f(\boldsymbol{H}) = \|\boldsymbol{H}\|_1$:

$$\text{LS:}\quad C_{SLS} = \frac{1}{2}\sum_{i,j}\left(\boldsymbol{V}_{i,j} - \tilde{\boldsymbol{\Lambda}}_{i,j}\right)^2 + \lambda f(\boldsymbol{H}) \quad (8)$$

$$\text{KLd:}\quad C_{SKL} = \sum_{i,j}\boldsymbol{V}_{i,j}\log\frac{\boldsymbol{V}_{i,j}}{\tilde{\boldsymbol{\Lambda}}_{i,j}} - \boldsymbol{V}_{i,j} + \tilde{\boldsymbol{\Lambda}}_{i,j} + \lambda f(\boldsymbol{H}) \quad (9)$$

In above, $\boldsymbol{V}_{i,j}$ is the log-frequency magnitude spectrogram, $\boldsymbol{V} \in \Re_+^{I\times J}$ is the data matrix of the sub-band TF mixture and $\tilde{\boldsymbol{\Lambda}} = \sum_{\tau}\sum_{\phi} \overset{\downarrow\phi}{\tilde{\boldsymbol{W}}}^{\tau} \overset{\rightarrow\tau}{\boldsymbol{H}}^{\phi}$ where $\tilde{\boldsymbol{W}}_{i,d}^{\tau} = \tilde{\boldsymbol{W}}_{i,d}^{\tau} \Big/ \sqrt{\sum_{\tau,i}\left(\tilde{\boldsymbol{W}}_{i,d}^{\tau}\right)^2}$. The derivatives of (8) with respect to $\boldsymbol{W}^{\tau}$ and $\boldsymbol{H}^{\phi}$ are given by:

$$\begin{aligned}
&\frac{\partial C_{SLS}}{\partial \boldsymbol{W}_{i',d'}^{\tau'}} \\
&= \frac{\partial}{\partial \boldsymbol{W}_{i',d'}^{\tau'}}\left(\frac{1}{2}\sum_{i,j}\left(\boldsymbol{Y}_{i,j} - \tilde{\boldsymbol{\Lambda}}_{i,j}\right)^2 + f(\boldsymbol{H})\right) \quad (10)\\
&= -\frac{1}{2}\sum_{\phi,j}\left(\boldsymbol{Y}_{i'+\phi,j} - \tilde{\boldsymbol{\Lambda}}_{i'+\phi,j}\right)\boldsymbol{H}_{d',j-\tau'}^{\phi}
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial C_{SLS}}{\partial \boldsymbol{H}_{d',j'}^{\phi'}} \\
&= \frac{\partial}{\partial \boldsymbol{H}_{d',j'}^{\phi'}}\left(\frac{1}{2}\sum_{i,j}\left(\boldsymbol{Y}_{i,j} - \tilde{\boldsymbol{\Lambda}}_{i,j}\right)^2 + f(\boldsymbol{H})\right) \quad (11)\\
&= -\frac{1}{2}\sum_{\tau,i}\tilde{\boldsymbol{W}}_{i-\phi',d'}^{\tau}\left(\boldsymbol{Y}_{i,j'+\tau} - \tilde{\boldsymbol{\Lambda}}_{i,j'+\tau}\right) + \frac{\partial f(\boldsymbol{H})}{\partial \boldsymbol{H}_{d',j'}^{\phi'}}
\end{aligned}$$

Thus, by applying the standard gradient decent approach, we have:

$$\begin{aligned}
\boldsymbol{W}_{i',d'}^{\tau'} &\leftarrow \tilde{\boldsymbol{W}}_{i',d'}^{\tau'} - \eta_W \frac{\partial C_{SLS}}{\partial \boldsymbol{W}_{i',d'}^{\tau'}} \\
\boldsymbol{H}_{d',j'}^{\phi'} &\leftarrow \boldsymbol{H}_{d',j'}^{\phi'} - \eta_H \frac{\partial C_{SLS}}{\partial \boldsymbol{H}_{d',j'}^{\phi'}}
\end{aligned} \quad (12)$$

where $\eta_{\boldsymbol{W}}$ and $\eta_{\boldsymbol{H}}$ are positive learning rates which can be obtained by following the approach of Lee and Seung [15], namely, $\eta_{\boldsymbol{W}} = \tilde{\boldsymbol{W}}^{\tau}_{i',d'} \Big/ \left(\sum_{\phi,j} \tilde{\boldsymbol{\Lambda}}_{i'+\phi,j} \boldsymbol{H}^{\phi}_{d',j-\tau'} \right)$ and $\eta_{\boldsymbol{H}} = \boldsymbol{H}^{\phi'}_{d',j'} \Big/ \left(\sum_{\tau,i} \tilde{\boldsymbol{W}}^{\tau}_{i-\phi',d'} \tilde{\boldsymbol{\Lambda}}_{i,j'+\tau} + \frac{\partial f(\boldsymbol{H})}{\partial \boldsymbol{H}^{\phi'}_{d',j'}} \right)$. Thus, in matrix notation, by using the multiplicative learning rules, the SNMF2D algorithm are summarized in **Table 1**. In these tables, the superscript "$\mathbf{T}$" denotes vector transpose, "$\bullet$" is the element-wise product and at each iteration, $\mathrm{diag}(\cdot)$ denotes a matrix with the argument on the diagonal. The column vectors of $\boldsymbol{W}^{\tau}$ will be factor-wise normalized to unit length.

After using mask, the sub-sources $\tilde{c}_k^{s_a}$ can be obtained:

$$\begin{aligned} \tilde{c}_k^{s_1} &= STFT^{-1}\left(\mathbf{mask}_k^{s_1} \bullet \boldsymbol{C}_k^{x}\right) \\ \tilde{c}_k^{s_2} &= STFT^{-1}\left(\mathbf{mask}_k^{s_2} \bullet \boldsymbol{C}_k^{x}\right) \end{aligned} \tag{13}$$

where the masks are determined element-wise by:

$$\mathbf{mask}^{s_a}_{k,i,j} = \begin{cases} 1, \text{if } \left| \left[\tilde{\boldsymbol{C}}_k^{s_a} \right]_{i,j} \right|^2 > \left| \left[\tilde{\boldsymbol{C}}_k^{s_b} \right]_{i,j} \right|^2 \\ 0, \text{otherwise} \end{cases} \tag{14}$$

3) *Clustering stage*: Once the sub-sources are obtained by using inverse STFT, the *k*-means based clustering method is used to group these sub-sources into clusters according to the number of sources. In this case, the *k*-means method aims to separate 2 *K* observations (*K* is number of sub-bands) into two clusters (corresponding to two sources). After convergence, all sub-sources will be grouped into their respective clusters which are given denoted as $\tilde{\boldsymbol{G}}_k^{s_2} = \left\{ \tilde{\boldsymbol{c}}_1^{s_2}, \tilde{\boldsymbol{c}}_2^{s_2}, \cdots, \tilde{\boldsymbol{c}}_{N_2}^{s_2} \right\}$ and $\tilde{\boldsymbol{G}}_k^{s_2} = \left\{ \tilde{\boldsymbol{c}}_1^{s_2}, \tilde{\boldsymbol{c}}_2^{s_2}, \cdots, \tilde{\boldsymbol{c}}_{N_2}^{s_2} \right\}$ which contains N_1 and N_2 number of sub-sources that belong to Source 1 and Source 2, respectively.

4) *Synthesis stage*: After up-sampling, the filterbank synthesis process is used to recombine all the sub-sources to form the estimated source signals. A series of expansions of the output can be reconstructed by using the time-shifted variants g_k (synthesis filter) [19-21]. The process is expressed as follow:

$$\begin{aligned} g_0 &= h_0\left(N-n-1\right), n = 1,2,\cdots,N \\ g_k\left(n\right) &= g_0\left(n\right)\cos\left[\left(k-\frac{1}{2}\right)\frac{n\pi}{K}\right] \end{aligned} \tag{15}$$

Finally, the recovered sources from each cluster can be estimated as

$$\hat{s}_d\left(t\right) = \sum_{k=1}^{K}\sum_{n=1}^{N} g_k\left(n\right)\tilde{c}_k^{s_d}\left(t-n\right) \tag{16}$$

where $t\left(t = l/M\right)$ denotes the time index at the restored sampling rate $\left(M\right)$.

3. Experimental Results

The proposed monaural source separation algorithm is tested on recorded audio signals. Several experimental studies have been designed to investigate the efficacy of the proposed approach. All simulations and analysis are conducted using a PC with Intel Core 2 CPU 6600 @ 2.4 GHz and 2 GB RAM. For mixture generation, two sentences of the target speakers (male and female) "fcjf0" and "mcpm0", were selected from TIMIT speech data-

Table 1. SNMF2D (LS and KL) algorithm.

A) Initialize $\boldsymbol{W}$ and $\boldsymbol{H}$ are nonnegative randomly matrix

B) $\tilde{\boldsymbol{W}}^{\tau}_{i,d} = \tilde{\boldsymbol{W}}^{\tau}_{i,d} \Big/ \sqrt{\sum_{\tau,i} \left(\tilde{\boldsymbol{W}}^{\tau}_{i,d}\right)^2}$

C) $\tilde{\boldsymbol{\Lambda}} = \sum_{\tau}\sum_{\phi} \overset{\downarrow\phi}{\tilde{\boldsymbol{W}}}{}^{\tau} \overset{\rightarrow\tau}{\boldsymbol{H}}{}^{\phi}$

D) $\boldsymbol{H}^{\phi} \leftarrow \boldsymbol{H}^{\phi} \bullet \dfrac{\sum_{\tau} \overset{\downarrow\phi}{\tilde{\boldsymbol{W}}}{}^{\tau\,\mathbf{T}} \overset{\leftarrow\tau}{\boldsymbol{V}}}{\sum_{\tau} \overset{\downarrow\phi}{\tilde{\boldsymbol{W}}}{}^{\tau\,\mathbf{T}} \overset{\leftarrow\tau}{\tilde{\boldsymbol{\Lambda}}} + \beta \dfrac{\partial f(\boldsymbol{H})}{\partial \boldsymbol{H}^{\phi}}}$ (LS) $\boldsymbol{H}^{\phi} \leftarrow \boldsymbol{H}^{\phi} \bullet \dfrac{\sum_{\tau} \overset{\downarrow\phi}{\tilde{\boldsymbol{W}}}{}^{\tau\,\mathbf{T}} \left(\overset{\leftarrow\tau}{\boldsymbol{V}} \Big/ \overset{\leftarrow\tau}{\tilde{\boldsymbol{\Lambda}}} \right)}{\sum_{\tau} \overset{\downarrow\phi}{\tilde{\boldsymbol{W}}}{}^{\tau\,\mathbf{T}} \boldsymbol{1} + \beta \dfrac{\partial f(\boldsymbol{H})}{\partial \boldsymbol{H}^{\phi}}}$ (KL)

E) $\tilde{\boldsymbol{\Lambda}} = \sum_{\tau}\sum_{\phi} \overset{\downarrow\phi}{\tilde{\boldsymbol{W}}}{}^{\tau} \overset{\rightarrow\tau}{\boldsymbol{H}}{}^{\phi}$

F) $\boldsymbol{W}^{\tau} \leftarrow \tilde{\boldsymbol{W}}^{\tau} \bullet \dfrac{\sum_{\phi} \overset{\uparrow\phi}{\boldsymbol{V}}{}^{\tau} \overset{\rightarrow\tau}{\boldsymbol{H}}{}^{\phi\,\mathbf{T}} + \tilde{\boldsymbol{W}}^{\tau} \mathrm{diag}\left(\sum_{\tau} \boldsymbol{1} \left(\left(\overset{\uparrow\phi}{\boldsymbol{V}} \overset{\rightarrow\tau}{\boldsymbol{H}}{}^{\phi\,\mathbf{T}} \right) \bullet \tilde{\boldsymbol{W}}^{\tau} \right) \right)}{\sum_{\phi} \overset{\uparrow\phi}{\tilde{\boldsymbol{\Lambda}}} \overset{\rightarrow\tau}{\boldsymbol{H}}{}^{\phi\,\mathbf{T}} + \tilde{\boldsymbol{W}}^{\tau} \mathrm{diag}\left(\sum_{\tau} \boldsymbol{1} \left(\left(\overset{\uparrow\phi}{\tilde{\boldsymbol{\Lambda}}} \overset{\rightarrow\tau}{\boldsymbol{H}}{}^{\phi\,\mathbf{T}} \right) \bullet \tilde{\boldsymbol{W}}^{\tau} \right) \right)}$ (Least Square) $\boldsymbol{W}^{\tau} \leftarrow \tilde{\boldsymbol{W}}^{\tau} \bullet \dfrac{\sum_{\phi} \left(\overset{\uparrow\phi}{\boldsymbol{V}} \Big/ \overset{\uparrow\phi}{\tilde{\boldsymbol{\Lambda}}} \right) \overset{\rightarrow\tau}{\boldsymbol{H}}{}^{\phi\,\mathbf{T}} + \tilde{\boldsymbol{W}}^{\tau} \mathrm{diag}\left(\sum_{\tau} \boldsymbol{1} \left(\left(\boldsymbol{1} \overset{\rightarrow\tau}{\boldsymbol{H}}{}^{\phi\,\mathbf{T}} \right) \bullet \tilde{\boldsymbol{W}}^{\tau} \right) \right)}{\sum_{\phi} \boldsymbol{1} \overset{\rightarrow\tau}{\boldsymbol{H}}{}^{\phi\,\mathbf{T}} + \tilde{\boldsymbol{W}}^{\tau} \mathrm{diag}\left(\sum_{\tau} \boldsymbol{1} \left(\left(\left(\overset{\uparrow\phi}{\boldsymbol{V}} \Big/ \overset{\uparrow\phi}{\tilde{\boldsymbol{\Lambda}}} \right) \overset{\rightarrow\tau}{\boldsymbol{H}}{}^{\phi\,\mathbf{T}} \right) \bullet \tilde{\boldsymbol{W}}^{\tau} \right) \right)}$ (Kullback Liebler)

G) Repeat from step B) until convergence.

base and the others including flute, bass and drum music. All mixtures are sampled at 16 kHz sampling rate and the length of all test signals was chosen to be (40,000 samples, approximately 2.5 s). The time-frequency representation was computed by normalizing the time-domain signals to unit power and we computed the STFT using 1024 point Hanning window FFT with 50% overlap. The spectrogram bins are grouped into 175 logarithmically spaced frequency bins in the range of 50 Hz to 8 kHz with 24 bins per octave, which corresponds to twice the resolution of the equal tempered musical scale. As for the filterbank, the parameter corresponding to the total number of filters K is set as 4 and the length of the hamming window is defined equal to 128. As for SNMF2D parameters, the convolutive components in frequency and time were selected as $\tau = \{0,\cdots,4\}$ and $\phi = \{0,\cdots,10\}$, respectively. The sparse regularization term is set to $\lambda = 6$. **Figure 2** shows the design of four sub-bands.

Using the filterbank is very useful and helpful for the separation stage. This is because one of original sources may centralize its basic frequency information in a specific sub-band such that the dominant source can be easier extracted using source separation algorithms such as the SNMF2D. In the separation stage, the observed signal in each sub-band is converted into the log-frequency spectrogram and decomposed by SNMF2D. The cost value of decomposing female speech mixed with bass music in each sub-band is shown in **Figure 3**. It is observed that the decomposition process converges to a low steady value after approximately 40 iterations for all sub-band mixtures by using the SNMF2D algorithm. **Figure 4** shows an example of $\boldsymbol{H}$ and $\boldsymbol{W}$ as decomposed in the fourth sub-band mixed signal. It is seen that the spectral bases and temporal codes of each source are distinguishable so that each spectral basis can represent the frequency patterns of one sub-sources. The example of final separation results are shown in **Figure 5**.

The measure distortion between the original source

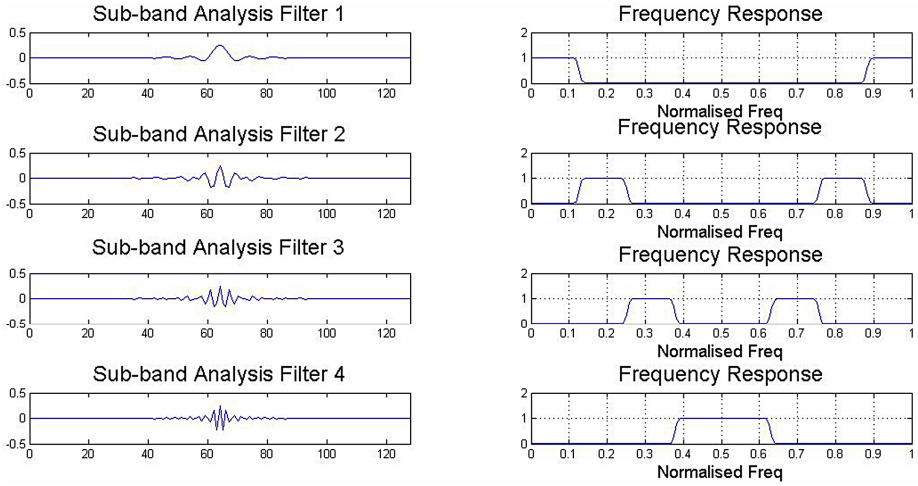


Figure 2. Filterbank designs.

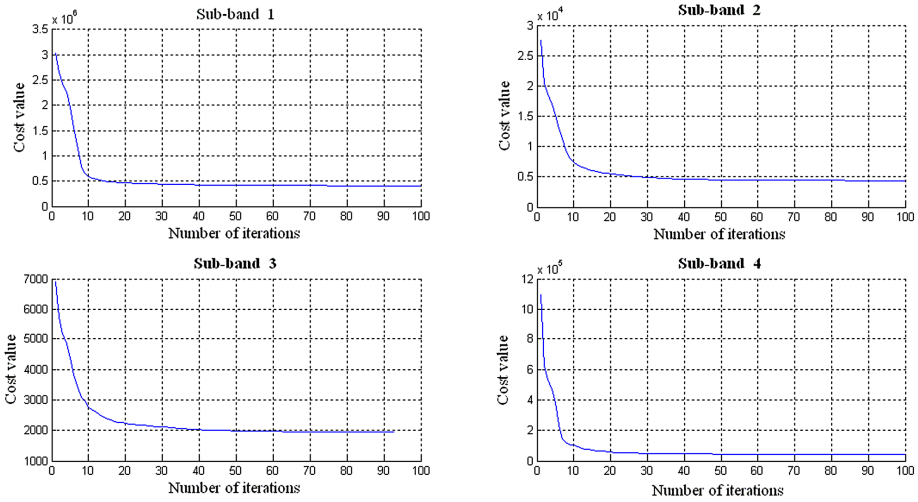


Figure 3. Convergence of LS cost function.

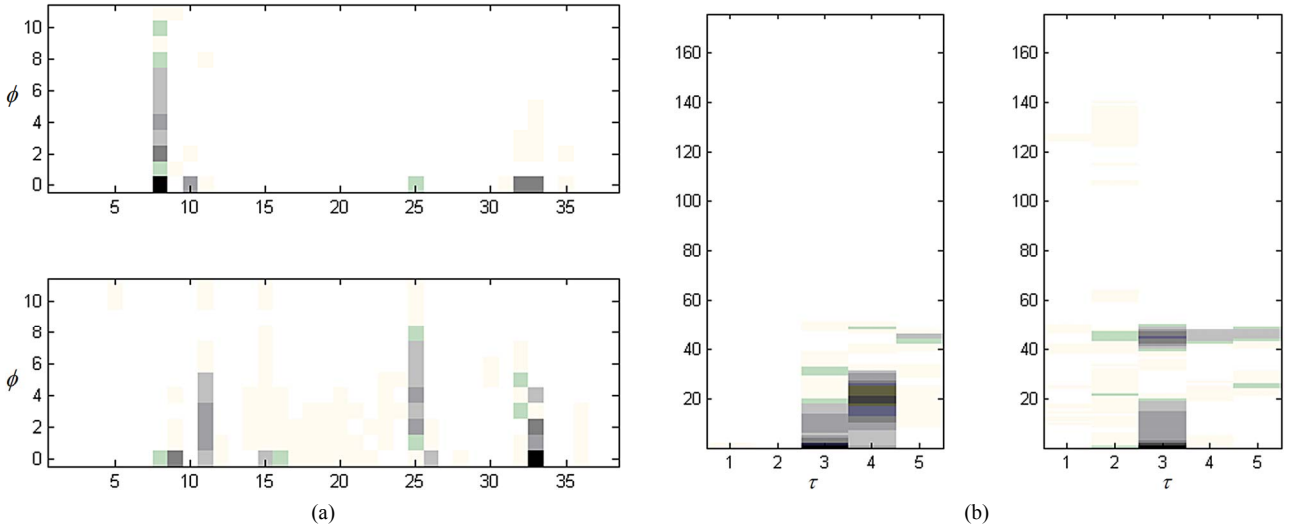


Figure 4. (a) H (in fourth sub-band); (b) W (in fourth sub-band).

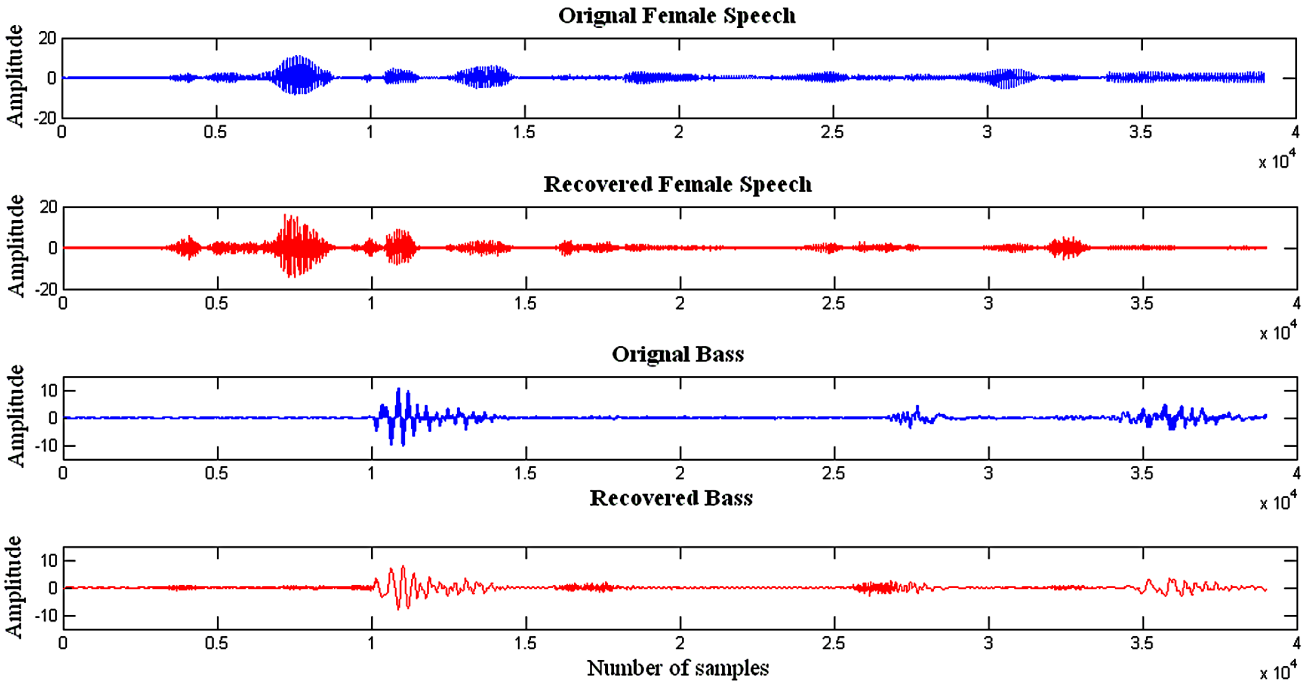


Figure 5. Original signals (blue) and recovered signals (red) using proposed method.

and the estimated one is computed by using the improvement of signal-to-noise ratio (ISNR) [22] defined as:

$$ISNR_d = 10\log_{10}\frac{\sum_t |s_d(t)|^2}{\sum_t |s_d(t)-\hat{s}_d(t)|^2} - 10\log_{10}\frac{\sum_t |s_d(t)|^2}{\sum_t |x(t)-s_d(t)|^2} \quad (17)$$

The ISNR is used as the quantitative measure of separation performance and the average ISNR will be tabulated in the evaluation tables. The ISNR represents the degree of suppression of the interfering signals to improve the quality of the target signal. It has been commonly used to measure the separation quality between the mixed and separated signal. The higher value of ISNR indicates better separation performance. In this paper, six types of mixture have been generated: 1) flute mixed with male speech; 2) flute mixed with female speech; 3) bass mixed male speech; 4) bass mixed female speech; 5) drum mixed male speech and 6) drum mixed female speech. All separation results have been summarized in **Figure 6** where $\{M, W, F, B, D\}$ represent

male speech, female speech, flute music, bass music and drum music. The separation of speech-bass music mixture is much better than those of other types of mixtures where the average ISNR has approached to 10 dB for recovered speech signal and 4 dB for recovered bass music.

Figure 6 summarizes the separation results of our proposed method. It is worth pointing out that because the frequency range of bass and drum music locate at very lower frequency region, the lower frequency bands are dominated with most energy from the bass or drum components through filterbank process. Hence, it is easier to extract these lower frequency components by using the SNMF2D. Thus, **Figure 6** shows the relatively better separation results when audio mixture contains bass or drum music. On the other hand, the frequency range of flute is very similar to speech sources (as indicated in **Figure 9**) and this particular mixture is very difficult to separate which explains the reason why the ISNR is relatively low. However, this performance is still substantially better than using the SNMF2D alone.

4. Discussion

4.1. Effects on Audio Mixtures Separation with/without Filterbank Preprocessing

The benefits filterbank preprocessing bring to SCSS is that since a filtered signal bounded within a particular range of sub-band frequencies, the complexity of the spectral and temporal patterns associated with each sub-band signal will be simpler and sparser than that of the mixed signal. This effectively means that there is a relatively clear distinction of the spectral and temporal patterns between the dominating source and the less dominating one in the TF domain in each sub-band. This is shown in **Figures 7** and **8**.

Figures 9 and **10** further show the time domain sub-band signal. It is clearly visible that the mixing at the different sub-band is dominated either by Source 1 or Source 2. In this example, it can be seen that flute music dominates the 1st sub-band while male speech dominates the 2nd-4thsub-band. The final comparison results of audio mixtures separation with/without filter-bank preprocessing are given in **Figure 11**.

4.2. Impact of Sparsity Regularization

In the separation stage, λ (sparse regularization), an essential parameter influences separation results. In **Figure 7**, we use an example-mixture of male speech and flute music for analyzing the impact of sparsity regularization. The separation results are concluded given different levels of sparse λ based on either LS or KLd cost functions. It is observed that the best ISNR has been found with the sparse factor $\lambda = 6$ by using the LS cost function and $\lambda = 20$ by using the KLd cost function. In addition, the LS cost function based decomposition reflects the local minimum whereas the KLd based decomposition returns the global minimum. However, our results have shown that both LS and KLd methods give comparable performance as shown in **Table 2**.

In this section, we develop a test to compare the separation performance between the proposed method and SNMF2D SCSS method. **Figure 11** shows that the ISNR results obtained using the proposed method which renders considerable improvements over the SNMF2D SCSS method. An average improvement of 1.8 dB per source is obtained across all the different type of mixtures for proposed method when compared to SNMF2D SCSS method. The specific comparison results are summarized as follows: 1) for mixture of speech and flute music, the average improvement is about 3.4 dB; 2) for mixture of speech and bass music, the improvement is 1.5 dB; 3) for mixture of speech and drum music, the average improvement is approximately 0.2 dB.

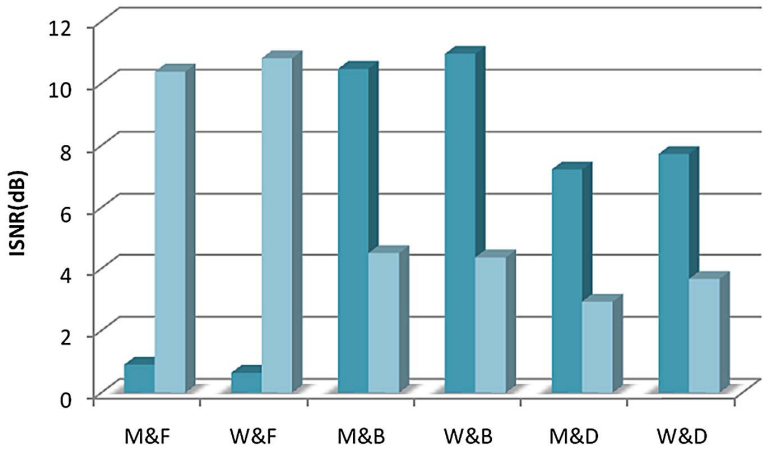
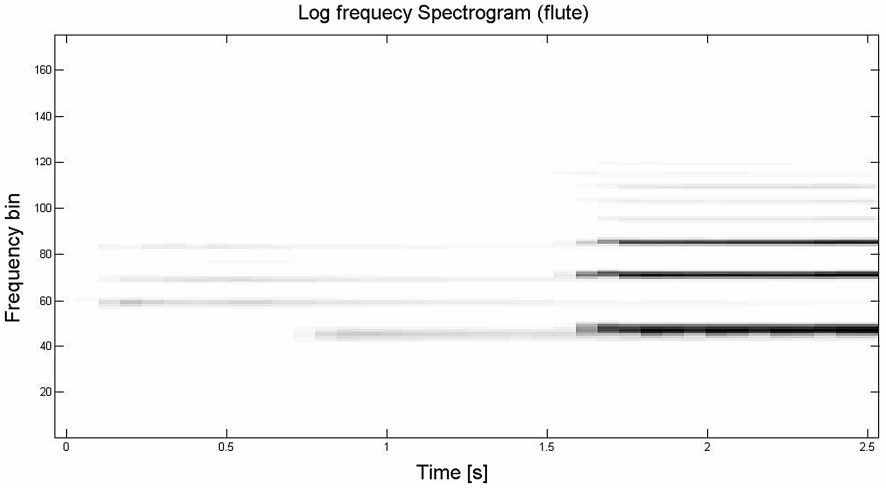
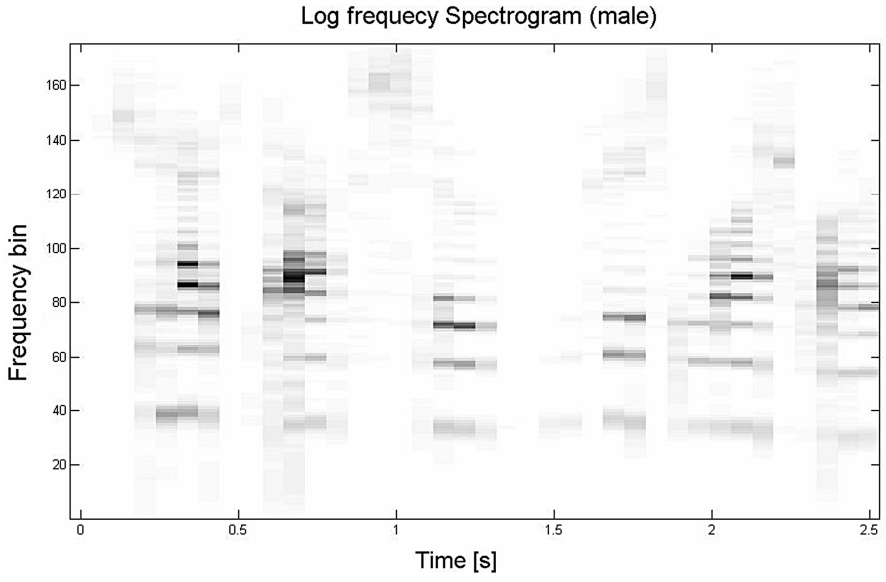


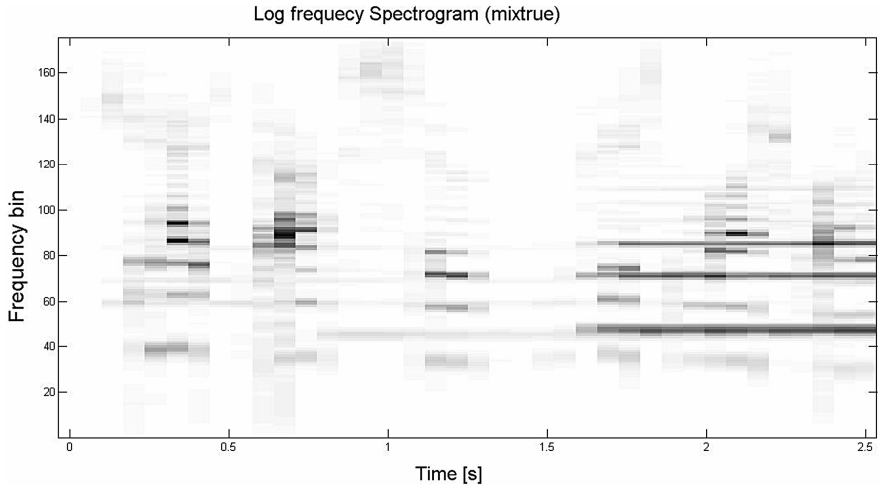
Figure 6. Separation results using the proposed method.



(a)



(b)



(c)

Figure 7. (a) and (b) denote the log-frequency spectrogram of flute music and male speech, respectively; (c) denotes the log-frequency spectrogram of mixed signal (flute + male).

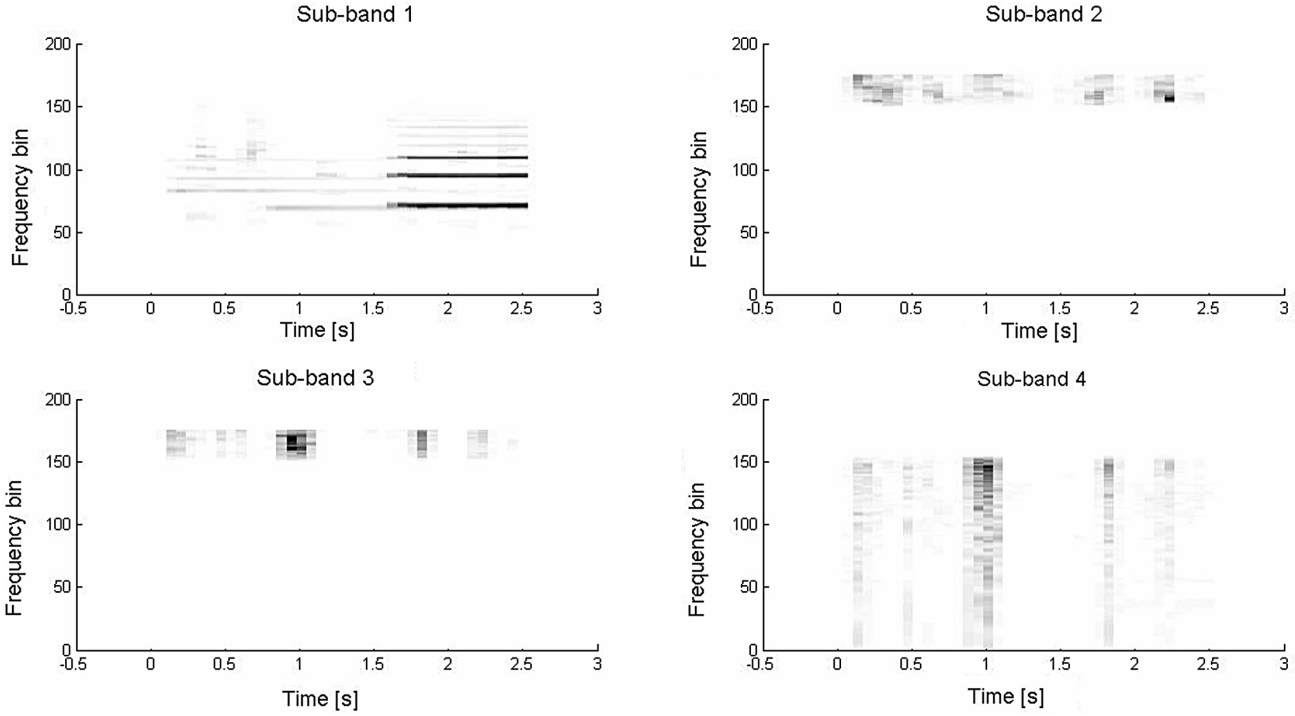


Figure 8. Log-frequency mixed spectrogram with filter-bank processing.

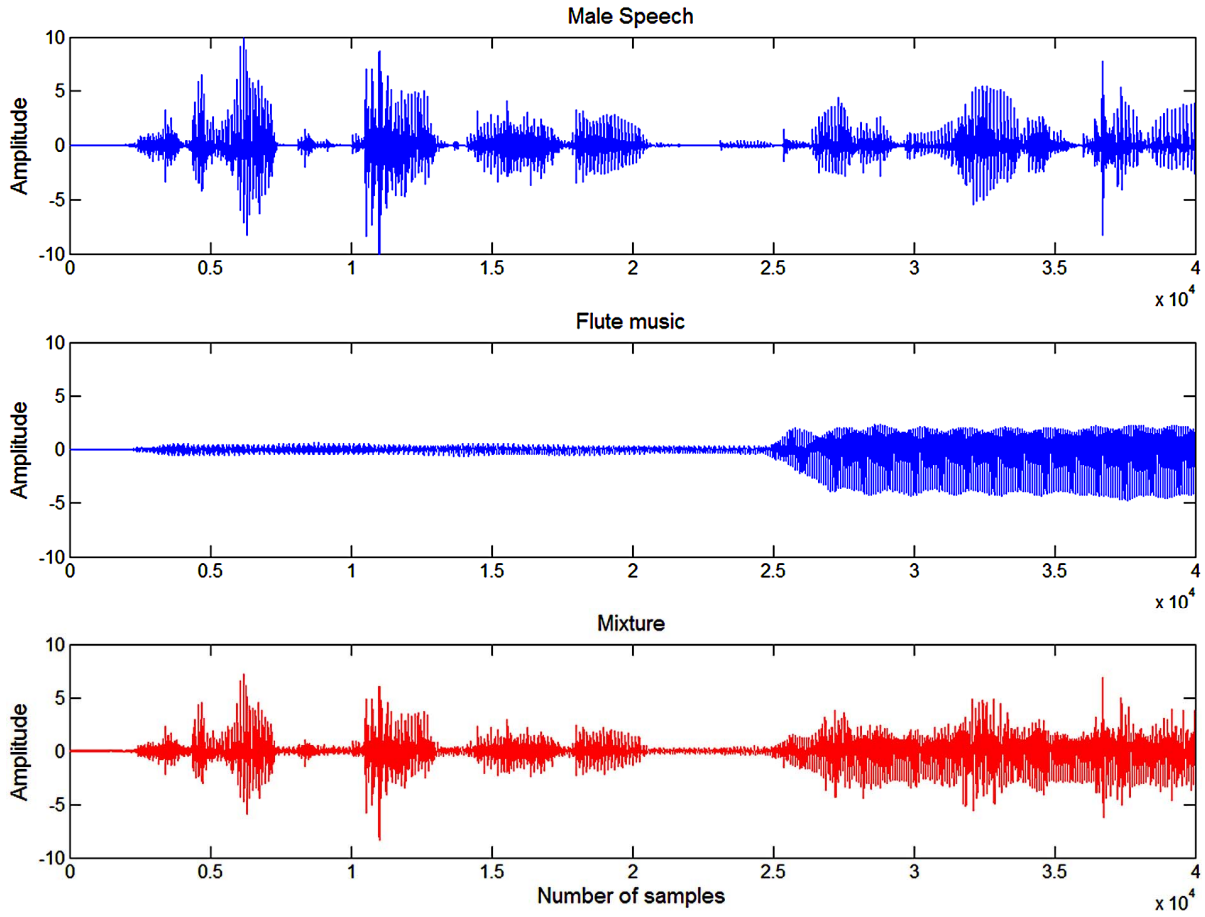


Figure 9. Time domain signals (flute music and male speech).

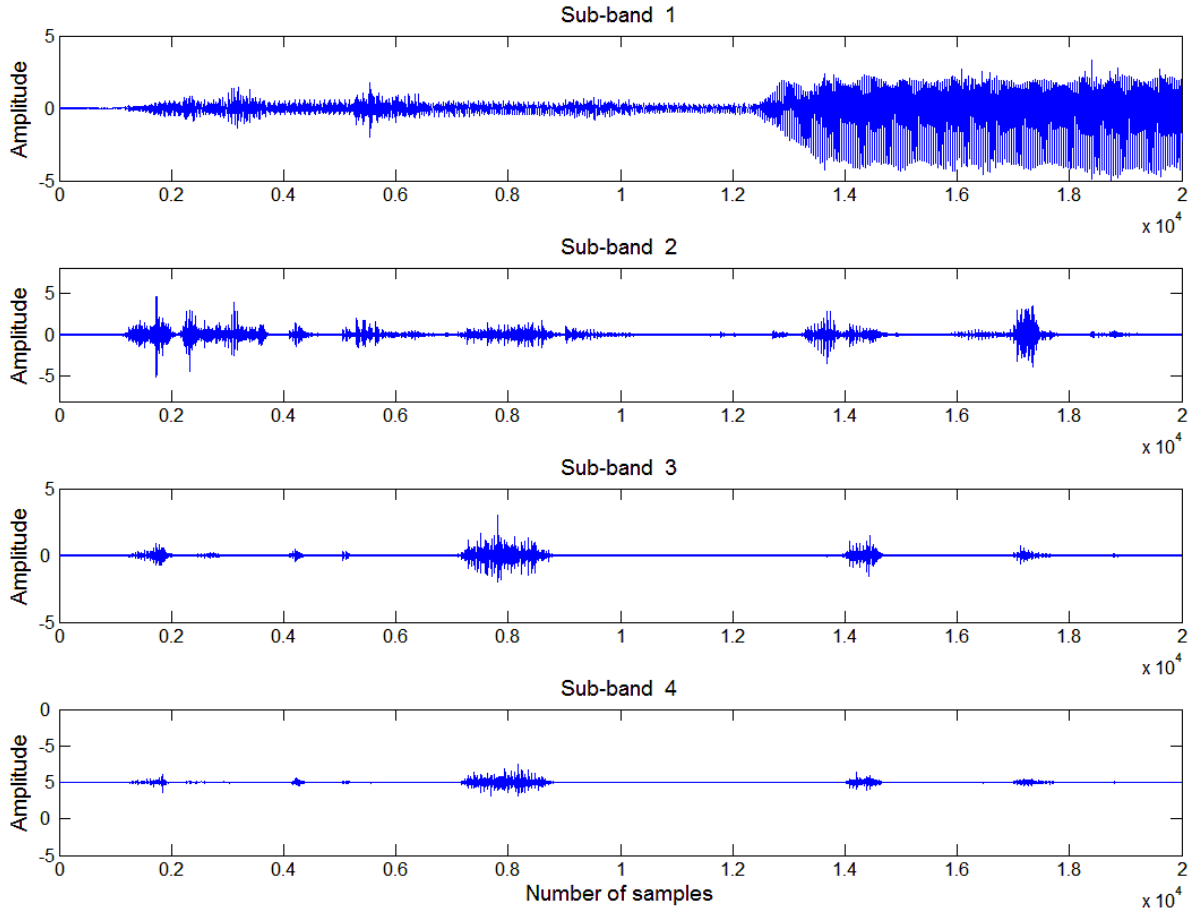


Figure 10. Time domain sub-band signals with filter-bank processing.

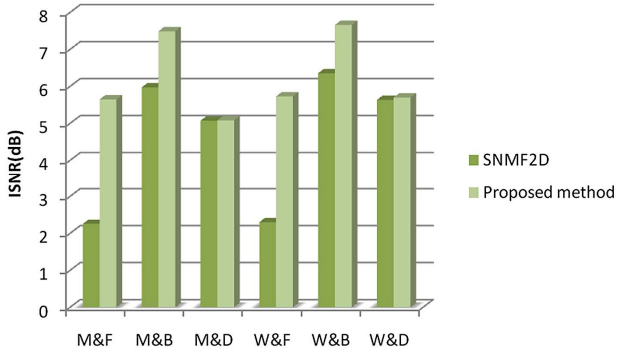


Figure 11. Overall comparison result.

Table 2. Separation results by using different sparse regularization.

λ	LS	KL
0	2.92	5.14
0.5	3.50	5.16
1.5	2.82	4.99
6	**5.64**	4.48
10	2.69	5.00
20	2.78	**5.17**
50	2.16	4.73

5. Conclusion

This paper has presented a novel framework of amalgammating filterbank technique with two-dimensional sparse non-negative matrix deconvolution (SNMF2D) for single channel source separation. Although proposed method and the SNMF2D SCSS method can extract sources from single channel mixture, the results obtained from our approach outperform that of using the SNMF2D. The strength of the proposed method: 1) it does not rely on training information so that it is more practical; 2) the degree of mixing ambiguity in each sub-band is less ambiguous than those in mixed signal; therefore the sub-band mixtures are simpler and sparser, and hence the spectral and temporal patterns can be efficiently extracted. Considerable improvements have been achieved in terms of ISNR by using our proposed method.

REFERENCES

[1] T. Kristjansson, H. Attias and J. Hershey, "Single Microphone Source Separation Using High Resolution Signal Reconstruction," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Quebec, 17- 21 May 2004, pp. 817-820.

[2] B. Gao, W. L. Woo and S. S. Dlay, "Single Channel Source Separation Using EMD-Subband Variable Regularized Sparse Features," *IEEE Transactions on Audio*, *Speech and Language Processing*, Vol. 19, No. 4, 2011, pp. 961-976. doi:10.1109/TASL.2010.2072500

[3] M. H. Radfa and R. M. Dansereau, "Single-Channel Speech Separation Using Soft Mask Filtering," *IEEE Transactions on Audio*, *Speech and Language Processing*, Vol. 15, No. 8, 2007, pp. 2299-2310.

[4] D. Ellis, "Model-Based Scene Analysis," In: D. Wang and G. Brown, Eds. *Computational Auditory Scene Analysis*: *Principles*, *Algorithms*, *and Applications*, Wiley/ IEEE Press, New York, 2006.

[5] T. Kristjansson, H. Attias and J. Hershey, "Single Micro-Phone Source Separation Using High Resolution Signal Reconstruction," *Proceedings of the International Conference on Acoustics*, *Speech*, *and Signal Processing*, Montreal, 17-21 May 2004, pp. 817-820.

[6] B. Gao, W. L. Woo and S. S. Dlay, "Adaptive Sparsity Non-Negative Matrix Factorization for Single Channel Source Separation," *IEEE Journal of Selected Topics in Signal Processing*, Vol. 5, No. 5, 2011, pp. 1932-4553.

[7] G. J. Brown and M. Cooke, "Computational Auditory Scene Analysis," *Computer Speech and Language*, Vol. 8, No. 4, 1994, pp. 297-336. doi:10.1006/csla.1994.1016

[8] M. Helén and T. Virtanen, "Separation of Drums from Polyphonic Music Using Nonnegative Matrix Factorization and Support Vector Machine," 13*th European Signal Processing Conference*, Turkey, 6 September 2005.

[9] P. Smaragdis and J. C. Brown, "Non-Negative Matrix Factorization for Polyphonic Music Transcription," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), 19-22 October 2003, pp. 177-180.

[10] R. Kompass, "A Generalized Divergence Measure for Nonnegative Matrix Factorization," *Proceedings of the Neuroinformatics Workshop*, Torun, September 2005.

[11] A. Cichocki, R. Zdunek, and S. I. Amari, "Csiszár's divergences for non-negative matrix factorization: family of new algorithms," *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Signal Separation*, Vol. 3889, Springer, Charleston, 2006, pp. 32-39. doi:10.1007/11679363_5

[12] P. D. O. Grady, "Sparse Separation of Under-Determined Speech Mixtures," Ph.D. Thesis, National University of Ireland Maynooth, Kildare, 2007.

[13] D. Lee and H. Seung, "Learning the Parts of Objects by Nonnegative Matrix Factorisation," *Nature*, Vol. 401, No. 6755, 1999, pp. 788-791. doi:10.1038/44565

[14] D. D. Lee and H. S. Seung, "Algorithms for Non-Negative Matrix Factorization," MIT Press, Cambridge, 2002, pp. 556-562.

[15] M. Mørup and M. N. Schmidt, "Sparse Non-Negative Matrix Factor 2-D Deconvolution for Automatic Transcription of Polyphonic Music," Technical University of Denmark, Lyngby, 2006.

[16] A. Mertins, "Signal Analysis Wavelets, Filter Banks, Time-Frequency Transforms and Applications," John Wiley & Sons, Hoboken, 1999, pp. 143-195.

[17] B. Gao, W. L. Woo and S. S. Dlay, "Variational Bayesian Regularized 2-D Nonnegative Matrix Factorization," *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 23, No. 5, 2012, pp. 703-716. doi:10.1109/TNNLS.2012.2187925

[18] Md. K. I. Molla and K. Hirose, "Single-Mixture Audio Source Separation by Subspace Decomposition of Hilbert Spectrum," *The IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 3, 2007, pp. 893-900.

[19] J. Taghia and M. Ali Doostari, "Subband-Based Single-Channel Source Separation of Instantaneous Audio Mixtures," *World Applied Sciences Journal*, Vol. 6, No. 6, 2009, pp. 784-792.

[20] K. Kokkinakis and P. C. Loiziu, "Subband-Based Blind Signal Processing for Source Separation in Convolutive Mixtures of Speech," *IEEE International Conference on Acoustic*, *Speech and Signal Processing*, Honolulu, 15-20 April 2007, pp. 917-920.

[21] P. P. Vaidyanathan, "Multirate Systems and Filter Banks," Prentice-Hall, Englewood Cliffs, 1993.

[22] E. Vincent, R. Gribonval and C. Fevotte, "Performance Measurement in Blind Audio Source Separation," *The IEEE Transactions on Audio*, *Speech and Language Processing*, Vol. 14, No. 4, 2005, pp. 1462-1469.

