Underdetermined Blind Source Separation using Binary Time-Frequency Masking with Variable Frequency Resolution

Amod J G Anandkumar[†], *MIEEE*, Aneesh Ghosh T. A., B. Teja Damodaram, Sumam David S.[‡] *SMIEEE*.
Department of Electronics and Communications Engineering, National Institute of Technology Karnataka,
Surathkal, Mangalore, 575 025, INDIA. Email: [†]amodjaiganesh@ieee.org, [‡]sumam@ieee.org

Abstract—**A novel method based on an algorithm by Pedersen *et. al.* using independent component analysis and binary time-frequency masking to iteratively segregate sources from a stereo recording is proposed here to improve the quality of estimated sources and reduce artifacts (musical noise). The inspiration comes from research in auditory scene analysis which indicates that the human auditory system uses a spectrogram-like time-frequency representation with variable frequency resolution for performing stream segregation and perceptual grouping. A time-frequency transform with a variable frequency based on the discrete cosine transform and amplitude modulation is suggested. A method for perfect reconstruction and a way to exploit short-time stationarity of the sources have also been suggested. Simulation results indicate significant improvements in objective evaluation criteria like percentage of energy loss and signal-to-noise ratio improvement. Subjective listening tests indicate a marked improvement in estimated signal quality with strongly reduced artifacts.**

I. Introduction

Blind source separation refers to the process of recovering the original signals from mixtures observed by spatially diverse sensors, without any information about either the mixing process or the source signals. Blind source separation has application in many scenarios, the most prominent of which involves recovering the speech of a particular speaker in a room of multiple simultaneous and independent speakers. In literature, this problem is known as the *cocktail party problem*. In this context, the term ‘blind’ stresses the fact that neither the mixing process nor the source signals are known. This could arise when undertaking the modeling of the system is intractable or when *a priori* knowledge of the mixing process is unavailable. This lack of *a priori* information is compensated by a statistically strong and physically plausible assumption of *independence* of the source signals. Blind source separation is presently quite a mature field with application in many areas of research and even has a few textbook-length reviews [1]–[4].

In the simplest mixing model, termed as *instantaneous mixing*, each recorded signal consists of a sum of differently weighted sources. But this is too simplistic to model many real-world scenarios, such as the cocktail-party problem which involves recovering speech of a particular speaker in a room of multiple simultaneous and independent speakers. The acoustics of such a room environment cannot be modeled as a simple instantaneous weighted mixture due to the presence of multipath propagation and reverberations since here the mixtures are weighted and delayed. Such filtered sums of different sources are called *convolutive mixtures*.

It is typically assumed that the number of sensors is no less than the number of sources, in which case, linear methods are sufficient to determine the mixing process. When the number of sensors exceeds the number of sources, the problem is called *overdetermined*,when it is equal it is called *fully-determined* otherwise it is called *underdetermined* (or *overcomplete*). The fully determined and overdetermined cases have been extensively studied by researchers over the past decade [5]. However, the underdetermined case has proven to be much more challenging and has received lesser attention. In the underdetermined case, linear methods are insufficient to recover the sources, even with the perfect knowledge of the mixing process [5]. In other words, the sources cannot be estimated completely as information is lost during the mixing process. Additional assumptions are needed to estimate the source signals. Our aim is to use novel ideas inspired from auditory scene analysis to improve underdetermined blind source separation algorithms. We have achieved significant improvements in objective evaluation criteria such as the percentage of energy loss and signal-to-noise ratio improvement. Subjective listening tests have shown that our method generates estimated sources with lesser noise and artifacts and slightly reduced interference.

Auditory scene analysis (ASA) is the process by which the human auditory system organizes complex mixtures of sound [6]. Human audition is surprisingly complex and intricate. For example, when we listen to an orchestra, we perceive music as a whole rather than hearing many individual instruments being played simultaneously. This is an example of *perceptual grouping*. On the other hand, in the cocktail party scenario, when there are many speakers around us in a room, we are still able to carry on a conversation with relative ease. This is an example of *stream segregation*. Computational auditory scene analysis (CASA) is the study of auditory scene analysis by computational means. In essence, CASA systems are machine listening systems that aim to separate mixtures of sound sources in the same way that human listeners do.

Related work and Contributions

Researchers have used various assumptions about source locations, distributions and other parameters to solve the underdetermined blind source separation problem. For example, sparsity in time, frequency or time-frequency domain may be assumed [7]. However, if there is an overlap, then binary time-frequency masking can be used to achieve good separation [8]. Time-frequency masking also has often been used in computational auditory scene analysis [9]. Pedersen *et. al.* [10]–[12] have utilized independent component analysis (ICA) and binary time frequency masking to devise an iterative algorithm to perform underdetermined blind source separation. We propose a method based on this approach for instantaneous mixtures which uses a time-frequency (T-F) transform having variable frequency resolution inspired from the human auditory system. Discrete cosine transform and amplitude modulation are used to generate a time-frequency transform with variable frequency resolution. Also, a method for perfect signal reconstruction and a way to exploit short time stationarity have been suggested.

The organization of the paper is as follows: we first present a brief discussion of general blind source separation principles, after which an overview of the method due to Pedersen *et. al.* is reviewed. Next, an novel method using a time-frequency transform with variable frequency resolution achieved using the discrete cosine transform and amplitude modulation along with a method for perfect signal reconstruction and a way to exploit short time stationarity are suggested. Results of simulations comparing our method with the original Pedersen's algorithm are then presented and finally, conclusions are drawn.

II. Basic Principles of Blind Source Separation

The general generative blind source separation model, called the *convolutive mixing model*, can be described as follows: At the discrete time index t, a mixture of N statistically independent sources $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]^T$ is assumed to be recorded at M spatially diverse sensors. The M real, zero-mean sensor signals in vector form $\mathbf{x}(t) = [x_1(t), \ldots, x_M(t)]^T$ are linear mixtures of filtered versions of each of the source signals, along with some additive sensor noise in vector form $\mathbf{v}(t) = [v_1(t), \ldots, v_M(t)]^T$. Each sensor signal can be represented as

$$x_m(t) = \sum_{n=1}^{N} \sum_{k=0}^{K-1} a_{mnk} s_n(t-k) + v_m(t) \qquad m = 1, \ldots, M \tag{1}$$

where a_{mnk} represents the mixing filter coefficient. In theory, the mixing filters may be of infinite length (implemented as IIR systems), however, in practice it is sufficient to assume channel length $K < \infty$ due to the nature of the in-room acoustic environment. In matrix form, the convolutive model can be written as

$$\mathbf{x}(t) = \sum_{k=0}^{K-1} \mathbf{A}_k \mathbf{s}(t-k) + \mathbf{v}(t) \tag{2}$$

where $\mathbf{A}_k$ is an $M \times N$ matrix which contains the k-th mixing filter coefficients.

Assuming a noise-free scenario with all the signals arriving at the sensors at the same instant without undergoing any filtering, the convolutive model in (2) simplifies to

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{3}$$

which is the *instantaneous mixing model*. Here $\mathbf{A} = \mathbf{A_0}$ is an $M \times N$ matrix containing the mixing coefficients.

The crux of solving the blind source separation problem is in estimating both $\mathbf{s}(t)$ and $\mathbf{A}$ given only the mixtures $\mathbf{x}(t)$. Basic independent component analysis (ICA) algorithms assume the instantaneous mixing model and work quite effectively in performing the separation. It is done by first estimating the mixing matrix $\mathbf{A}$ and then computing its inverse $\mathbf{W}$. The estimated original sources $\mathbf{y}(t)$ are then computed using

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t) \tag{4}$$

Separation based on the convolutive mixing model is much more complex and difficult as the unmixing filter $\mathbf{W}$ must first be estimated after which the sources can be estimated as

$$\mathbf{y}(t) = \sum_{l=0}^{L-1} \mathbf{W}_l \mathbf{x}(t-l) \tag{5}$$

III. Overview of Under-determined Blind Source Separation using Binary T-F Masking

Pedersen *et. al.* use the output of a 2 × 2 ICA algorithm and binary time-frequency masking to separate stereo mixtures having an unknown number of sources [10]–[12]. A knowledge about the source geometry and a degree of sparsity, either in spatial, temporal or time-frequency domain is assumed. Time-frequency binary masking has the advantage that only binary decisions have to be made. It is an iterative procedure that runs until all the source signals are estimated. The results are improved by grouping similar signals. This algorithm has the advantage that it does not require previous knowledge of the number of sources in a stereo mixture. The algorithm has three stages: a *core procedure*, a *separation stage* and a *merging stage*. The inputs the algorithm are the two mixed signals $x_1(t)$ and $x_2(t)$ of duration N_s.

A. Core Procedure

The core procedure accepts a binary mask $BM(\omega, t)$ and two mixed signals x_a and x_b as input. For the initial iteration, these inputs are the stereo mixture $x_1(t)$ and $x_2(t)$ and the binary mask $BM(\omega, t) = 1, \forall\ \omega, t$. These two signals are separated into two independent components y_1 and y_2 by a 2×2 ICA algorithm. The scaling ambiguity of the ICA algorithm is overcome by normalizing y_1 and y_2 to get $\hat{y}_1$ and $\hat{y}_2$ respectively. These are then transformed into the T-F domain using the short-time Fourier transform (STFT) to get the two spectrograms,

$$\hat{y}_1 \longrightarrow \mathbf{Y}_1(\omega, t) \tag{6}$$

$$\hat{y}_2 \longrightarrow \mathbf{Y}_2(\omega, t) \tag{7}$$

where ω is the frequency bin and t is the time window index. Two binary masks $BM_1(\omega, t)$ and $BM_2(\omega, t)$ are estimated by comparing the magnitudes of the two spectrograms at each T-F unit using

$$BM_1(\omega,t) = \begin{cases} BM(\omega,t), \text{if } |Y_1(\omega,t)| > \tau|Y_2(\omega,t)| \\ 0, \qquad\qquad\;\; \text{if } |Y_2(\omega,t)| > \tau|Y_1(\omega,t)| \end{cases} \quad \forall\; \omega,t \tag{8}$$

$$BM_2(\omega,t) = \begin{cases} BM(\omega,t), \text{if } |Y_2(\omega,t)| > \tau|Y_1(\omega,t)| \\ 0, \qquad\qquad\;\; \text{if } |Y_1(\omega,t)| > \tau|Y_2(\omega,t)| \end{cases} \quad \forall\; \omega,t \tag{9}$$

where τ is a parameter and $|Y(\omega,t)|$ denotes the magnitude of the spectrogram $Y(\omega,t)$. The amount of interfering signal removed at each iteration is decided by the sparsity of the mask which is controlled by the parameter τ in 8 and 9. When $\tau = 1$ the two estimated masks together contain the same number of retained T-F units as the previous mask. If $\tau > 1$, the resulting mask is more sparse than the previous mask and the convergence is faster. The case $0 < \tau < 1$ is not considered, as some T-F units would be assigned the value 1 in both estimated masks.

Each of these binary masks are then multiplied with the spectrograms of the two input signals to get four output spectrograms $X_{1a}(\omega,t)$, $X_{1b}(\omega,t)$, $X_{2a}(\omega,t)$ and $X_{2b}(\omega,t)$, which are then transformed to the time domain using the inverse STFT to get four time domain signals,

$$X_{1a}(\omega,t) \quad \longrightarrow \quad x_{1a}(t) \tag{10}$$

$$X_{1b}(\omega,t) \quad \longrightarrow \quad x_{1b}(t) \tag{11}$$

$$X_{2a}(\omega,t) \quad \longrightarrow \quad x_{2a}(t) \tag{12}$$

$$X_{2b}(\omega,t) \quad \longrightarrow \quad x_{2b}(t) \tag{13}$$

Thus, two binary masks and two pairs of masked output signals are generated by each instance of the core procedure.

B. Separation Stage

The separation stage is the repeated and iterative application of the core procedure. At the end of each instance of the core procedure, each of the masked output signals is classified into one of the following categories, based on the stopping criterion:

1) The masked signal is of poor quality.
2) The masked signal consists of mainly one source signal.
3) The masked signal consists of more than one source signal.

If the stopping criterion indicates that the mask of a signal has too few T-F units, the signal will have many artifacts (musical noise). It is marked as poor quality signal and stored for later use. If the signal is in the second category, it is stored as a candidate for a separated source signal. For the first two categories, no further processing in the separation stage is done. If it is the third case, further separation is done by processing it through another instance of core procedure. After each instance of core procedure, the masks become sparser. This process is continued till there are no signals containing more than one source signal.

Stopping Criterion: Consider the noisy instantaneous mixing model in vector form,

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{v} \tag{14}$$

where $\mathbf{v}$ is the sensor noise. Assuming that the noise is independent with variance σ^2, the covariance matrix $\mathbf{R_{xx}}$ can be written as function of the mixing matrix and the source signals as

$$\begin{aligned} \mathbf{R_{xx}} &= E\{\mathbf{x}\mathbf{x}^T\} \\ &= \mathbf{A}E\{\mathbf{s}\mathbf{s}^T\}\mathbf{A}^T + E\{\mathbf{v}\mathbf{v}^T\} \\ &= \mathbf{A}\mathbf{R_{ss}}\mathbf{A}^T + \sigma^2\mathbf{I} \end{aligned} \tag{15}$$

where $E\{.\}$ is the statistical expectation operation. It is assumed that the masked sensor signal consists of a single source if the condition number (based on the 2-norm) is greater than a threshold τ_c, *i.e.*,

$$cond(\mathbf{R_{xx}}) > \tau_c \tag{16}$$

A high condition number indicates that the matrix is close to being singular. Since $\mathbf{R_{xx}}$ is symmetric and positive definite,

$$cond(\mathbf{R_{xx}}) = \frac{max \ \mathrm{eig}(\mathrm{R_{xx}})}{min \ \mathrm{eig}(\mathrm{R_{xx}})} \tag{17}$$

where $\mathrm{eig}(\mathrm{R_{xx}})$ is the vector of eigenvalues of $\mathbf{R_{xx}}$.

C. Merging Stage

So far in this procedure, there is no guarantee that multiple estimated masks do not result in the same source signal. To improve the possibility of segregating all source signals and reduce the possibility of segregating the same source repeatedly, the merging stage is applied. Merging stage also helps improve the quality of the separated signals.

The output of the separation stage consists of the k segregated sources $\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_k$, the l segregated signals of poor quality $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_l$, and their corresponding binary masks $BM_{\hat{s}_i}$ and $BM_{\hat{p}_i}$ respectively. To identify whether two masks have led to the same signal, the correlation between the signals in the time domain is considered by computing the normalized correlation coefficients. If the correlation coefficient is found to be greater than some threshold τ_{C1}, a new binary mask is created by applying the logical OR operation to the two masks associated with the two correlated signals. Now a new signal is generated using this new binary mask. This step reduces the total number of signals containing segregated sources. Next, the poor quality signals are used to improve the segregated source estimates. The normalized correlation coefficients between the poor quality signals and signals having one source are computed. If it is found to be greater than some threshold τ_{C2}, then the binary mask of the poor quality signal is merged with that of the signal containing one source by applying the logical OR operation. The new mask will be less sparse and thus is expected to improve the quality of the estimated source.

There is a possibility that some of the original sources in the mixture have not been assigned to one of the segregated sources $\hat{s}_i$. The T-F units containing these unassigned sources are assigned to a *background mask*. The background mask is computed by

$$BGM = \overline{BM_{\hat{s}_1} \vee BM_{\hat{s}_2} \vee \cdots \vee BM_{\hat{s}_N}} \qquad (18)$$

This background mask is applied to the original stereo mixture to remove these sources. The new signals now generated are fed into the separation stage and the process is repeated. This is continued till the background mask generated does not change from the background mask of the previous iteration.

IV. Proposed Method for Underdetermined Blind Source Separation

A. Discrete Cosine Transform with Amplitude Modulation

Research in Auditory Scene Analysis (ASA) and human audition have given us much insight into the way humans quite successfully perform source separation in the cocktail party scenario. It is now known that a long coiled ribbon-like structure called the basilar membrane is present in the inner ear and is responsible for converting the acoustic energy into specific neural signals [6]. Different segments of this structure resonate at different frequencies and thus activate nerves near them only if the sound has frequency components near its resonant frequency. The amplitude of oscillation is also indicative of the magnitude of that particular frequency component. This suggests that the brain gets information in the T-F domain similar to a spectrogram. Research on perceptual grouping effects and stream segregation indicate that the separability of two tones is directly related to how far they are separated in the frequency domain, but on a logarithmic scale [6]. It has also been found that frequency resolution of the basilar membrane is logarithmic-like. *i.e.*, the frequency resolution at higher frequencies is lower than the frequency resolution at lower frequencies [6].

This information indicates that a T-F transformation with such a variable frequency resolution may perform better than using the STFT. The discrete cosine transform (DCT) has a strong energy compaction property and expands high frequencies and compresses low frequencies [13]. This is exactly opposite of what we are aiming to achieve. To counter this, amplitude modulation (AM) is performed on the stereo mixtures so that the spectra of the input signals are reversed. Now, when these signals are transformed to the T-F domain using short time DCT, the frequency components originally at the lower end of the spectrum are expanded and the components originally at the higher end are compressed. This results in a frequency resolution that is similar to the human auditory system in that the resolution at higher frequencies is lower than that at lower frequencies of the input signals. After the computations in the T-F domain, amplitude modulation is again performed to recover the signals. However, computing the DCT takes about twice as long as the STFT as the DCT is not symmetrical.

B. Method for Perfect Signal Reconstruction

When computing the transform from the time domain to the time-frequency domain, a STFT with rectangular window and no overlap between successive blocks of data is used. Overlapping is not used because when transforming the signal from the T-F domain back to the time domain, it leads to problems as there is no longer a one-one correspondence between the blocks in the time and T-F domains.

However, if overlapping between successive blocks is used, it may reduce the amount of artifacts in the resulting estimates source signals. To achieve a 75% overlap between successive blocks, each block is multiplied with the absolute value of a sinusoid having a period equal to the block length (say T) before applying the frequency transform. After the necessary computations are performed in the T-F domain, each block is transformed back to the time domain and again multiplied with the absolute value of the sinusoid used earlier. Any quarter of a block in the original signal is now present in four consecutive blocks. These may be added to get a scaled version of the original signal. If x_i is an element in a quarter of a block in the original signal, it is multiplied by $|\sin(\theta)|$, $|\sin(\theta + \frac{\pi}{2})|$, $|\sin(\theta + \pi)|$ and $|\sin(\theta + \frac{3\pi}{2})|$ respectively over four blocks. It is again multiplied by the same factors on returning to the time domain. Thus, when they are added we get

$$\begin{aligned} & x_i.\big(|\sin(\theta)|^2 + |\sin(\theta + \frac{\pi}{2})|^2 + |\sin(\theta + \pi)|^2 \\ & \qquad + |\sin(\theta + \frac{3\pi}{2})|^2\big) \qquad (19) \\ & = x_i.\big(2\sin^2(\theta) + 2\cos^2(\theta)\big) \\ & = 2x_i \end{aligned}$$

Thus, this procedure allows for perfect signal reconstruction after using overlapping windows for the T-F transform.

C. Temporal Mask Continuity

It can be assumed that the mask contents will not change significantly between consecutive blocks as the time duration of the step size from one block to the next is very small (of the order of 0.05s). This short range temporal stationarity is accounted when calculating the masks of any particular block by adding attenuated values of the nearby masks.

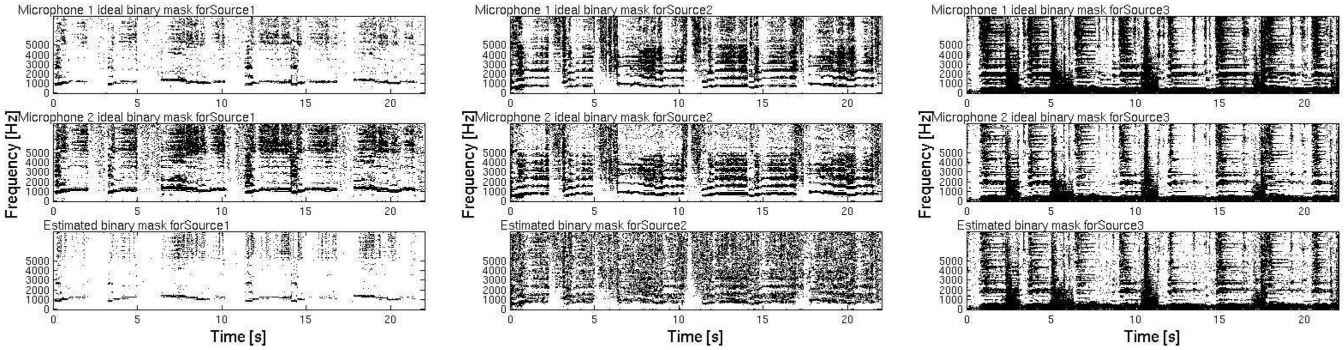
V. Results

A. Evaluation criteria

It is not possible to perfectly reconstruct the signals after separation using binary masks as the signals overlap. Hence, the concept of an *ideal mask* has been suggested as a suitable computational goal for separation [14]. The ideal binary mask for a signal is found for each T-F unit by comparing the energy of the signal to the energy of all the interfering signals. Whenever the signal energy is higher within a T-F unit, the T-F unit is assigned the value 1 and whenever the combined interfering signals have more energy, the T-F unit is assigned the value 0.

We use improvement in signal-to-noise ratio (SNR) in dB (ΔSNR), percentage of energy loss (P_{EL}) and the percentage of noise residue (P_{NR}) as the objective performance measures. Improvement in SNR is defined as

$$\Delta SNR = SNR_o - SNR_i \qquad (20)$$



(a) Mask outputs for Source1

(b) Mask outputs for Source2

(c) Mask outputs for Source3

Fig. 1. Typical binary mask outputs for a simulation with 3 sources

TABLE I
SIMULATION RESULTS WITH DIFFERENT ICA ALGORITHMS

Algorithm	Max. ΔSNR (dB)	Average ΔSNR (dB)	Min. P_{EL} (%)	Average P_{EL} (%)	Min. P_{NR} (%)	Average P_{NR}(%)
STFT with JADE	24.23	13.41	1.43	16.64	0.12	8.18
STFT with FastICA	23.97	14.04	1.03	14.02	0.11	6.58
STFT with Infomax ICA	23.56	13.68	1.49	16.54	0.17	8.37
DCT+AM with JADE	22.62	14.10	1.55	11.07	0.52	8.56
DCT+AM with FastICA	22.89	14.09	0.59	9.03	0.53	10.00
DCT+AM with Infomax ICA	22.35	14.37	1.04	9.47	0.49	8.15

where SNR_o is the output SNR after running the algorithm and SNR_i is the SNR before separation. The output SNR is defined as

$$SNR_o = 10\log_{10}\left[\frac{\sum_n I^2(n)}{\sum_n (I(n) - O(n))^2}\right] \tag{21}$$

where $O(n)$ is the estimated signal and $I(n)$ is the signal re-synthesized after applying the ideal mask. The SNR before separation is defined as the ratio between the desired signal and the interfering signals in the recorded masked mixtures. Percentage of energy loss and percentage of noise residue are respective defined as

$$P_{EL} = \frac{\sum_n e_1^2(n)}{\sum_n I^2(n)} \tag{22}$$

$$P_{NR} = \frac{\sum_n e_2^2(n)}{\sum_n O^2(n)} \tag{23}$$

where $e_1(n)$ denotes the signal present in $I(n)$ but absent in $O(n)$ and $e_2(n)$ denotes the signal present in $O(n)$ but absent in $I(n)$. P_{EL} can be considered as a weighted sum of the T-F unit power present in the ideal mask but absent in the estimated mask and P_{NR} as a weighted sum of the T-F unit power present in the estimated mask but absent in the ideal mask.

B. Simulation Results

We tested our algorithm with different speech mixtures and musical stereo mixtures obtained from [15], [16]. We ran extensive simulations of more than 1500 input signal combinations for each of the following variations of the algorithm in Matlab:

1) STFT with JADE
2) STFT with FastICA
3) STFT with Infomax ICA
4) DCT+AM with JADE
5) DCT+AM with FastICA
6) DCT+AM with Infomax ICA

JADE, FastICA and Infomax implementations in Matlab were downloaded from ICA Central website [17]. The value $\tau = 1$ was used for the parameter in (8) and (9). Typical binary mask outputs generated by the algorithm for a simulation with 3 sources is shown in Figure 1. The top two rows of masks are the ideal masks for the respective source and the last row shows the estimated masks. It can be seen that the estimated masks are very similar to the ideal masks, especially at the lower frequencies, due to the higher frequency resolution there. The simulation results are presented in Table I. It can be seen that the performance of both the original and the improved method does not depend much on the ICA algorithm used. Our method shows upto 5% increase in average improvement in SNR while the peak SNR improvement is reduced by upto 4%. Significant improvements of upto 42% are seen in average percentage of energy loss and upto 33% in peak percentage of energy loss, in the case of the Infomax ICA algorithm. This means that the estimated signals have much more spectral content of the original sources. However, this has simultaneously resulted in increased percentages of noise residue indicating that increased signal content is at the cost of increased residual interference from other sources.

Subjective listening tests have shown a clear improvement in the clarity of the estimated sources using the improved algorithm, with lesser background interference from other sources.

Musical noises (artifacts) are markedly reduced, particularly due to the temporal mask continuity.

VI. Conclusion

Underdetermined blind source separation is among the more complex problems in the area of blind source separation. There are currently existing algorithms that perform underdetermined blind source separation of stereo mixtures using independent component analysis and binary time-frequency masking. Research in auditory scene analysis has shown that human audition uses time-frequency representation with variable frequency resolution. Inspired by this, a time-frequency transform with a variable frequency based on the discrete cosine transform and amplitude modulation is suggested in this paper. Also, a method for perfect signal reconstruction and a way to exploit short time stationarity have been suggested. Simulation results indicate significant improvements in objective criteria like percentage of energy loss and SNR improvement. Subjective listening tests indicate a marked improvement in estimated signal quality with slightly reduced interference and greatly reduced artifacts. As further research, this method can be extended to the convolutive mixing model. Also, time-frequency transforms other than the DCT can be used to generate a variable frequency resolution.

Acknowledgment

The authors would like to thank Prof. Jonathon Chambers, Head of Advanced Signal Processing Group, Loughborough University, UK for his valuable guidance and feedback during the course of our research.

References

[1] A. Cichocki and S. I. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. J. Wiley and Sons, 2002.

[2] S. Makino, T. W. Lee, and H. Sawada, *Blind Speech Separation (Signals and Communication Technology)*. Springer, 2007.

[3] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. J. Wiley and Sons, 2001.

[4] T. Lee, *Independent component analysis: theory and applications*. Kluwer Academic Publishers, 1998.

[5] M. S. Pedersen, J. Larsen, U. Kjems, and L. Parra, “A survey of convolutive blind source separation methods,” *Springer Handbook on Speech Processing and Speech Communication*, Sept., 2007.

[6] A. S. Bregman, *Auditory Scene Analysis*. MIT Press, 1990.

[7] P. D. OGrady, B. A. Pearlmutter, and S. T. Rickard, “Survey of sparse and non-sparse methods in source separation,” *IJIST (International Journal of Imaging Systems and Technology), special issue on Blind Source Separation and Deconvolution in Imaging and Image Processing.*, 2005.

[8] N. Roman, D. L. Wang, and G. J. Brown, “Speech segregation based on sound localization,” *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, October 2003.

[9] D. L. Wang and G. J. Brown, “Separation of speech from interfering sounds based on oscillatory correlation,” *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 684–697, May 1999.

[10] M. S. Pedersen, D. L. Wang, J. Larsen, and U. Kjems, “Overcomplete blind source separation by combining ICA and binary time-frequency maskin,” in *Proc. MLSP workshop*, Sept., 2005.

[11] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems, “Two-microphone Separation of Speech Mixtures,” *IEEE Tran. Neural Network*, 2008.

[12] M. S. Pedersen, T. Lehn-Schiler, and J. Larsen, “BLUES from music: BLind Underdetermined Extraction of Sources from Music,” in *Proc. ICA*, 2006.

[13] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete time signal processing*. Prentice Hall.

[14] D. L. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” in *Speech Separation by Humans and Machines, P. Divenyi, Ed. Norwell*. MA: Kluwer, 2005, pp. 181–197.

[15] http://www.imm.dtu.dk/pubdb/p.php?4399.

[16] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, “First stereo audio source separation evaluation campaign: data, algorithms and results,” in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation*, 2007.

[17] http://www.tsi.enst.fr/icacentral/algos.html.