PERCEPTUALLY MOTIVATED BLIND SOURCE SEPARATION OF CONVOLUTIVE AUDIO MIXTURES WITH SUBSPACE FILTERING METHOD

Rammohana Reddy Guddeti and Bernard Mulgrew

Institute for Digital Communications
School of Engineering & Electronics
The University of Edinburgh
Edinburgh EH9 3JL U.K
E-mail: ram.guddeti@ed.ac.uk

ABSTRACT

In this paper, a perceptually motivated subspace filtering method is proposed for solving the permutation ambiguity of frequency-domain independent component analysis when the mixing environment is noisy and highly reverberant. In this method, perceptually irrelevant frequencies are first removed from the speech spectrum using block based perceptual masking (simultaneous frequency masking) before applying the subspace method followed by frequency-domain independent component analysis. After source separation in frequency domain, a physical property of the mixing matrix, i.e., the coherency in adjacent frequencies while checking the similarity mesure among spectral envelopes of the separated output for reduced frequencies, is utilized as a post processing tool for solving the permutation ambiguity. From the simulation results it appears that the perceptual masking avoids the permutation problem.

1. INTRODUCTION

Blind source separation (BSS) aims to recover independent sources from their multiple observed mixtures using independent component analysis (ICA). However, when applying BSS to audio mixture problem such as a number of people talking in a room, the performance of the system is greatly reduced by the effect of the room reflections and ambient noise. Humans deal with this real cocktail party effect very effi ciently by using only two ears (sensors). These perceptual masking techniques have been already exploited in successful development of MPEG audio coding standard (MP3 players).

Asano et al [1] have proposed the subspace method for reducing the effect of room reflections and ambient noise. Since the subspace method works in the frequency-domain, we must employ frequency-domain ICA (FDICA). The drawback of FDICA is permutation and scaling problem. For the scaling problem, the method proposed by Murata et al [2], in which the separated output is fi ltered by the inverse of the separation fi lter, shows good performance.

For the permutation problem, Asano et al [3] proposed a method that utilizes both the coherency of the mixing matrices and the correlation between spectral envelopes at several adjacent frequencies (denoted as inter frequency coherency (IFC)).

The authors [4] previously proposed a perceptually motivated FDICA method for solving the permutation problem. This method uses the simultaneous frequency masking (MPEG psychoacoustic model 1 [5]) for the complete omission of a signal componet at the given frequency.

In this paper, a perceptually motivated FDICA system with subspace approach for solving the permutation problem is proposed. This method utilizes both the simulatneous frequency masking for the complete omission of a signal at the given frequency and thereby using the subspace method for further reduction of room reflections.

This paper is organized as follows. In Section 2, an outline of the proposed perceptually motivated FDICA system with subspace method is presented for solving the permutation ambiguity. In Section 3, simulation results of experiments using both synthetic and real room recording speech data to evaluate the proposed perceptually motivated FDICA system are reported.

2. PERCEPTUAL FDICA SYSTEM

The flow of the proposed perceptual FDICA system is summarized in Fig.1.

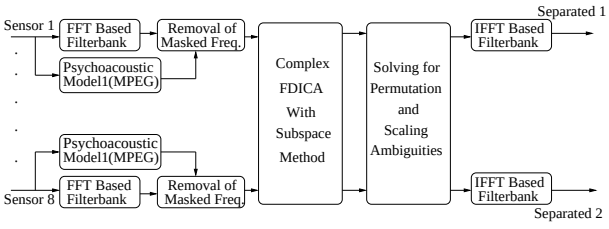


Figure 1: Perceptually Motivated FDICA System

First, the short time Fourier transform (STFT) of the multichannel input signal, $\mathbf{x}(\omega, t)$, is obtained with an appropriate time shift and Hann window function.
Next, psychoacoustic model 1 [5] is used to determine the masking threshold for each segment of speech and thereby obtaining a binary mask for each frequency.
A straightforward means to remove the masked frequency bins would be the multiplication of the complex spectrum of the input speech frame by the binary mask at each frequency bin. Thus, the thresholding in a stereo environment is described by logical AND operation.
The subspace method is then applied to the perceptually relevant spectral components of the input signal. In this stage, room reflections and ambient noise are reduced in advance of the application of FDICA. Next, the FDICA algorithm (complex Infomax [6–9]) is applied to the output of the subspace stage to obtain the separation filter.
After obtaining this filter, permutation and scaling problem is solved by processing the output of separation filter with the permutation and scaling matrices.
Finally, the filter matrices obtained in the above stages are transformed into the time domain and the input speech signal is processed with this time-domain filter network.

2.1. Model of Signal

Let us consider the case when there are *N* sound sources in the mixing environment with M sensors. By taking STFT of the sensor inputs, we obtain the input vector

$$\mathbf{x}(\omega, t) = [X_1(\omega, t), ..., X_M(\omega, t)]^T \tag{1}$$

Here, $\mathbf{X}_m(\omega, t)$ is STFT of the input signal in the *t*th time frame at the *m*th sensor. Further, the input signal is assumed to be modeled as

$$\mathbf{x}(\omega, t) = \mathbf{A}(\omega)\mathbf{s}(\omega, t) + \mathbf{n}(\omega, t) \tag{2}$$

$\mathbf{A}(\omega)$ is the mixing matrix and its (m, n) element, $A_{m,n}(\omega)$, being the transfer function from the *n*th source to the *m*th sensor as $A_{m,n}(\omega) = H_{m,n}(\omega)e^{-j\omega\tau_{m,n}}$. $\mathbf{s}(\omega, t)$ consists of the source spectra as $\mathbf{s} = [S_1(\omega, t), ..., S_N(\omega, t)]^T$.

2.2. Psychoacoustic Model 1

The ISO MPEG-1 [5] psychoacoustic model 1 uses a 512 point FFT for high resolution spectral analysis, then selects the perceptually relevant spectral components in each frame of the input speech by means of thresholding. This model assumes that the masking effects are additive.
In perceptual audio coding, thresholding sets the quanization level, here we set a threshold for further processing of the frequencies by ICA according to their psychoacoustic relevance and thereby reducing the computational complexity of solving the permutation problem.
While this thresholding is a nonlinear activity which might at first sight appeared to destroy the linear convolutive properties of the BSS, but it can also be viewed as an irregular sampling rate strategy which is linear. It will however alter the pdf of the signals presented to ICA.
Simultaneous masking refers to a frequency domain phenomenon which has been observed within critical bands. Sharp signal transients create premasking (backward temporal masking) and postmasking (forward temporal masking) in time during which a listener will not perceive signals beneath the elevated audible masking thresholds.
We didn't consider temporal masking based on the fact that our model is principally oriented to the speech signal that is stationary for a period shorter than 50 m sec.

2.3. Modified FDICA Algorithm

Whenever the perceptually masked input speech $\mathbf{x}(\omega, t)$ in one of the channels contains no values, the subspace filter matrix (special case of principal component analysis (PCA) with $M \gg N$, where *M* and *N* denote the number of nodes (channels) of the input and the output of PCA, respectively) $\mathbf{W}(\omega)$ is singular, resulting in rank deficiency.
Without loss of generality we have assumed identity matrix of order *N* for each pair of input nodes as the rank of subspace filter matrix $\mathbf{W}(\omega)$ to avoid this problem while retaining the whitening property of subspace filter.
Then, apply the complex Infomax algorithm for those frequency components of masked input speech, $\mathbf{y}(\omega, t)$, that contains nonzero values in both the channels in order to overcome the rank deficiency of ICA filter, $\mathbf{U}(\omega)$.
Thus the processing of ICA can be avoided whenever the masked input speech in one of the channels is zero.
In the ICA stage, the input signal $\mathbf{y}(\omega, t)$ is processed with the filter matrix $\mathbf{U}(\omega)$ as $\mathbf{z}(\omega, t) = \mathbf{U}(\omega, t)\mathbf{y}(\omega, t)$.
The ICA learning rule is given by

$$\mathbf{U}(\omega, t+1) = \mathbf{U}(\omega, t)+\eta[\mathbf{I}-\varphi(\mathbf{z}(\omega, t))\mathbf{z}^H(\omega, t)]\mathbf{U}(\omega, t) \tag{3}$$

Then, solve the scaling problem of FDICA by filtering individual output of the separation filter, $\mathbf{B}(\omega)$, (product of $\mathbf{W}(\omega)$ and $\mathbf{U}(\omega)$), by its pseudo inverse due to employment of the subspace method [3].
Finally, solve the permutation problem by utilizing both similarity measure among spectral envelopes of the separated output for frequencies that are perceptually relevant and the coherency of perceptually masked mixing matrices in several adjacent frequencies.
Without loss of generality assume zero cross correlation between spectral envelopes of the separated output when one of the channels does not contain any values and thereby avoiding rank deficiency of permutation matrix.

The cost function $F(\mathbf{P})$ is defined as

$$F(\mathbf{P}) = \frac{1}{N}\sum_{n=1}^{N}\cos\theta_n \tag{4}$$

Where, the cosine of the angle θ_n between the two location vectors in the adjacent frequencies, $\bar{\mathbf{a}}_n(\omega)$ and $\bar{\mathbf{a}}_n(\omega_0)$, of estimated mixing matrix is defined as

$$\cos\theta_n = \frac{\bar{\mathbf{a}}_n^H(\omega)\bar{\mathbf{a}}_n(\omega_0)}{\|\bar{\mathbf{a}}_n(\omega)\|.\|\bar{\mathbf{a}}_n^H(\omega_0\|} \tag{5}$$

In order to get reliable value of the cost function $F(\mathbf{P},k)$ at $\omega_0 = \omega - k.\Delta\omega$, for $k = 1,\cdots,K$, the confidence measure defined as

$$C(k) = \max_{\mathbf{P}\in\Omega}[F(\mathbf{P},k)] - \max_{\mathbf{P}\in\Omega'}[F(\mathbf{P},k)] \tag{6}$$

Here, Ω denotes the set of all possible $\mathbf{P}$ while Ω' denotes Ω without $\hat{\mathbf{P}} = \arg\max_{\mathbf{P}\in\Omega}[F(\mathbf{P},k)]$. The permutation is then solved at $\omega_0 = \omega - \hat{k}.\Delta\omega \quad (\hat{k} = \max_{\mathbf{P}}[C(k)])$ as

$$\hat{\mathbf{P}} = \arg\max_{\mathbf{P}}[F(\mathbf{P},\hat{k})] \tag{7}$$

The main contribution of this perceptual auditory masking and subspace method based preprocessor is not only the reduction of frequencies that are processed by ICA algorithm, but also the reduction of frequencies where the similarity to be checked for solving the permutation.

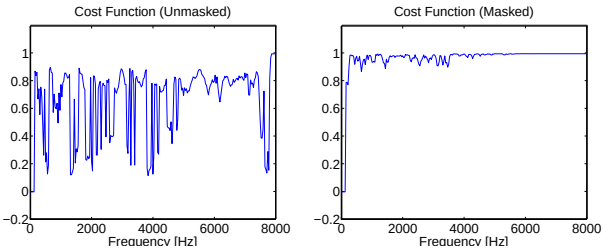
3. SIMULATION RESULTS

3.1. Experiment 1

This experiment was conducted with two speech sources (4 s at 16 kHz) and a circular microphone array (M = 8 and dia = 0.5 m) for simulating the room acoustic environment with reverberation time of 0.4 sec for both the weak and strong early reflection cases [3].

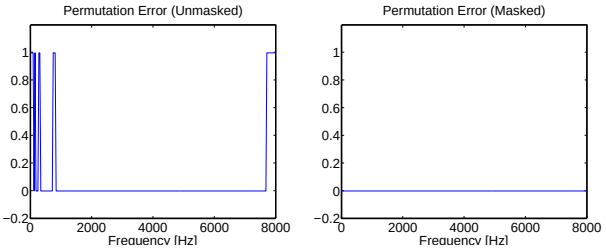
3.1.1. Weak Early Reflection Case

From the Fig.2(a), it can be seen that there are many vertical lines in the measured value of the cost function when unmasked FDICA is considered. These vertical lines show that it is necessary to exchange the output at those frequencies where the permutation problem exists.
From the Fig.2(b), it is clearly evident that the measured value of the cost function is almost unity for all the frequencies except for very low frequencies when the speech is perceptually masked.
Permutation error is defined as the case when the result of inter frequency coherency (IFC) differs from that of source output crosscorrelation (SOC) [3]. It is evident



(a) Unmasked (b) Masked

Figure 2: Measured Value of Cost Function for $k = 5$



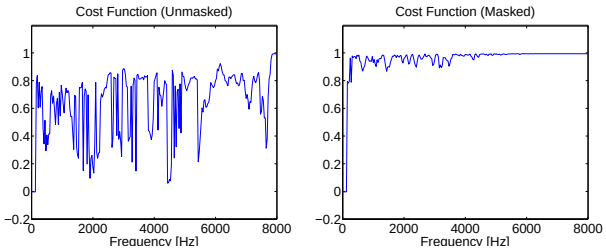
(a) Unmasked (b) Masked

Figure 3: Measured Value of Permutation Error for $k = 5$

from the Fig.3(a) that there are many vertical lines for frequencies below 2 kHz and a very few vertical lines for frequencies above 6 kHz in the measured permutation error when the perceptual masking is not considered.
It is clearly evident from the Fig.3(b) that the measured value of the permutation error is zero for all the frequencies when the speech is perceptually masked.

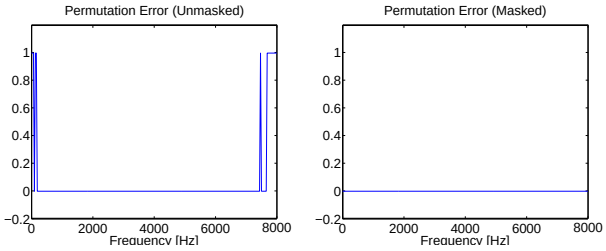
3.1.2. Strong Early Reflection Case

From the Fig.4(a), it can be seen that there are many vertical lines in the measured value of the cost function when unmasked FDICA is considered. These vertical lines show that it is necessary to exchange the output at those frequencies where the permutation problem exists.
From the Fig.4(b), it is clearly evident that the measured value of the cost function is almost unity for all the frequencies except for very low frequencies when the speech is perceptually masked.
It is evident from this Fig.5(a) that there are a very few vertical lines for frequencies below 1 kHz and a few vertical lines for frequencies above 6 kHz in the measured value of the permutation error when the perceptual auditory masking is not at all taken into account.
It is clearly evident from the Fig.5(b) that the measured



(a) Unmasked (b) Masked

Figure 4: Measured Value of Cost Function for $k = 5$



(a) Unmasked (b) Masked

Figure 5: Measured Value of Permutation Error for $k = 5$

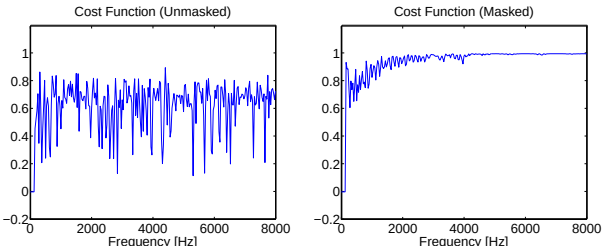
value of the permutation error is zero for all the frequencies when the speech is perceptually masked.

3.2. Experiment 2

The second experiment was chosen to test the algorithm's ability in real room recording condition. To do this, we used real room recorded speech signals (6 s at 16 kHz).
The permutation error cannot be computed in this real room recording case as the original sources are unknown. Cost function shown in Fig.6 is similar to that of previous experiment for both unmasked and masked systems.

4. CONCLUSIONS

A perceptually motivated FDICA scheme with subspace method, proposed in the paper, reduces the frequency components that are perceptually irrelevant by exploiting the masking properties of speech.
This system also reduces the computation complexity of similarity measure among spectral envelopes of separated signals for solving the permutation ambiguity.
Further, the crosstalk suppression ratio has been improved by 5 dB when perceptual masking is taken into account.



(a) Unmasked (b) Masked

Figure 6: Measured Value of Cost Function for $k = 5$

The measured permutation error is 7.8% and 6.6% for unmasked FDICA system under both weak and strong early reflection conditions respectively.
On the other hand, the permutation error is zero for perceptually masked FDICA system for both the cases of weak and strong and reflection conditions.

5. REFERENCES

[1] F. Asano et al, "Speech Enhancement Based on the Subsapce Method," *IEEE Trans. Speech, Audio Processing*, Vol. 8, pp. 497-507, Sept. 2000.

[2] N. Murata, S. Ikeda and A. Ziehe, "An Approach to BSS Based on Temporal Structure of Speech Signals," *Neurocomputing*, Vol. 41, pp. 1-24, Oct. 2001.

[3] F. Asano et al, "Combined Approach of Array Processing and ICA for Blind Separation of Acoustic Signals," *Proc. IEEE Trans. Speech and Audio Processing*, Vol. 11, No. 3, pp. 204-215, May 2003.

[4] Rammohana Reddy Guddeti and Bernard Mulgrew, "Perceptually Motivated Blind Source Separation of Convolutive Mixtures," *Proc.of the Int. Conf. IEEE ICASSP2005*, Philadelphia, PA, USA, March 2005.

[5] T. Painter and A. Spanias, "Perceptual Coding of Digital Audio," *Proc. of IEEE*, Vol. 88, No. 4, pp. 451-513, April 2000.

[6] A. Bell and T. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," *Proc. Neural Comput.*, Vol. 7, pp. 1129-1159, 1995.

[7] S. Amari, A. Cichocki and A. A. Yang, "A New Learning Algorithm for Blind Signal Separation," *Proc. NIPS'95*, pp. 752-763, 1996.

[8] P. Smaragdis, "Blind Separation of Convolved Mixtures in the Frequency Domain," *Proceedings of Neurocomputing*, Vol. 22, pp. 21-34, 1998.

[9] S. Ikeda and N. Murata, "A Method of ICA in Time-Frequency Domain," *Proc. ICA'99*, pp. 365-371, January 1999.