PERCEPTUALLY MOTIVATED BLIND SOURCE SEPARATION OF CONVOLUTIVE MIXTURES

Rammohana Reddy Guddeti and Bernard Mulgrew

Institute for Digital Communications
School of Engineering & Electronics
The University of Edinburgh
Edinburgh EH9 3JL U.K

ABSTRACT

A perceptually motivated method is proposed for solving the permutation ambiguity of frequency-domain independent component analysis when the mixing environment is noisy and reverberant. In this method, perceptually irrelevant frequencies are removed from the speech spectrum using block based perceptual masking (simultaneous frequency masking) and then independent component analysis is applied. After source separation in frequency domain, a physical property of the mixing matrix, i.e., the coherency in adjacent frequencies, is utilized to solve the permutation ambiguity. From the simulation results it appears that the perceptual masking avoids the permutation problem.

1. INTRODUCTION

The framework of blind source separation (BSS) based on independent component analysis (ICA) can be used to separate multiple signals without any previous knowledge of the sound sources and the mixing environment [1]. However, when applying to the cocktail party effect the performance of the BSS system is greatly reduced by the effect of the room reflections and ambient noise. Humans deal with this cocktail party effect very effectively by using only two ears (sensors). These perceptual masking techniques have been already exploited in successful development of MPEG audio coding standard which is the backbone of MP3 players.

In general, convolutive BSS methods can be classified into time domain ICA (TDICA) and frequency domain ICA (FDICA). TDICA has the disadvantage of being rather computational expensive due to computing many convolutions. The biggest obstacle in the FDICA is the permutation and scaling problem. For the scaling problem, the method proposed by Murata et al [2, 3], in which the separated output is filtered by the inverse of the separation filter.

For the permutation problem, Asano et al [4] have 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)). In this paper, a perceptually motivated FDICA approach for solving the permutation problem is proposed. This method utilizes the block based perceptual masking for the complete omission of a signal at the given frequency that is perceptually irrelevant.

This paper is organized as follows: In Section 2, an outline of the proposed perceptually motivated FDICA system is presented in order to solve the permutation problem. In Section 3, simulation results of experiments using both synthetic and real 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 the form of block diagram as shown in Fig.1.

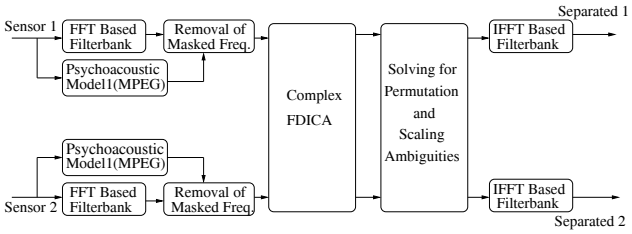


Fig. 1. Proposed Perceptually Motivated FDICA System

First, the short time Fourier transform (STFT) of the multichannel input signal, $\mathbf{x}(\omega, t)$, is obtained with an appropriate time shift and window function. Next, psychoacoustic model 1 (MPEG 1, layer I) [5] is used to determine the perceptual masking threshold for each segment of speech and thereby producing 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.

Then, the FDICA algorithm (complex Infomax with feedforward architecture [2,8,9]) is applied to the spectral components that are perceptually relevant for obtaining the separation filter. Next, the permutation and the scaling problem is solved by processing the output of the separation filter with the permutation and the scaling matrices. Finally, the filter matrices are transformed into the time domain and the input speech signal is processed with these filters.

2.1. Model of Signal

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

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

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

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

$\mathbf{A}(\omega)$ 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 $\mathbf{s} = [S_1(\omega,t), ..., S_D(\omega,t)]^T$.

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 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.

2.2.1. Power Spectrum

First, the sensor input, $\mathbf{x}(n)$, is segmented into frames of size of 512 samples using an appropriate time shift and Hann window function. A power spectral density (PSD), $P(k)$, for $\left(0 \le k \le \frac{N}{2}\right)$ is then obtained using a 512-point FFT as

$$P(k) = PN + 10\log_{10}\left|\sum_{n=0}^{N-1} w(n)x(n)e^{-j\frac{2kn\pi}{N}}\right|^2. \quad (3)$$

The power normalization term *PN*, fixed at 96 dB, is used to estimate the sound pressure level (SPL) conservatively from the input signal and $w(n)$ is Hann window function.

2.2.2. Global Masking Threshold

The absolute threshold of hearing is characterised by the amount of energy needed in a pure tone such that it can be detected by a listener in a noiseless environment. The quiet threshold is well approximated by

$$T_q(f) = \begin{array}{l} 3.64\left(\frac{f}{1000}\right)^{-0.8} \\ -6.5e^{0.6\left(\frac{f}{1000}-3.3\right)^2} \text{ (dB SPL)} \\ +10^{-3}\left(\frac{f}{1000}\right)^4 \end{array} \quad (4)$$

Simultaneous masking refers to a frequency domain phenomenon which has been observed within critical bands. Masking also occurs in the time domain. Sharp signal transients create pre- and post- masking regions in time during which a listener will not perceive signals beneath the elevated audibility thresholds produced by a masker. We didn't take into account temporal masking. This is due to the fact that our model is principally oriented to the speech signal that is stationary for a period shorter than 50 m sec.

Since masking refers to a psychoacoustic phenomenon, the masking threshold will be calculated in Barks. The Bark scale, in fact, refers to the critical bands of hearing. The conversion from frequency to Bark is given by

$$Bark(f) = \begin{array}{l} 13\arctan(.00076f) \\ +3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right] \end{array} \quad (5)$$

From the PSD of equation 3 we detect all the local maxima, then we replace any two maxima in a 0.5 Bark sliding window by the stronger of the two. Once the tone and noise maskers are calculated, a decimation process takes place before calculating the global masking threshold according to the following scheme:

$$i = \begin{cases} k & 1 \le k \le 48 \\ k + (k.\text{mod}2) & 49 \le k \le 96 \\ k + 3 - ((k-1) \bmod 4) & 97 \le k \le 232 \end{cases} \quad (6)$$

where *k* is the FFT index and *i* the decimation index. This process reduces the number of bins for the calculation of the global masking threshold, without loss of maskers. Having obtained a decimated set of tonal and noise maskers, individual tone and noise masking thresholds are computed next. Each individual threshold represents a masking contribution at frequency bin *i* due to the tone or noise masker located at bin *j*. Tonal masker thresholds, $T_{TM}(i,j)$ are expressed in (dB SPL) as

$$T_{TM}(i,j) = P_{TM}(j) - 0.275z(j) + SF(i,j) - 6.025 \quad (7)$$

where $P_{TM}(j)$ denotes the SPL of the tonal masker in frequency bin *j*, $z(j)$ denotes the Bark frequency of bin *j*, and the spread of masking from masker bin *j* to maskee bin *i*, $SF(i,j)$, is modeled by the expression in (dB SPL)

$$SF(i,j) = \tag{8}$$

$$\begin{cases} 17\Delta_z - 0.4P_{TM}(j) + 11, & -3 \le \Delta_z < -1 \\ (0.4P_{TM}(j) + 6)\,\Delta_z, & -1 \le \Delta_z < 0 \\ -17\Delta_z, & 0 \le \Delta_z < 1 \\ (0.15P_{TM}(j) - 17)\,\Delta_z - 0.15P_{TM}(j), & 1 \le \Delta_z < 8 \end{cases}$$

Individual noise masker thresholds (dB SPL) are given by

$$T_{NM}(i,j) = P_{NM}(j) - 0.175z(j) + SF(i,j) - 2.025 \tag{9}$$

where $P_{NM}(j)$ denotes the SPL of the noise masker in frequency bin *j*, $z(j)$ denotes the Bark frequency of bin *j*, and $SF(i,j)$ is obtained by replacing $P_{TM}(j)$ with $P_{NM}(j)$ everywhere in equation 8.

The global masking threshold, $T_g(i)$, is therefore obtained in dB by computing the sum

$$T_g(i) = 10\log_{10}\left(\begin{array}{l} 10^{0.1T_q(i)} + \sum_{l=1}^{L} 10^{0.1T_{TM}(i,l)} \\ + \sum_{m=1}^{M} 10^{0.1T_{NM}(i,m)} \end{array}\right) \tag{10}$$

where $T_q(i)$ is the absolute hearing threshold for frequency bin *i*, $T_{TM}(i,l)$ and $T_{NM}(i,m)$ are the individual masking thresholds and *L* and *M* are the number of tonal and noise maskers , respectively.

2.3. FDICA Algorithm

Whenever the perceptually masked input speech $\mathbf{x}(\omega,t)$ in one of the channels contains no values, the PCA filter matrix $\mathbf{W}(\omega)$ is singular, resulting in rank deficiency. Without loss of generality we have assumed identity matrix of order *M* as the rank of $\mathbf{W}(\omega)$ to avoid this problem. Then, the Infomax algorithm is applied to the output of the PCA filter, $\mathbf{y}(\omega,t)$ to obtain the ICA filter $\mathbf{U}(\omega)$. The separation filter $\mathbf{B}(\omega)$ is expressed as the product of $\mathbf{W}(\omega)$ and $\mathbf{U}(\omega)$. 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

$$\mathbf{U}(\omega,t+1) = \mathbf{U}(\omega,t) + \eta[\mathbf{I} - \varphi(\mathbf{z}(\omega,t))\mathbf{z}^H(\omega,t)]\mathbf{U}(\omega,t) \tag{11}$$

where the score function for $\varphi(\mathbf{z})$ is defined as

$$\varphi(\mathbf{z}) = [\varphi(z_1), \cdots, \varphi(z_d), \cdots, \varphi(z_D)]^T \tag{12}$$

$$\varphi(z_d) = 2\tanh(\Re(z_d)) + 2j\tanh(\Im(z_d)) \tag{13}$$

The symbol z_d is the *d*th element of the vector $\mathbf{z}(\omega,t)$. The matrix $\mathbf{I}$ is an identity matrix. The symbol $.^H$ denotes the Hermitian transpose. The constant η (.0001) is termed the learning rate. Here also we have avoided ICA filtering when the input of ICA filter in one of masked channels is zero in order to overcome the rank deficiency of ICA filter matrix.

The scaling problem can be solved by filtering individual output of the separation filter $\mathbf{B}(\omega)$ by its inverse [3]. The permutation problem can be solved by minimizing the sum of the angles $\{\theta_1, \cdots, \theta_D\}$ between the location vectors in the adjacent frequencies. The cosine of the angle θ_n between the two vectors, $\bar{\mathbf{a}}_n(\omega)$ and $\bar{\mathbf{a}}_n(\omega_0)$, of estimated mixing matrix is defined as [4]

$$\cos\theta_n = \frac{\bar{\mathbf{a}}_n^H(\omega)\bar{\mathbf{a}}_n(\omega_0)}{\|\bar{\mathbf{a}}_n(\omega)\|.\|\bar{\mathbf{a}}_n^H(\omega_0\|} \tag{14}$$

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

$$F(\mathbf{P}) = \frac{1}{D}\sum_{n=1}^{D}\cos\theta_n \tag{15}$$

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 [4]

$$C(k) = \max_{\mathbf{P}\in\Omega}[F(\mathbf{P},k)] - \max_{\mathbf{P}\in\Omega'}[F(\mathbf{P},k)] \tag{16}$$

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 [4]

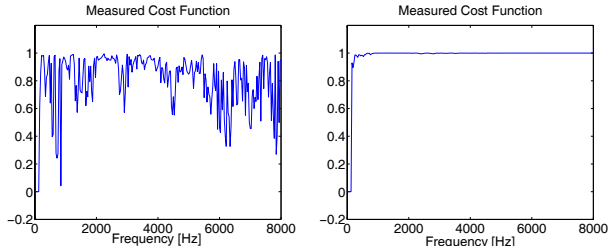
$$\hat{\mathbf{P}} = \arg\max_{\mathbf{P}}[F(\mathbf{P},\hat{k})] \tag{17}$$

The main contribution of this perceptual filtering is not only the reduction of frequencies that are processed by ICA, but also the reduction of frequencies where the similarity has to be checked for solving the permutation problem.

3. SIMULATION RESULTS

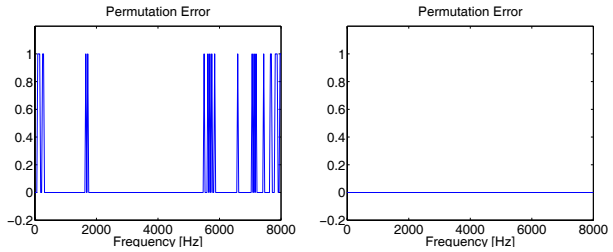
3.1. Experiment 1

In the first experiment, we created a synthetic convolutive mixture of two speech sources (7 s at 16 kHz) and we used Westner's [6] room acoustic data with reverberation time of 0.5 sec to simulate reverberant condition. 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 IFC differs from that of



(a) Unmasked (b) Masked

Fig. 2. Measured Value of Cost Function for $k = 5$



(a) Unmasked (b) Masked

Fig. 3. Measured Value of Permutation Error for $k = 5$

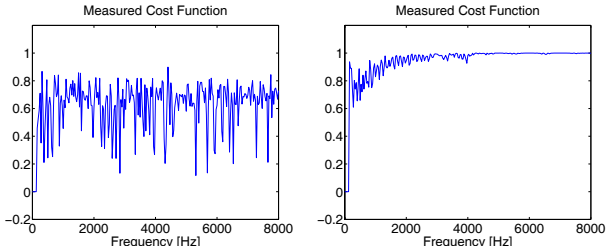
source output crosscorrelation (SOC) [4]. It is evident from this Fig.3(a) that there are many verticle lines in the measured permutation error when the speech is unmasked. It is clearly evident from the Fig.3(b) that the permutation error is zero for all the frequencies when the speech is masked.

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. Real room recording results shown in Fig.4 are similar to that of previous experiment from the cost function point of view.

4. CONCLUSIONS

Incorporating the proposed perceptual solution for the permuation problem in the FDICA system produced good separation results in terms of the measured values of the cost function and the permutation error.



(a) Unmasked (b) Masked

Fig. 4. Measured Value of Cost Function for $k = 5$

5. REFERENCES

[1] A. Hyvarinen, J. Karhunen and E. Oja, "Independent Component Analysis,"*Wiley Inter-Science*, 2001.

[2] N. Murrata and S. Ikeda, "A Method of ICA in Time-Frequency Domain," *Proc. ICA'99*, pp. 365-370, Jan. 1999.

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

[4] F. Asano, S. Ikeda, M. Ogawa, H. Asoh and N. Kitawaki, "Combined Approach of Array Processing and ICA for Blind Separation of Acoustic Signals," *IEEE Trans. Speech and Audio Processing*, 11 (3), pp. 204-215, May 2003.

[5] T. Painter and A. Spanias, "A Review of Algorithms for perceptual Coding of Digital audio Signals," *Proc. DSP1997*, vol. 1, pp. 179-208, July 1997.

[6] A. G. Westner, "Object Based Audio Capture: Separating Acoustic Sounds," *M.S. Thesis, MIT Media Laboratory*, 1998.

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

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

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