A frequency bin-wise nonlinear masking algorithm in convolutive mixtures for speech segregation

Tai-Shih Chi, Ching-Wen Huang, and Wen-Sheng Chou
Department of Electrical Engineering, National Chiao Tung University, Hsinchu 300, Taiwan
tschi@mail.nctu.edu.tw, chingwen.cm98g@nctu.edu.tw, s900925@gmail.com

Abstract: A frequency bin-wise nonlinear masking algorithm is proposed in the spectrogram domain for speech segregation in convolutive mixtures. The contributive weight from each speech source to a time-frequency unit of the mixture spectrogram is estimated by a nonlinear function based on location cues. For each sound source, a non-binary mask is formed from the estimated weights and is multiplied to the mixture spectrogram to extract the sound. Head-related transfer functions (HRTFs) are used to simulate convolutive sound mixtures perceived by listeners. Simulation results show our proposed method outperforms convolutive independent component analysis and degenerate unmixing and estimation technique methods in almost all test conditions.

PACS numbers: 43.60.Gk, 43.60.Hj, 43.60.Dh [JC]
Date Received: February 01, 2012 **Date Accepted:** February 27, 2012

1. Introduction

Segregating sound streams from a sound mixture is a typical task for people in daily lives but is very challenging for machines. Solving this "cocktail party problem" has drawn a lot of interests and is referred to as the blind source separation (BSS) research field. The "blind" means that only the mixed signals are available while the source signals and the mixing process are all unknown. Based on different assumptions, many algorithms have been proposed for the BSS problem. For instance, independent component analysis (ICA) algorithms assume that source signals are mutually independent and aim to maximize the independence between separated signals.[1–5] The well-known FastICA (Ref. 2) and InfomaxICA (Ref. 3) algorithms were developed using kurtosis and mutual information as measurements of independence. These algorithms were originally developed for the time-domain instantaneous mixing condition where sound mixtures are linearly combined from individual sound signals. To deal with the convolutive mixing condition, conventional ICA algorithms were modified to transform signals from the time domain to the frequency domain to convert the time-domain convolutive mixtures to the frequency bin-wise instantaneous mixtures. Such an approach is referred to as the convolutive ICA (cICA). Although the idea is intuitive, certain problems such as permutation and amplitude scaling need to be carefully addressed.[4,5]

Another approach to solve the BSS problem is the sparse component analysis (SCA),[6–8] which adopts the sparse assumption of speech signals to segregate sound streams. A signal is called sparse when most of its values are zero or close to zero. The property that the probability of two or more sparse sources being active simultaneously is very low leads to a favorable condition for speech segregation. SCA algorithms developed for the instantaneous mixing condition generally estimate the mixing matrix by identifying the principal directions in the scatter plot. To deal with the convolutive mixing condition, SCA algorithms also transform sound mixtures into the frequency domain using short-time Fourier transform (STFT). Based on the sparse assumption of spectrograms, SCA algorithms estimate the source index of each time-frequency (T-F) unit and generate a binary mask to extract each sound stream. For example, the

degenerate unmixing and estimation technique (DUET) estimates source indexes using magnitude and phase differences between received spectrograms of sound mixtures.[8] However, the binary mask would inevitably generate musical noise in separated sound streams.

In this paper, we propose a frequency bin-wise nonlinear masking (NM) algorithm to generate a *non-binary* mask to segregate sound streams with less musical noise. Simulation results show our proposed algorithm outperforms cICA and DUET in almost all test conditions. The rest of this paper is organized as follows. Backgrounds of the convolutive mixing model and the sparse assumption are given in Sec. 2. Our proposed algorithm is described in detail in Sec. 3, and experimental results are given in Sec. 4. We end in Sec. 5 with a conclusion and future works.

2. Mixing model and sparse assumption

Consider N source signals $s_1,\ldots, s_N$ and M received mixtures $x_1,\ldots, x_M$ in the time domain. The corresponding vector forms are written as an Nx1 vector $\mathbf{s}=[s_1,\ldots, s_N]^{\mathrm{T}}$ and a Mx1 vector $\mathbf{x}=[x_1,\ldots, x_M]^{\mathrm{T}}$. The instantaneous mixing model can be written as $\mathbf{x}(t)=\mathbf{A}\mathbf{s}(t)$, where $\mathbf{A}$ is the MxN mixing matrix. The goal of conventional ICA algorithms is to find a de-mixing matrix $\mathbf{W}$ such that $\mathbf{W}\approx\mathbf{A}^{-1}$. In that case, the separated signal $\mathbf{y}$(t) can be written as $\mathbf{y}(t)=\mathbf{W}\mathbf{x}(t)=\hat{\mathbf{s}}(t)$.

2.1 Convolutive mixing

In real world environments, however, speech signals are mixed in a convolutive manner due to reverberations. The convolutive mixing model is written as

$$x_i(t)=\sum_k b_{ik}(t)*s_k(t)=\sum_k\sum_{\delta=0}^{\infty} b_{ik}(\delta)s_k(t-\delta) \tag{1}$$

where $*$ is the time domain convolution and $b_{ik}(t)$ is the room impulse response measured from source k to microphone i. Equation (1) can be transformed to the frequency domain using STFT so that convolutive mixtures become instantaneous mixtures in each frequency bin ω_j (Refs. 4, 5):

$$\mathbf{X}(\omega_j,t)=\mathbf{B}(\omega_j)\mathbf{S}(\omega_j,t) \tag{2}$$

where the capital letters stand for the Fourier domain representation of the signals in lower case letters. Therefore, ICA algorithms for instantaneous mixing can be applied to each frequency bin of convolutive mixtures. At the end, spectrograms of separated signals can be obtained by assembling de-mixing results from all frequency bins.

2.2 Sparse assumption

If two time-domain signals are sparse, they are seldom active at the same time, and their scatter plot would produce two orthogonal lines as shown in the left panel of Fig. 1. After applying the instantaneous mixing matrix **A,** the principal directions in the scatter plot will change accordingly. Therefore identifying the principal directions in the scatter plot is a typical SCA approach for separating instantaneous mixtures.

Speech signals are also assumed sparse in transformed domains by some SCA algorithms. For example, the DUET assumes spectrograms of two speech signals are W-disjoint orthogonal,[8] i.e., for a given window function $w(t)$, the supports of the windowed Fourier transforms of $s_1(t)$ and $s_2(t)$ are disjoint. The STFT of $s_i(t)$ is defined as

$$S_i(\omega,\tau)=\sum_t s_i(t)w(t-\tau)e^{-j\omega t} \tag{3}$$

where τ is the frame index in the time domain. The W-disjoint orthogonality can be expressed as

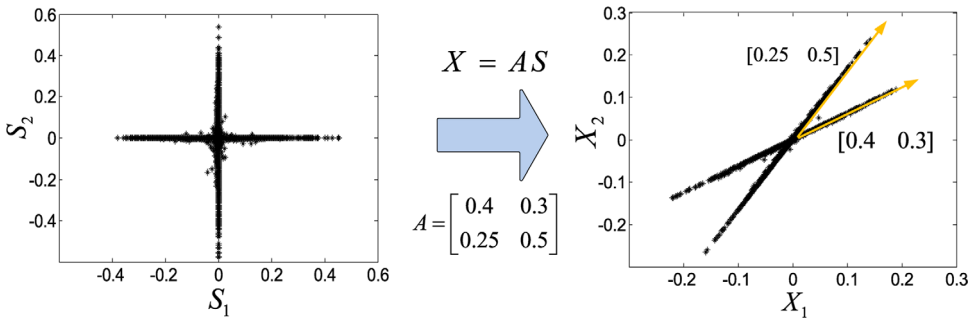

FIG. 1. (Color online) Scatter plots of sparse speech mixtures before (left panel) and after (right panel) the time-domain instantaneous mixing. Obviously, principal directions encode the instantaneous mixing matrix.

$$S_1(\omega,\tau)S_2(\omega,\tau)=0, \quad \forall\omega,\tau, \tag{4}$$

which says T-F units of spectrograms of $s_1(t)$ and $s_2(t)$ are not active simultaneously.

3. Proposed nonlinear masking algorithm

Without loss of generality, we describe our algorithm in segregating two sources from two mixtures in this section. The algorithm can be extended to segregate multiple sources from two mixtures. First, the two mixtures are transformed into spectrograms $X_1(\omega,t)$ and $X_2(\omega,t)$ using the 512-point STFT. The two magnitude spectrograms at each frequency bin ω_j are then treated as a new set of mixtures, i.e., $\mathbf{Z}(\omega_j,t)=\left[\left|X_1(\omega_j,t)\right|,\left|X_2(\omega_j,t)\right|\right]$. As in cICA algorithm, the $\left|X_1(\omega_j,t)\right|$ and $\left|X_2(\omega_j,t)\right|$ at frequency bin ω_j are now instantaneous mixtures. Moreover, based on the sparse assumption of mixed spectrograms, there shall be only two principal directions in the scatter plot of $\mathbf{Z}(\omega_j,t)$ for the two sources. The nonlinear projection column masking (NPCM) technique[9] is used to detect principal directions in the scatter plot.

3.1 Principal direction detection using NPCM

For the frequency bin ω_j, the sample point at time instant t in the scatter plot can be represented by the vector $\mathbf{Z}_t$, which is short for $\mathbf{Z}(\omega_j, t)$. If the principal direction is $\boldsymbol{\alpha}$, the projection length of the sample point $\mathbf{Z}_t$ onto the direction vector $\boldsymbol{\alpha}$ can be written as $y_t=\|\mathbf{Z}_t\|\left|\cos\measuredangle(\boldsymbol{\alpha},\mathbf{Z}_t)\right|$, where $\measuredangle(\boldsymbol{\alpha},\mathbf{Z}_t)$ is the angle between $\boldsymbol{\alpha}$ and $\mathbf{Z}_t$. Principal component analysis (PCA) is then used to search a vector from all possible $\boldsymbol{\alpha}$ such that $\mathrm{J}(\boldsymbol{\alpha})=\mathrm{E}[y_t^2]$ is maximized. However, the chosen $\boldsymbol{\alpha}$ using PCA might not be close to the true principal direction due to interferences from far-away samples. To deal with this problem, the NPCM technique was proposed to approximate true principal directions for instantaneously mixtures by masking any far-away samples.[9]

The NPCM technique modifies the projection length y_t as $y_t=\|\mathbf{Z}_t\|f(\cos\measuredangle(\boldsymbol{\alpha},\mathbf{Z}_t))$, where $f(\cdot)$ is a nonlinear decaying function to mask samples far away from the vector $\boldsymbol{\alpha}$. The exponential function $f(r)=\exp(-\rho r^2)$ with $\rho=10^7$ is used in our algorithm. The following criterion is then considered to find the principal direction $\boldsymbol{\alpha}$.

$$\max_a J(\boldsymbol{\alpha})=\sum_t \|\mathbf{Z}_t\|f(\cos\measuredangle(\boldsymbol{\alpha},Z_t)). \tag{5}$$

The left panel of Fig. 2 shows an example of the scatter plot of $\mathbf{Z}_t$ and the estimated principal directions using $\rho=10^7$. The right panel shows the trajectory of $\mathrm{J}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}=[\cos\varphi,\ \sin\varphi]^{\mathrm{T}}$ and $\varphi\in[0,\pi/2]$ is the angle between the principal direction and the x-axis of the scatter plot. One can observe that the two principal directions correspond to the two local maxima of the trajectory of $\mathrm{J}(\boldsymbol{\alpha})$.

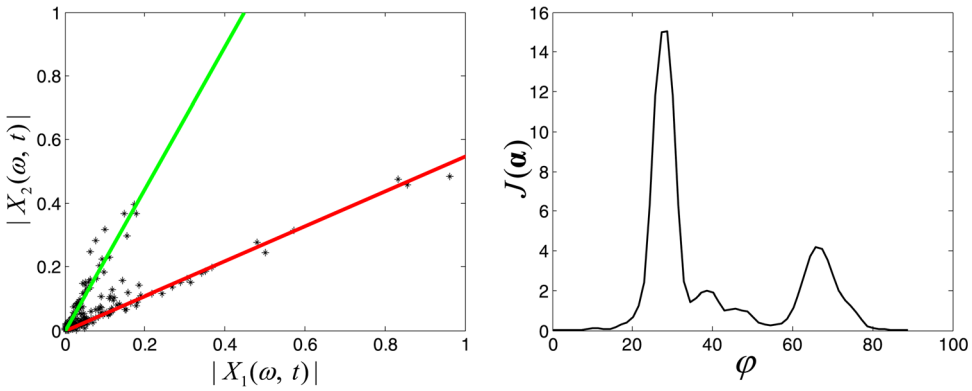

FIG. 2. (Color online) Detecting principal directions in a certain frequency bin of magnitude spectrograms of sample convolutive speech mixtures. Left panel shows two principal directions superimposed on the scatter plot of magnitude spectrograms with $\rho = 10^7$. Right panel shows the trajectory of J($\boldsymbol{\alpha}$). The two principal directions correspond to the two local maxima of the trajectory of J($\boldsymbol{\alpha}$).

3.2 Nonlinear masking

According to the W-disjoint orthogonality property, only one source is assumed active at each T-F unit of spectrograms. A reasonable approach would assume T-F units close to a principal direction in the scatter plot belong to that specific sound source. In other words, to extract a sound stream, we only need to synthesize T-F units close to a particular principal direction vector $\boldsymbol{\alpha}$ while mask other T-F units. However, the spectrogram built using the binary decision on "closeness" would inevitably produce musical noise as in all other binary mask algorithms, such as the DUET.

To diminish musical noise, a *non-binary* mask is generated based on measures of the "closeness" of T-F units to a particular principal direction by a nonlinear function. The masking value of the T-F unit $\mathbf{Z}_t$ to the source k at the frequency ω_j is formulated as:

$$u_k(\omega_j, t) = \exp(-\kappa \cos^2 \measuredangle(\boldsymbol{\alpha}_k, \mathbf{Z}_t)) \tag{6}$$

where κ serves as a masking constant, and the larger the κ, the faster the exponential function decays. The non-binary masking value u_k becomes larger or smaller when $\mathbf{Z}_t$ is closer to or further away from $\boldsymbol{\alpha}_k$. After calculating the masking function $u_k(\omega_j, t)$, we estimate the magnitude spectrogram of source k ($k = 1, 2$) at frequency ω_j by

$$\hat{S}_k(\omega_j, t) = u_k(\omega_j, t) \odot Z_{IP}^{(k)}(\omega_j, t) \tag{7}$$

where $\odot$ represents the point-wise multiplication between matrices and $Z_{IP}^{(k)}(\omega, t)$ is the magnitude spectrogram of source k received by the ipsilateral ear. The $Z_{IP}^{(k)}(\omega, t)$ is determined by

$$Z_{IP}^{(k)}(\omega, t) = \begin{cases} |X_1(\omega, t)| & \text{if } \varphi_k \le \pi/4 \\ |X_2(\omega, t)| & \text{otherwise.} \end{cases} \tag{8}$$

After assembling the target magnitude spectrogram from all frequency bins, the overlap-and-add method is applied to restore the magnitude spectrogram $\hat{S}_k(\omega, t)$ combined with the original phase spectrogram received by the ipsilateral ear to sound $\hat{s}_k(t)$.

Figure 3 shows an original clean spectrogram and non-binary masks generated for a sample convolutive mixing condition using $\kappa = 10^2$, 10^4, and 10^6, respectively. One can observe that the non-binary mask preserves fewer T-F units of the target signal when κ is larger and the mask keeps more residue noise when κ is smaller. In our

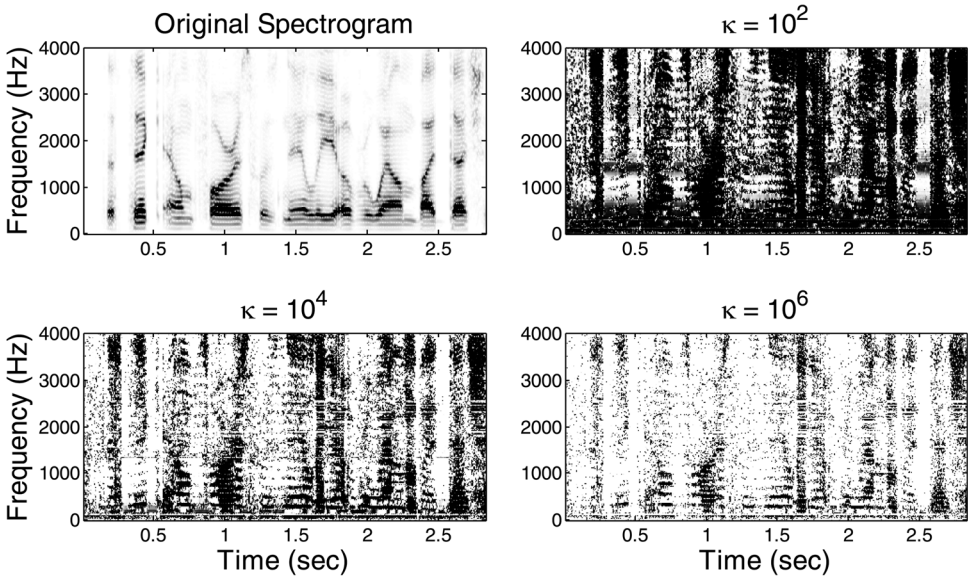

FIG. 3. Original spectrogram of a sample speech and non-binary masks generated by the proposed algorithm for different κ.

algorithm, the parameter κ is set to 10^4 to have a proper balance between target speech distortion and residue noise in segregated speech streams. Discussions about the choice of κ are given in Sec. 4.

3.3 Permutation problem

The cICA or other frequency bin-wise methods need to deal with the permutation problem by correctly assigning separated signals from each frequency bin to the target source. Conventional approaches, such as considering direction-of-arrival (DOA) or correlation across frequency bins, require extra computations and always show a trade-off between robustness and correctness.[10] In our proposed algorithm, the permutation problem is fairly easy to deal with. Although the direction vectors $\boldsymbol{\alpha}$ corresponding to a particular speech source at all frequency bins may not be exactly the same, they are actually quite similar across frequencies because they encode the spatial information of that source. Therefore we only need to group T-F units associated with close principal directions across frequency bins as from a specific speech source.

4. Experimental evaluations

To model sound mixtures perceived by listeners, we used HRTFs to simulate the convolutive mixtures of speech signals. The locations at the left, front, and right sides of the head were assigned $-90°$, $0°$, and $90°$, respectively, in our simulations. In following discussions, the (θ_1, θ_2) in degree is used to represent the locations of source 1 and source 2. Rather than using the signal-to-interference ratio (SIR), we used the signal-to-interference distortion ratio (SDR) as the performance measure. The SDR takes both speech quality and suppression of interference into consideration, while the SIR only considers suppression of interference.[11] The SDR was defined as

$$SDR = 10\log_{10}\frac{\left\|s_{\text{target}}\right\|^2}{\left\|e_{\text{interf}} + e_{\text{artif}}\right\|^2} \tag{9}$$

where $s_{\text{target}}(t)$ is the estimated target source, $e_{\text{interf}}(t)$ is the interference from unwanted sources, and $e_{\text{artif}}(t)$ indicates artificial noise (such as musical noise) induced by

TABLE 1. Average SDR (in dB) by different algorithms at various location (θ_1, θ_2) settings.

(θ_1, θ_2)	(−5°, 25°)	(40°, 10°)	(−60°, −40°)	(30°, −30°)	(15°, 50°)
DUET	0.557	−1.649	−6.365	−6.362	−1.286
cICA	3.996	3.367	−2.570	7.824	0.384
NM	7.287	8.065	−0.504	6.877	7.552

separation algorithms. Both s(t) and e(t) were decomposed from the segregated signal $\hat{s}(t)$ using the BSS_EVAL Toolbox.[12]

Four sentences (by 2 male and 2 female speakers) and another eight sentences (by 4 male and 4 female speakers) were selected from the TIMIT database as source 1 and source 2, respectively. All sentences were extracted from different speakers of the database. The average SDR of two segregated signals from all 32 mixing pairs at five location settings are given in Table 1 for DUET, cICA, and the proposed NM algorithm. We tested parameter κ from 10^1, 10^2 to 10^9, and the overall averaged SDR from all locations raised from 1 dB ($\kappa = 10^1$), topped around 6 dB ($\kappa = 10^4$) and dropped to −3 dB ($\kappa = 10^9$). Therefore, κ was set to 10^4 in our simulations.

The results in Table 1 show the proposed NM method outperforms cICA and DUET by a large margin in almost all test conditions except being 1 dB less than cICA in the (30°, −30°) condition. It is worth noting that the second peak of magnitude and phase histograms in DUET analysis is weaker than the first peak such that more artificial noise emerges in the second segregated speech stream and damages the average SDR score. On the other hand, our proposed method with non-binary masks achieves a better balance between speech quality and interference suppression for all segregated sounds.

5. Conclusion

A frequency bin-wise nonlinear masking algorithm is proposed for speech segregation in convolutive mixtures in this paper. Unlike the DUET, the proposed algorithm generates non-binary masks based on estimated "closeness" in the scatter plot to reduce musical noise in segregated speech. Simulation results show SDR scores of the proposed algorithm outperform scores of cICA and DUET in almost all test conditions. However, our algorithm strongly depends on principal directions detected from frequency bin-wise scatter plots. If too many sound sources are present or sound sources are too close to each other, our algorithm would perform poorly due to the wrongly estimated principal directions such as in the (−60°, −40°) test condition. However, these conditions are also very difficult for other algorithms.

Our algorithm only utilizes magnitude spectrograms for speech segregation while the DUET considers both magnitude and phase information of spectrograms. To enhance the spatial resolution of our algorithm, phase cues will be considered in the future.

Acknowledgments

This work is supported by the National Science Council, Taiwan under Grant No. NSC 100-2220-E-009-004.

References and links

[1]J. V. Stone, *Independent Component Analysis: A Tutorial Introduction* (MIT Press, Cambridge, MA, 2004).

[2]A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," Neural Comput. **9**, 1483–1492 (1997).

[3]A. J. Bell and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," Neural Comput. **7**, 1129–1159 (1995).

[4]P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," Neurocomputing **22**, 21–34 (1998).
[5]S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," in *Proceedings of ICA* (1999), pp. 365–371.
[6]P. D. O'Grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," Int. J. Imaging Syst. Technol. **15**(1), 18–33 (2005).
[7]M. Zibulevsky, P. Kisilev, Y. Y. Zeevi, and B. A. Pearlmutter, "Blind source separation via multinode sparse representation," Adv. Neural Inf. Process. Syst. **14**, 1049–1056 (2002).
[8]O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process. **52**(7), 1830–1847 (2004).
[9]Z. Guoxu, Y. Zuyuan, X. Shengli, and Y. Jun-Mei, "Mixing matrix estimation from sparse mixtures with unknown number of sources," IEEE Trans. Neural Networks **22**, 211–221 (2011).
[10]S. Araki, R. Mukai, and S. Makino, "The fundamental limitation of frequency-domain blind source separation for convolutive mixtures of speech," IEEE Trans. Speech Audio Process, **11**, 109–116 (2003).
[11]E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," IEEE Trans. Audio, Speech, Language Process. **14**, 1462–1469 (2006).
[12]C. Fevotte, R. Gribonval, and E. Vincent, "BSS EVAL Toolbox User Guide," IRISA Technical Report 1706, Rennes, France, 2005, http://www.irisa.fr/metiss/bsseval/.