A MULTISTAGE APPROACH FOR BLIND SEPARATION OF CONVOLUTIVE SPEECH MIXTURES

Tariqullah Jan†*, Wenwu Wang*†*, and DeLiang Wang*‡ *

† Centre for Vision, Speech and Signal Processing, University of Surrey, UK
Email: {t.jan, w.wang}@surrey.ac.uk
‡ Department of Computer Science and Engineering & Centre for Cognitive Science,
The Ohio State University, Columbus, USA
Email: dwang@cse.ohio-state.edu

ABSTRACT

In this paper, we propose a novel algorithm for the separation of convolutive speech mixtures using two-microphone recordings, based on the combination of independent component analysis (ICA) and ideal binary mask (IBM), together with a post-filtering process in the cepstral domain. Essentially, the proposed algorithm consists of three steps. First, a constrained convolutive ICA algorithm is applied to separate the source signals from two-microphone recordings. In the second step, we estimate the IBM by comparing the energy of corresponding time-frequency (T-F) units from the separated sources obtained with the convolutive ICA algorithm. The last step is to reduce musical noise caused typically by T-F masking using cepstral smoothing. The performance of the proposed approach is evaluated based on both reverberant mixtures generated using a simulated room model and real recordings. The proposed algorithm offers considerably higher efficiency, together with improved speech quality while producing similar separation performance as compared with a recent approach.

Index Terms— Independent component analysis (ICA), ideal binary mask (IBM), estimated binary mask, cepstral smoothing, musical noise

1. INTRODUCTION

Human listeners show remarkable ability to segregate target speech from complex auditory mixtures, such as in a cocktail party environment. However, it remains extremely challenging for machines to replicate even part of such functionalities. This problem has been studied for decades. A recent technique is to address this problem under the blind source separation (BSS) framework where the mixing process is described as a linear convolutive model, and independent component analysis (ICA) can then be applied to separate the convolutive mixtures either in the time-domain [4] or in the transform domain [1, 3, 5]. Nevertheless, the separation performance of many developed algorithms is still limited, and leaves a large room for improvement. This is especially true when dealing with reverberant and noisy mixtures.

A recent technique, called IBM, originated from computational auditory scene analysis (CASA) [2], has shown promising properties in suppressing interference and improving quality of target speech. IBM is usually obtained by comparing the T-F representations of target speech and background interference, with 1 assigned to a T-F unit where the target energy is stronger than the interference energy and 0 otherwise. However, without the clean target speech and interfering sound, it is a difficult task to directly estimate an accurate IBM from the mixtures only.

In this paper we propose to obtain the target and background sounds from the mixtures using a constrained convolutive ICA algorithm [1], whose outputs are then used to estimate the IBM. This method can effectively address the aforementioned problems associated with the individual methods. However, errors introduced during the estimation of the IBM may give rise to isolated T-F units and hence result in fluctuating artifacts, i.e., the so-called musical noise [6]. To overcome this problem, the spectral smoothing in the cepstral domain is applied to the estimated binary mask. Different levels of smoothing are applied to the binary mask across different frequencies which are determined by pitch information estimated directly from the segregated signals, in contrast to the method in [6]. The following sections in this paper present our proposed multistage approach, experimental evaluation results, conclusions and future work.

2. A MULTISTAGE APPROACH

2.1. BSS of Convolutive Mixtures in the Frequency Domain

In a cocktail party environment, *N* speech signals are recorded by *M* microphones, described mathematically by,

$$x_j(n) = \sum_{i=1}^{N} \sum_{p=1}^{P} h_{ji}(p) s_i(n-p+1) \qquad (j = 1, ..., M) \quad (1)$$

*We are very grateful to M. S. Pedersen for his matlab code and help in this work. Part of the work was conducted while W. Wang was visiting OSU. D. L. Wang was supported in part by an AFOSR grant (FA9550-08-1-0155) and an NSF grant (IIs-0534707). Thanks to the NWFP UET Peshawar for funding T. Jan.

where s_i and x_j are the source and mixture signals respectively, h_{ji} is a *P*-point room impulse response [9]. Also we consider a two input two output system (TITO) system, i.e., *N*=*M*=2. The BSS problem for convolutive mixtures in the time domain is converted to multiple instantaneous problems in the frequency domain [1, 3] with equation (2). By applying short time Fourier transform (STFT) to equation (1), and using matrix notations we get

$$\mathbf{X}(k,n) = \mathbf{H}(k)\mathbf{S}(k,n) \tag{2}$$

where *k* represents the frequency index and *n* is the discrete time index. The mixing matrix $\mathbf{H}(k)$ is assumed to be invertible and time invariant.

To find the sources, we can effectively apply an unmixing filter $\mathbf{W}(k)$ to the mixtures.

$$\mathbf{Y}(k,n) = \mathbf{W}(k)\mathbf{X}(k,n) \tag{3}$$

where $\mathbf{Y}(k,n)$ represents the estimated source signals. $\mathbf{W}(k)$ is obtained when the estimated sources become mutually independent. Many algorithms have been developed for this purpose [3, 4, 5]. In this work we use a constrained convolutive ICA approach in [1] for the separation in this stage. Similar to many existing ICA approaches, e.g. [5], the separation performance of [1], especially the quality of the separated speech is still limited by the amount of interference. The performance steadily degrades with the increment of the reverberation time (RT). To improve the quality of the separated speech signals, we further apply the IBM technique from the CASA domain.

2.2. Combining Convolutive ICA and Binary Masking for the Segregation of Speech Signals

Applying an inverse Fourier transform, $\mathbf{Y}(k,n)$ can be converted back to the time domain denoted as

$$\mathrm{y}(n) = [\mathrm{y}_1(n) \quad \mathrm{y}_2(n)]^T \tag{4}$$

Scaling is applied to the $\mathrm{y}_1(n)$ and $\mathrm{y}_2(n)$ in order to get the normalized outputs $\tilde{\mathrm{y}}_1(n)$ and $\tilde{\mathrm{y}}_2(n)$. After this we transform the two normalized outputs into the T-F domain using STFT,

$$\tilde{\mathrm{y}}_1(n) \rightarrow \tilde{\mathbf{Y}}_1(k,n) \tag{5}$$

$$\tilde{\mathrm{y}}_2(n) \rightarrow \tilde{\mathbf{Y}}_2(k,n) \tag{6}$$

By comparing the energy of each T-F unit of the above two spectrograms, the two binary masks are estimated as,

$$\mathbf{M}_1^f(k,n) = \begin{cases} 1 & \text{if } \mid \tilde{\mathbf{Y}}_1(k,n) \mid > \tau \mid \tilde{\mathbf{Y}}_2(k,n) \mid, \\ 0 & \text{otherwise} \qquad \forall k,n. \end{cases} \tag{7}$$

$$\mathbf{M}_2^f(k,n) = \begin{cases} 1 & \text{if } \mid \tilde{\mathbf{Y}}_2(k,n) \mid > \tau \mid \tilde{\mathbf{Y}}_1(k,n) \mid, \\ 0 & \text{otherwise} \qquad \forall k,n. \end{cases} \tag{8}$$

where τ is a threshhold for controlling the sparseness of the mask, and τ=1 has been used in our experiment. The masks are applied to the T-F representation of the original two-microphone recordings in order to recover the source signals, as follows

$$\mathbf{Y}_i^f(k,n) = \mathbf{M}_i^f(k,n).\mathbf{X}_i(k,n) \qquad i = 1,...,N \tag{9}$$

In the next step source signals are recovered in the time domain using inverse STFT.

$$\mathbf{Y}_i^f(k,n) \rightarrow \mathrm{y}_i^t(n) \qquad i = 1,...,N \tag{10}$$

As observed in our experiments, the estimated IBM considerably improves the separation performance by suppressing the interference to a much lower level, leading to the separated speech signals with considerably improved quality over the outputs obtained in the section 2.1. However, a typical problem with the binary T-F masking is the introduction of the errors in the estimation of the masks causing fluctuating musical noise [6]. To mitigate this problem, we employ a cepstral smoothing technique [6] as detailed in the next section.

2.3. Cepstral Smoothing of the Binary Mask

The basic idea is to tranform the estimated IBM into the cepstral domain, and after the different smooothing levels have been applied to the transformed mask, it is converted back to the T-F domain. Essentially, the musical artifacts are reduced according to the speech production machanism which has the advantage of preserving the broadband structure and pitch information of the speech signal [6, 7]. Representing the binary masks of equation (7) and (8) in the cepstral domain we have,

$$\mathbf{M}_i^c(l,n) = DFT^{-1}\{ln(\mathbf{M}_i^f(k,n)) \mid_{k=0,..,K-1}\} \tag{11}$$

where l and k are the quefrency bin index and the frequency bin index respectively [6]. DFT is for the discrete Fourier transform and K represents the length of the DFT. After applying smoothing, the resultant smoothed mask is given as,

$$\overline{\mathbf{M}}_i^s(l,n) = \gamma_l \overline{\mathbf{M}}_i^s(l,n-1) + (1-\gamma_l)\mathbf{M}_i^c(l,n) \quad i = 1,..,N \tag{12}$$

where γ_l is a parameter for controlling the smoothing level. The selection of γ_l is important, therefore, according to the different values of l, the selection criterion for the value of γ_l is given as,

$$\gamma_l = \begin{cases} \gamma_{env} & \text{if } l \in \{0,...,l_{env}\}, \\ \gamma_{pitch} & \text{if } l = l_{pitch}, \\ \gamma_{peak} & \text{if } l \in \{(l_{env}+1),...,K\} \setminus l_{pitch} \end{cases} \tag{13}$$

where $0 \leq \gamma_{env} < \gamma_{pitch} < \gamma_{peak} \leq 1$, l_{env} is the quefrency bin index that represents the spectral envelop of the mask $\mathbf{M}^f(k,n)$ and l_{pitch} is the quefrency bin index showing the structure of the pitch harmonics in $\mathbf{M}^f(k,n)$. The principle employed for this range of γ_l is illustrated as follows. $\mathbf{M}^c(l,n)$, $l \in \{0,..,l_{env}\}$ basically represents the spectral envelop of the mask $\mathbf{M}^f(k,n)$. In this region, the value selected for γ_l is relatively low to avoid the distortion in the

envelop. Similarly, low smoothing is applied if l is equal to l_{pitch}, so that harmonic structure of the signal is maintained. High smoothing is applied in the last range selected for γ_l, which is able to reduce the artifacts without harming the pitch information and the structure of the spectral envelop. Different from [6], we calculate pitch frequency by using the segregated speech signal obtained in the section 2.2. Specifically, pitch frequency is computed as,

$$l_{pitch} = argmax_l\{sig^c(l,n) \mid l_{low} \leq l \leq l_{high}\}, \quad (14)$$

where $sig^c(l,n)$ is the cepstral domain representation of the segregated speech signal $\mathrm{y}^t(n)$. The range l_{low}, l_{high} is chosen so that it can accommodate pitch frequencies of human speech in the range of 50 to 500 Hz. The final smoothed version of the spectral mask is given as,

$$\overline{\mathbf{M}}_i^f(k,n) = exp(DFT\{\overline{\mathbf{M}}_i^s(l,n) \mid_{l=0,...,K-1}\}), \quad (15)$$

This smoothed mask is then applied to the output segregated speech signal obtained in section 2.2, as follows,

$$\overline{\mathbf{Y}}_i^f(k,n) = \overline{\mathbf{M}}_i^f(k,n).\mathbf{Y}_i^f(k,n) \qquad i = 1,...,N \quad (16)$$

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we evaluate the performance of the proposed method using simulations. The algorithm is applied to both artificially mixed signals and real room recordings.

3.1. Experimental setup

Speech signals from a pool of 12 different speakers including six male and six female utterances with 11 different languages are used in the experiments [8]. All the signals have the same loudness level. The hamming window is used with an overlap factor set to 0.75. The duration of the speech signal is 5 seconds with a sampling rate of 10 KHz. Rest of the parameters are set as: l_{env}=8, l_{low}=16, l_{high}=120, γ_{env}=0, γ_{pitch}=0.4, and γ_{peak}=0.8. Performance indices used in evaluation include signal to noise ratio (SNR), the percentage of energy loss (PEL) and the percentage of noise residue (PNR) [8]. Notations mSNR$_i$, mSNR$_o$ and $\triangle$SNR are also used in the evaluation. SNR$_i$ is the ratio between the desired signal and the interfering signal taken from the mixture. SNR$_o$ is the ratio between the desired signal resynthesized from the ideal binary mask to the difference of the desired resynthesized signal and the estimated signal [8]. mSNR$_i$ and mSNR$_o$ are the average results for 50 random tests. $\triangle$SNR is given by, $\triangle$SNR=mSNR$_o$-mSNR$_i$.

3.2. General evaluation

Firstly we mix the sources artificially using a simulated room model [9]. A series of experiments were carried out to evaluate the proposed method changing with the parameters. In the first experiment with the above parameters setup, for the RT equal to 100 ms, experiments have been performed for different window length of 256, 512, 1024 and 2048. The results are given in Table 1. The average behaviour is shown for 50 different convolutive mixtures, with each consisting of two speech sources randomly picked up from a pool of 12 speech signals [8]. It can be seen that the highest $\triangle$SNR is obtained for the window length of 512. Therefore, the window length is fixed to 512 for all the subsequent experiments.

The performance of the proposed method for different FFT frame lengths is given in Table 2. Results show that by increasing the FFT frame length from 512 to 2048, performance of the algorithm in terms of SNR, PEL and PNR is improved. The best performance is obtained at 2048. Hence FFT frame length used for the subsequent experiments is fixed to 2048. In Table 3 average results for PEL, PNR and $\triangle$SNR are given for different values of RT. A noticeable change with this table is that the performance degrades with increasing RT.

In above experiments, we have not considered microphone noise in the mixtures. Now, we conduct experiments for noisy mixtures by adding white noise with noise level calculated with respect to the level of the mixtures, with a weaker noise corresponding to a smaller number [8]. The average $\triangle$SNR values for different noise levels are given in Table 4. It can be observed that the performance of the algorithm decreases as the noise level is increased, and similar to [8], the algorithm can tolerate the noise levels up to -20 dB.

The above experiments provide a general view for the objective performance of the proposed approach. We observed from our experiments that cepstral smoothing does not improve the objective performance in terms of SNR measurement. In fact, it decreases slightly (hence negligible) the SNR gain from the output of IBM. However, the cepstral smoothing does improve considerably the quality of the separated speech. To show this, a subjective listening test has been conducted for 10 listeners. Each listener awarded a score from one (musical noise clearly audible) to five (not audible) for the final segregated speech signals. The average results of mean opinion score (MOS) for the ten listeners is shown for RT equal to 30 and 100 ms respectively. The MOS is given in Table 5. It indicates that using cepstral smoothing gives higher MOS as compared the methods withouting using smoothing, e.g., [8], suggesting improved quality of the separated speech.

3.3. Comparison

The performance of our method is compared with the recent algorithm in [8]. For making this comparison, we have performed experiments using the real recordings in a reverberant room with RT=400 ms. Two omnidirectional microphones vertically placed and closely spaced are used for the recordings. Different loud speaker positions are used to estimate the room impulse responses. Clean speech signals from the pool of 12 speakers were convolved with the room impulses to generate the source signals [8]. The system used here has the given specifications. The processor used is Intel(R) Xeon(TM) 3.00GHz. Memory of the system is 31.48 GBytes. The results are given in Table 6. The results show that our proposed algorithm is 18 times faster than the recent method.

Table 1. Results for Different Window Lengths

Window Length	PEL	PNR	mSNR$_i$	mSNR$_o$	$\triangle$SNR
256	9.10	15.30	1.10	7.11	6.01
512	8.60	14.48	1.10	7.44	6.34
1024	9.30	14.70	1.10	7.11	6.01
2048	10.92	15.92	1.12	6.32	5.20

Table 2. Results for Different FFT Frame lengths

NFFT	PEL	PNR	mSNR$_i$	mSNR$_o$	$\triangle$SNR
512	9.06	14.96	1.10	7.17	6.06
1024	8.65	14.53	1.10	7.40	6.30
2048	8.60	14.48	1.10	7.44	6.34

Table 3. Results for Different RT

RT	PEL	PNR	mSNR$_i$	mSNR$_o$	$\triangle$SNR
40	2.16	2.24	1.13	13.22	12.08
60	3.79	4.12	1.15	10.94	9.79
80	5.50	8.30	1.14	9.42	8.27
100	8.60	14.48	1.10	7.44	6.34
120	10.99	19.53	1.03	6.30	5.26
140	13.36	24.14	0.94	5.48	4.53
150	13.86	25.38	0.90	5.29	4.39

Table 4. Results for Different Noise Levels

Noise	PEL	PNR	mSNR$_i$	mSNR$_o$	$\triangle$SNR
-10dB	9.46	16.49	1.09	6.91	5.81
-20dB	8.62	14.52	1.10	7.43	6.33
-30dB	8.60	14.48	1.10	7.44	6.34
-40dB	8.60	14.48	1.10	7.45	6.34

Table 5. Results for Musical Noise

RT	MOS before smoothing	MOS after smoothing	MOS for the Method in [8]
30	3.30	3.92	3.07
100	2.04	2.55	2.26

Table 6. Comparison Between Proposed and the Method in [8]

Algorithm	PEL	PNR	$\triangle$SNR	Total time	Time per test
Proposed	30.56	9.73	2.50	40min	0.8min
Method in [8]	17.14	49.33	2.64	700min	14min

The method in [8] requires 700 minutes for 50 random tests and 14 minutes per test. In contrast, our proposed method is faster and requires 40 minutes for 50 tests and 0.8 minutes per test. Although the results for $\triangle$SNR are comparable, our method outperforms significantly the method in [8] in terms of computational efficiency. Listening tests also suggest that our results have a better quality than those in [8]. Some demos are available on the website [10] for both real and artificial recordings.

4. CONCLUSIONS AND FUTURE WORK

A novel multistage approach has been presented for the segregation of speech signals from their convolutive mixtures using two-microphone recordings. The convolutive mixtures are first separated using a constrained convolutive ICA algorithm. The separated sources are then used to estimate the IBM, which are further applied to the T-F representation of original mixtures. In order to reduce the musical noise induced by T-F masking, cepstral smoothing is applied to the estimated IBM. The segregated speech signals are observed to have considerably improved quality and reduced musical noise. Future work involves the segregation of speech signals in underdetermined cases with high RT.

5. REFERENCES

[1] W. Wang, S. Sanei, and J. A. Chambers, "Penalty function-based joint diagnolization approach for convolutive blind separation of nonstationary sources," *IEEE Trans. Signal Process.*, vol. 53, pp. 1654–1669, May 2005.

[2] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis", in Divenyi P.(ed.), *Speech Separation by Humans and Machines*, pp. 181–297, Kluwer Academic, Norwell MA, 2005.

[3] S. Araki, R. Mukai, S. Makino, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixture of speech," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 109–116, March 2003.

[4] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*, Wiley Press, 2002.

[5] L. Parra and C. Spence, "Convolutive blind separation of non stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 320–327, May 2000.

[6] N. Madhu, C. Breithaupt, and R. Martin, "Temporal smoothing of spectral masks in the cepstral domain for speech separation," in *Proc. IEEE ICASSP*, pp. 45–48, March 2008.

[7] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice Hall, New Jersey, 1975.

[8] M. S. Pedersen, D. L. Wang, J. Larsen, and U. Kjems, "Two-microphone separation of speech mixtures," *IEEE Trans. on Neural Networks*, vol. 19, pp. 475–492, March 2008.

[9] N. D. Gaubitch, "Allen and Berkley image model for room impulse response, Imperial college london," [Online]. Available: http://www.commsp.ee.ic.ac.uk/%7Endg/downloadfiles/mcsroom.m

[10] W. Wang, [Online]. Available: http://personal.ee.surrey.ac.uk/Personal/W.Wang/demondata.html