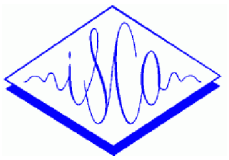

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Modulation domain blind source separation for noisy speech mixture

Yi Zhang and Yunxin Zhao

Department of Computer Science
University of Missouri, Columbia, Missouri, 65211, USA
Email: yzcb3@mail.missouri.edu, ZhaoY@missouri.edu

ABSTRACT

In this paper, we propose a noise-robust blind speech separation (BSS) method by using two microphones. We first use modulation domain real and imaginary spectral subtraction (MRISS) to enhance both magnitude and phase spectra of the speech mixture inputs. We then estimate the direction of arrivals (DOAs) of the speech sources and perform time-acoustic-modulation frequency masking to recover the source signals. Our experimental results in five types of noise conditions have showed the superior performance of the proposed method in comparison with the conventional acoustic domain DOA based separation method.

Index Terms — time-frequency masking, direction of arrival, modulation frequency, blind speech separation

10.21437/Interspeech.2012-435

1. INTRODUCTION

Blind speech separation (BSS) is an approach of estimating source speech signals from the observed mixtures of source speech. The BSS problems can be categorized according to the numbers of sources and sensors. When the number of sources is greater than the number of sensors, the source separation problem becomes more challenging. One of the workable methods for solving this problem is time-frequency (T-F) masking [1-4], which is based on the sparseness assumption that the energies of independent speech signals rarely overlap in the T-F domain.

Although many BSS methods work well when background noise is low, their performance is significantly degraded when background noise is high. Several methods have been proposed to deal with noisy conditions for BSS. Joho et. al. [5] proposed a two-stage algorithm, where principal component analysis (PCA) was used for noise reduction and independent component analysis (ICA) was used for blind source separation. They showed good results by using 5-20 sensors to separate a 5 source mixture at input SNR of 15dB. Vu and Umbach [6] proposed a BSS algorithm for directional noise. They combined T-F sparseness with the generalized eigenvalue decomposition of the power spectral density of noisy speech, and successfully separated 2 sources by using an 8-microphone array at the input SNR of 0dB and reverberation time 0 ~500 ms. Choi and Cichocki [7] used the method of joint diagnoalisation of multiple time-delayed correlation matrices of the observation data to estimate the source mixing matrix, and they achieved good results when the input SNR was 10-15 dB. Aichner et al. [8] presented a real-time implementation for BSS of convolutive mixtures, and produced a high separation performance in a noisy car environment at SNR of 0 dB.

In this current work, we focus on the problem of blind speech separation in noisy conditions and propose a robust modulation domain time-frequency (T-F) masking method to carry out the task. The key mechanism of our approach is recovering speech phase in noise and from which to derive DOA information for source separation, and performing source separation in the modulation frequency domain to reduce distortion in the separated speech. The proposed method consists mainly of two steps. In the first step, the noisy speech mixture signal received at each sensor is enhanced by a modulation domain real and imaginary spectral subtraction method (MRISS) proposed in [11], where both the magnitude and the phase spectra of the mixture signals are enhanced from the noise. In the second step, the DOAs of the speech sources are estimated by using an asymmetric Laplacian mixture density (ALMD) model fitting, and time-acoustic-modulation frequency masks are computed and applied to the modulation spectra for source separation and signal recovery. We conducted experiments to evaluate the performance of the proposed method by using the criteria of perceptual evaluation of speech quality (PESQ), segmental signal-to-distortion ratio (SDR), and signal-to-interference ratio (SIR) gain. We compared the performance of the proposed method with the conventional acoustic frequency domain DOA based source separation and showed the superior performance of the proposed method.

This paper is organized as follows. In Section 2, we discuss the background of DOA based BSS; in Section 3 we describe our proposed BSS method; in Section 4 we present experimental results, and in Section 5 we give a conclusion.

2. DOA BASED BLIND SPEECH SEPARATION

2.1 Far field signal model

In a sound field of N simultaneous speech sources and two sensors, the signal received by the ith sensor is modeled as

$$x_i(t) = \sum_{n=1}^{N} \sum_{l} s_n(t-l) h_{i,n}(l), \quad i = 1,2 \tag{1}$$

where $s_n(t)$ denotes the nth source, and $h_{i,n}(t)$ is the impulse response from the nth source to the ith sensor.

In the far field model, a plane-wave is assumed for speech sound, and in the absence of reverberation and attenuation, the DFT of the impulse response of an acoustic path is simplified as

$$H_{i,n}(k) \approx \exp\{-j2\pi k \tau_{i,n}\} \tag{2}$$

where k is the frequency index and $\tau_{i,n}$ is the time delay from the nth source to the ith sensor. Accordingly, the acoustic spectra of the signals at the two sensors are

$$X_i(t,k) = \sum_{n=1}^{N} S_n(t,k) \exp\{-j2\pi k \tau_{i,n}\}, \quad i = 1,2 \tag{3}$$

2.2 DOA Estimation

In histogram based DOA estimation, the far field model and the sparseness property of speech are utilized. The sparseness property states that the energies of independent speech signals rarely overlap in time-frequency domain, and therefore at each T-F element the signal energy is dominated by one source. Assume that at a T-F element (t,k) the signal energy is dominated by the nth source. Expressing the relative time delay τ_n as a function of the speed of sound c, the sensors spacing d, and the arrival angle θ_n leads to [9]

$$\frac{X_1(t,k)}{X_2(t,k)} \approx \exp\{j2\pi k \tau_n\} = \exp\left\{\frac{j2\pi k d \cos\theta_n}{c}\right\} \tag{4}$$

where τ_n is referred to as the inter-sensor time delay (ITD) at frequency k and time frame t, $2\pi k d \cos\theta_n / c$ as the inter-sensor

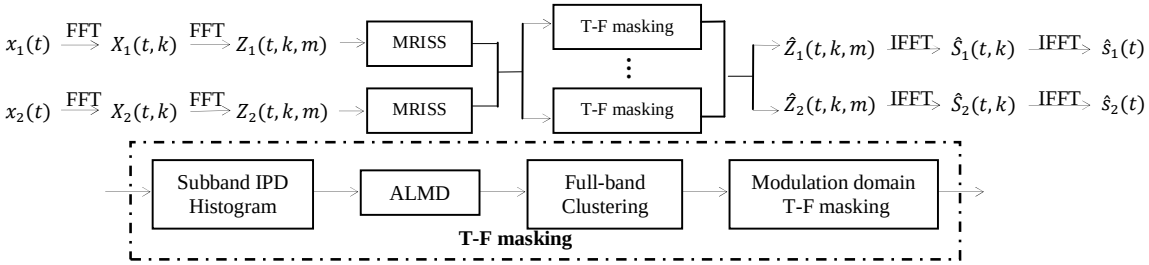

Fig. 1 *Block diagram of the proposed method*

phase difference (IPD), and $2\pi d cos\theta_n/c$ as the frequency normalized IPD. Histogram can then be generated for the normalized IPD of the T-F elements over a block of time frames (we used a block length of 2 seconds with 62 frames per block for a reasonable tradeoff between tracking signal changes and having sufficient frames for histogram generation.).

2.3 Speech Separation

For a speech mixture input $x(t)$ at a microphone sensor, speech separation can be performed based on a mixture density modeling of the clustering structure of the IPD data. Based on the model, the posterior probabilities that the signal energy at a T-F element is associated with the different source directions are computed to generate the T-F masks $M_n(t,k)$ for the source signals $s_n, n = 1,\dots,N$. Speech separation can then be performed by extracting the source signals according to Eq. (5):

$$\hat{S}_n(t,k) = M_n(t,k)X(t,k) \quad (5)$$

where $\hat{S}_n(t,k)$ is the extracted signal component of the source n. The source speech signals are obtained by inverse transforming $\hat{S}_n(t,k)$ into the time domain. In this paper, we used the separation results derived from one of the two sensors.

3. PROPOSED METHOD

The procedure of the proposed method for blind speech separation in noise is shown in Fig. 1. Sensor inputs of the speech mixture signals $x_i(t)$ are first transformed into the modulation domain and MRISS enhancement is performed to recover the complex modulation spectra in each channel. For each fixed modulation frequency referred to as a modulation frequency layer, DOA is estimated by using the ALMD model fitting on a subband IPD histogram, and a full-band clustering is performed to obtain the T-F masks. The separated speech spectra are transformed back to the time domain to recover the source speech signals.

3.1 Modulation domain IPD distribution

We assume that at each acoustic frequency and within each modulation time window the dominant source is mostly consistent, and we refer to this property as sparsity in time-acoustic-modulation frequency. When the sparsity property holds, $\exp\{-j2\pi k\tau_{i,n}\}$ is a constant within a modulation window at a fixed acoustic frequency bin k, since the positions of the source n and sensor i are fixed. From Eq. (3), we can then derive:

$$Z_i(t,k,m) = FFT\{S_n(t,k)\}\exp(-j2\pi k\tau_{i,n}), i = 1,2 \quad (6)$$

which leads to

$$\frac{Z_1(t,k,m)}{Z_2(t,k,m)} \approx \exp\{j2\pi k\tau_n\} = \exp\left\{\frac{j2\pi k d cos\theta_n}{c}\right\} \quad (7)$$

We utilize the source DOA information given by Eq. (7) to perform source separation in the modulation domain.

In our experiment, we found that this sparsity assumption held well for clean and low reverberation (< 0.3 seconds) speech mixtures when the modulation window length ranged from 80 ~ 160 ms, and the modulation domain IPDs on each modulation layer showed similar distributions as in the acoustic frequency domain.

We used the entropy and Gini measures [10] to compare the speech sparsity properties in the acoustic and modulation domains. The overall sparsity score in the acoustic domain is computed as $\overline{W} = \frac{1}{TK}\sum_k\sum_t W(t,k)$, where T and K are the numbers of time frames and acoustic frequency bins, respectively, and $W(t,k)$ is one of the Entropy or Gini scores. Similarly, the overall sparsity score in the modulation domain is computed as $\overline{W} = \frac{1}{TKM}\sum_t\sum_k\sum_m W(t,k,m)$, where M is the number of modulation frequency bins. The results are shown below in Table 1 (modulation window length was 120 ms). Note that smaller $\overline{W}$ in entropy and larger $\overline{W}$ in Gini represent stronger sparsity property.

Table 1 *Sparsity measures in acoustic and modulation domains*

	Entropy	Gini
Acoustic domain	0.0612	-0.0511
Modulation domain	0.0522	-0.0421
Relative difference	14.7%	17.6%

The result in Table 1 confirmed that the sparsity property not only holds in the modulation domain, and it is even stronger than that in the acoustic domain by about 15% relative.

3.2 Real and imaginary modulation spectral subtraction

In order to preserve the IPD information for DOA based source separation in noise, we use our proposed MRISS method to enhance both magnitude and phase spectra of the speech mixture signals. For details of MRISS, please refer to [11].

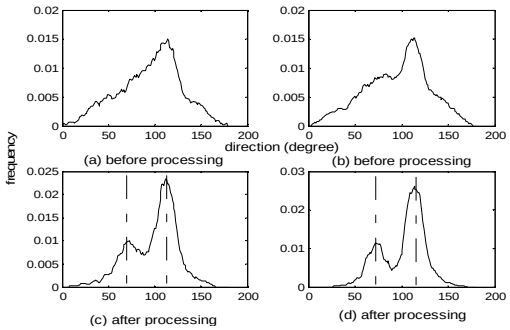

Fig. 2 *DOA histogram (left: white noise, right: babble noise)*

Fig. 2 shows four IPD histograms derived before and after the MRISS processing in a scenario of 2 sensors and 2 sources, where Fig. 2 (a) and (c) are for the case of white noise, and Fig. 2 (b) and (d) are for the case of babble noise, with SNR of 0 dB. Without enhancement, the histograms (top) could not show two source directions, while after the MRISS enhancement, the histograms

(bottom) each showed two peaks clearly, from which one could easily distinguish the two source directions (the dash-dotted lines represent the true source directions).

3.3 Asymmetric Laplacian mixture density

The distribution of IPD data in each cluster often has long tails and is asymmetric around the mean, especially when the sources are close to each other. Hence the commonly used K-means clustering [12] or Gaussian mixture density (GMD) [13] are not a good fit in such scenarios. We propose to use a mixture of asymmetric Laplacian density function to model the distribution of IPD data instead.

The probability density function of an asymmetric Laplacian random variable x is defined as [14]

$$p(x;\mu,\sigma,q)=\frac{q(1-q)}{\sigma}exp\left\{-\frac{x-\mu}{\sigma}[q-I(x\le\mu)]\right\} \quad (8)$$

where $0<q<1$ is the skew parameter, $\sigma>0$ is the scale parameter, μ is the mean parameter, and $I(\cdot)$ is the indication function with $I(A)=1$ if A is true, and $I(A)=0$ if A is false.

A mixture of K asymmetric Laplacian density functions is then defined as

$$p(x|\theta)=\sum_{i=1}^{K}\pi_i p(x|\mu_i,\sigma_i,q_i) \quad (9)$$

We have extended the parameter estimation for Eq.(8) of [14] into the mixture density of Eq.(9) by using the EM algorithm. However, due to the limited space here, the details are omitted.

3.4 Full-band clustering

For each modulation layer, an ALMD is estimated, and the full-band IPDs are clustered around the ALMD μ parameters after scaling them by the acoustic frequency k and taking into account of phase unwrapping, i.e., $\varphi_{i,k,m}=\mu_{i,m}k\pm 2n\pi$, where $\pm 2n\pi$ are phase unwrapping terms needed for high frequency bins. The posterior probability that an IPD sample at (t,k,m) belongs to the ith cluster is computed as

$$P_r(i|IPD(t,k,m))=\frac{\pi_{i,k,m}p\left(IPD(t,k,m)\middle|\varphi_{i,k,m},\sigma_{i,k,m},q_{i,k,m}\right)}{\sum_{j=1}^{K}\pi_{j,k,m}p\left(IPD(t,k,m)\middle|\varphi_{j,k,m},\sigma_{j,k,m},q_{j,k,m}\right)} \quad (10)$$

where $\pi_{i,k,m}$, $\sigma_{i,k,m}$ and $q_{i,k,m}$ were estimated by using the EM estimation in Section 3.3.

4. EXPERIMENTS

4.1 Experiment setting

We evaluated the performance of the proposed method in noisy conditions with the number of sources and sensors being 2 and 2 (the underdetermined case were not studied due to the limited space here). The anechoic room impulse responses (RIR) of the RWCP dataset [15] were used to generate the speech mixture data, where the two sensors were mounted on a circular array with a spacing of 5.85cm, and the two speakers were about 2m away from the sensors at the directions of 70^0 and 110^0. The target and interference speech source signals came from the TIMIT dataset with a sampling rate of 8k Hz. The target speech includes 40 sentences, from 2 male and 2 female speakers, and each speaker contributed 10 sentences. The target and interference speech signals were of different genders. The input SIR was 0 dB. In generating the noisy speech mixtures, the two source speech signals were first convolved with the respective RIRs and summed at the corresponding microphones, and the microphone signals were then corrupted by spatially incoherent additive noises. We used five kinds of noises: white, babble, pink, volvo and factory2 from the NOISEX92 database. The SNR ranged from 0dB to 10dB, where the SNR was computed as the ratio of the clean mixture power over the noise power.

The baseline was the conventional DOA based acoustic-domain separation method [9] without pre-processing. The proposed method used the MRISS enhancement, and was implemented in the modulation frequency domain. In both the baseline and the proposed methods ALMD was used and the source number was assumed to be known. In the next section, the observed mixture is referred to as 'mix,' the baseline method as 'baseline,' and the proposed method as 'proposed.'

We found that for every noise condition and evaluation criterion, our proposed method delivered the best performance. We therefore conducted a statistical significance test on the performance difference between the proposed method and the baseline method using a one-sided student-t test, with n-1 = 39 degrees of freedom at the significance level of $\alpha=0.05(t_\alpha=1.686)$ [16].

4.2 Experiment results and discussion

We evaluated the performance of speech separation by using the criteria of PESQ, segmental SDR and SIR gain. All the results were averaged over the two channels.

(1) PESQ

PESQ is widely adopted for automated assessment of speech quality as experienced by a listener. Our evaluation results were generated by using the PESQ routine of [17].

Table 2 *PESQ results under different noise conditions*

Noise (SNR dB)		mix	baseline	proposed
white	0	1.351	1.323	1.859
	5	1.517	1.922	2.082
	10	1.641	2.147	2.267
babble	0	1.504	1.715	1.890
	5	1.622	2.137	2.160
	10	1.691	2.261	2.339
pink	0	1.476	1.802	1.960
	5	1.617	2.084	2.168
	10	1.703	2.270	2.336
volvo	0	1.770	2.511	2.600
	5	1.786	2.601	2.672
	10	1.793	2.629	2.697
Factory2	0	1.645	2.101	2.203
	5	1.717	2.302	2.378
	10	1.755	2.457	2.486

From Table 2, we see that the proposed method outperformed the baseline in every noise condition. In white noise of 0 dB, the baseline failed to work while the proposed method still gained 0.5 in PESQ score. When SNR was low, the improvement of the proposed method over the baseline was larger, which showed the robustness of the proposed method. The improvements were significant for every noise types at 0 and 5 dB SNRs, and for white, babble and pink noises at 10 dB SNR.

(2) Segmental SDR

Segmental SDR measures the distortion between the recovered signal and the reference target signal, and it is defined as the average SDR values calculated from short segments of speech:

$$SegSDR=\frac{1}{N}\sum_{t=0}^{N-1}10\log_{10}\sum_{k=0}^{K-1}\frac{|S(t,k)|^2}{|S(t,k)-\hat{S}(t,k)|^2}$$

where t is the segment index and k is the frequency index, $S(t,k)$ and $\hat{S}(t,k)$ are the reference speech and recovered speech,

respectively. In computing the SegSDR values, we used a conventional speech analysis window length of 32 ms.

Table 3 *Segmental SDR results (dB) in different noise conditions*

Noise (SNR dB)		mix	baseline	proposed
white	0	-8.622	-6.541	1.130
	5	-3.383	4.342	4.773
	10	1.433	6.548	7.202
babble	0	-8.592	-1.163	0.762
	5	-3.336	3.378	4.491
	10	1.348	6.403	7.004
pink	0	-8.604	-0.551	1.100
	5	-3.378	3.375	4.770
	10	1.438	6.436	7.189
volvo	0	-8.477	-0.360	2.290
	5	-3.321	3.804	5.536
	10	1.430	6.345	7.676
Factory2	0	-8.562	-0.654	1.566
	5	-3.324	3.710	5.054
	10	1.370	6.531	7.329

The Segmental SDRs in Table 3 showed a similar trend as the PESQs in Table 2. The proposed method was consistently the best in every noise condition and at every SNR. When SNR was high, both the baseline and the proposed method worked well, but when the SNR decreased, the performance of the baseline degraded much faster than the proposed method did. Under this criterion, the improvements by the proposed method over the baseline were statistically significant in all the noise conditions and at all SNRs.

(3) SIR gain

The SIR (dB) is defined as

$$SIR = 10log_{10}\left(\frac{\sum_t s_T^2(t)}{\sum_t s_I^2(t)}\right)$$

where s_T and s_I are the time-domain target speech and interference speech, respectively. The SIR gain was computed as the SIR of the output signal minus the SIR of the input signal, and it reflects the improvement in SIR due to source separation.

Table 4 *SIR gain results (dB) in different noise conditions*

Noise (SNR dB)		baseline	proposed
white	0	1.031	17.336
	5	14.013	17.332
	10	16.142	17.289
babble	0	10.935	15.870
	5	16.012	17.573
	10	16.518	17.362
pink	0	13.764	17.740
	5	16.163	17.577
	10	16.246	17.513
volvo	0	15.987	17.211
	5	16.155	17.159
	10	16.047	17.136
Factory2	0	15.750	17.394
	5	16.295	17.373
	10	16.263	17.288

The proposed method produced consistent SIR gains for all noise types and SNRs, where the gain became larger when the SNR became lower. The baseline method obtained good results when SNR was high, but its performance at low SNR was not always satisfactory. This shows that the proposed method is more robust to the studied noise conditions than the baseline method. It is worth noting that in white noise at 0dB SNR, the baseline method almost failed completely, while the proposed method obtained an SIR gain of 17.336 dB, which confirmed MRISS's positive effect on DOA estimation as illustrated in Fig. 2. The improvements of the proposed method over the baseline are statistically significant in all the cases except the case of volvo noise at 10 dB SNR.

5. CONCLUSION

In this work we have proposed a robust blind speech separation method for noisy speech mixtures. Experimental results have shown the capability of the proposed method in extracting the DOA information from two-channel noisy speech mixtures, and in improving the performance of speech separation measured by PESQ, Segmental SDR, and SIR gain. The improved performance was due to the phase spectra recovery by the real-imaginary enhancement pre-processing as well as the reduced speech distortion by the modulation domain processing. In the future work, we shall investigate problems of source number estimation, underdetermined speech source separation, and speech separation in reverberant conditions.

6. REFERENCE

[1] N. Roman, D. Wang, and G. Brown, "Speech segregation based on sound localization," J. Acoust. Soc. Am., vol. 114, no. 4, pp. 2236-2252, 2003

[2] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process., vol. 52, no. 7, pp. 1830-1847, 2004

[3] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation maximization sources separation and localization," IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 2, pp. 382-394, 2010

[4] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," IEEE Trans. Audio, Speech, Lang. Process., vol. 19, pp. 516-527, 2011

[5] M. Joho, H. Mathis, and R. H. Lambert, "Overdetermined blind source separation: using more sensors than source signals in a noisy mixture," Proc. ICA and BSS, pp. 81-86, 2000

[6] D. H. T. Vu and R. H. Umbach, "Blind speech separation in presence of correlated noise with generalized eigenvector beamforming," Proc. Voice Communication, pp. 1-4, 2008

[7] S. Choi and A. Cichocki, "Blind separation of nonstationary sources in noisy mixtures," Electronics Letters, vol. 36, pp. 848-849, 2000

[8] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments," Signal Processing, vol. 86, pp. 1260-1277, 2006

[9] S. Araki, H. Sawada, R. Mukai, S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," Proc. ICASSP, vol. 5, pp. 33-35, 2006

[10] N. Hurley, and S. Rickard, "Comparing measures of sparsity," Proc. MLSP, pp. 55-60, 2008

[11] Y. Zhang, and Y. Zhao, "Spectral subtraction on real and imaginary modulation spectra," Proc. ICASSP, pp. 4744-4747, 2011

[12] Y. Li, A. Cichocki, and S. Amari, "Analysis of sparse representation and blind source separation," Neural Computation, vol. 16, pp. 1193-1234, 2004

[13] A. Koutvas, E. Dermatas, and G. Kokkinakis, "Blind speech separation of moving speakers in real reverberant environments," Proc. ICASSP, vol. 2, pp. 1133-1136, 2000

[14] K. Yu, and J. Zhang, "A three parameter asymmetric laplace distribution and its extension," Communications in Statistics – Theory and Methods, vol. 34, pp. 1867-1879, 2005

[15] RWCP Sound Scene Database in Real Acoustic Environments, ATR Spoken Language Translation Research Laboratory, Japan 2001.

[16] A. Papoulis, Probability, Random Variables, and Stochastic Processes, 3rd ed. McGraw-Hill, New York, 1991

[17] http://www.utdallas.edu/~loizou/speech/software.htm