TARGET SPEECH EXTRACTION IN COCKTAIL PARTY BY COMBINING BEAMFORMING AND BLIND SOURCE SEPARATION

Lin Wang[1,2], Heping Ding[2] and Fuliang Yin[1]

[1] **School of Electronic and Information Engineering, Dalian University of Technology, P. R. China**
[2] **Institute for Microstructural Sciences, National Research Council Canada**

wanglin_2k@sina.com, heping.ding@nrc-cnrc.gc.ca, flyin@dlut.edu.cn

Due to the ambient noise, interferences, reverberation, and the speakers moving and talking concurrently, it is a challenge to extract a target speech in a real cocktail-party environment. Emulating human auditory systems, this paper proposes a two-stage target speech extraction method which combines fixed beamforming and blind source separation. With the target speaker remaining in the vicinity of a fixed location, several beams from a microphone array point at an area containing the target, then the beamformed output is fed to a blind source separation scheme to get the target signal. The fixed beamforming preprocessing enhances the robustness to time-varying environments and makes the target signal dominant in the beamformed output and hence easier to extract. In addition, the proposed method does not need to know the knowledge of source positions. Simulations have verified the the effectiveness of the proposed method.

INTRODUCTION

Extracting a desired speech signal from its corrupted observations is essential for tremendous applications of speech processing and communication [1]. One of the hardest situations to handle is the extraction of a desired speech signal in a "cocktail party" condition - from mixtures picked up by microphones placed inside a noisy and reverberant enclosure. In this case, the target speech is immersed in ambient noise and interferences, and distorted by reverberation. Further more, the environment may be time-varying. Generally, there are two well-known techniques that may achieve the objective: blind source separation (BSS) and beamforming.

Assuming mutual independence of the sources, BSS is a technique for recovering them from observed signalss with the mixing process unknown [2, 3]. Nevertheless, BSS may not appropriate for target signal extraction in a cocktail-party condition. First, under-determined situations can result from the fact that there is only a limited number of microphones. Second, BSS processes the target signal and interference equally; it can be difficult to separate many signals simultaneously and also a waste of computational power if we want only one target. Third, BSS performs poorly in high reverberation, where the mixing filters are very long.

With a microphone array, beamforming is a well known technique for target extraction. It can be implemented as a data-independent fixed beamforming or data-dependent adaptive one [4, 5]. Fixed beamforming is more preferred in complicated environments due to its robustness. It achieves a directional response by coherently summing signals from multiple sensors based on a model of the wavefront from acoustic sources. It can enhance signals from the desired direction while suppressing ones from other directions. Thus, fixed beamforming can be used for both noise suppression and dereverberation. However, its performance also degrades in cocktail-party conditions. First, the performance is closely related to the microphone array size - a large array is usually required to obtain a satisfactory result but may not be practically feasible. Second, beamforming cannot reduce reverberation coming from the desired direction.

Because of the reasons above, few methods proposed in recent years have good separation results in a real cocktail-party environment. In contrast, a human has a remarkable ability to focus on a specific speaker in that case. This selective listening capability is partially attributed to binaural hearing. Two ears work as a beamformer which enables directive listening [6], then the brain analyzes the received signals to extract sources of interest from the background, just as blind source separation does. Stimulating this principle, we propose to extract the target speech by combining beamforming and blind source separation. In fact, the idea of combining both technologies has been proposed by several researchers [7, 8]. In [8], the beamforming as a preprocessor of BSS, forms a number of beams each pointing at a source. This makes subsequent separation easier. However, it requires that prior knowledge of all source positions, which is seldom available in real life. We extend the work in [8] by applying it to a special case of blind source extraction problem in noisy cocktail party environments, where only one source is of interest. Instead of focusing on all the sources, the proposed method forms just several fixed beams at an area containing the target source. The beamforming enhances the robustness of the algorithm to time varying environments. After that, the

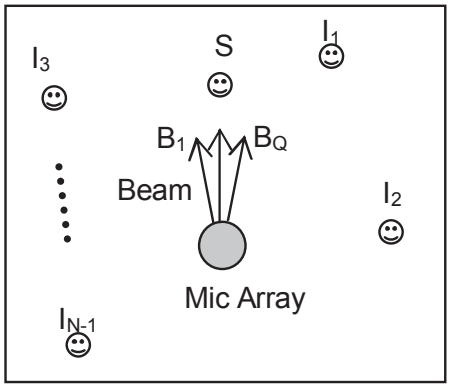

Figure 1. Illustration of the proposed method

target source becomes dominant in the beamformed output and it is easier for a blind source separation algorithm to extract it. Since the proposed method only needs the position of the target to do beamforming, it can be more practical.

PROPOSED METHOD

In a cocktail party, each speaker may move and talk freely. While this is a most difficulty for source separation, it is often in such case that the target speaker stays in a position or moves slowly and the noisy environment around it is time-varying, *e.g.*, moving interfering speakers and the ambient noise. For this specific situation, a target speech extraction method with a microphone array is proposed. It is illustrated in Fig. 1, where the target source S and N-1 interfering sources $I_1, \cdots, I_{N-1}$, are convolutively mixed and observed at an array of M microphones. To extract the target, Q beams ($Q \leq N$) are formed at an area containing it, with a small separation angle between adjacent beams; then the Q beamformed outputs are fed a blind separation scheme. Using beamforming as a preprocessor for BSS, the method possesses the advantages of both while complementing their weakness. In particular,

1) the residuals of interference at the output of beamforming are further reduced by BSS;
2) the poor separation performance of BSS in reverberant environments is compensated for by beamforming, which deflates the reflected paths and shortens the mixing filters;
3) the beamformer enhances the source that is in its path and suppresses the ones outside. It provides a cleaner output for the BSS to process; and
4) the fact that there are fewer beams than sources reduces the dimensionality of the problem and saves computation.

In a word, the target signal becomes dominant in the beamformed output and is hence easier to extract. Meanwhile, as seen in Fig. 1, the beams are pointing at an area containing the target, as opposed to the interfering sources. This is very important for operation under a time-varying condition, because

1) when the target speaker remains in a constant position while others move, it is impractical to know all speakers' positions and steer a beam at each of them;
2) there is no need to steer the beams at individual speakers since only the target speaker is of interest;
3) the target signal is likely to become dominant in at least one of the beamformed output channels if the beams point at an area containing the target speaker. Thus, it is possible to extract it as an independent source even if the number of beams is less than the sources [1]. This feature is very important for the proper operation of the proposed method; and
4) a seamless beam area will be formed by several beams with each covering some beamwidth. It is possible to extract the target signal even if it moves slightly inside this area. This feature may improve the robustness of the proposed method.

In a nutshell, beamforming makes primary use of spatial information while BSS utilizes statistical information contained in signals, and combining both technologies may help get a better extraction result. The signal flow of the proposed method is shown in Fig. 2. The implementation details are given in the two subsections to follow.

Beamforming

A superdirective fixed beamformer is designed in the frequency domain, using a circular microphone array. The principle of a filter-and-sum beamformer is shown in Fig. 3. Suppose a beamformer model with a target source r(t) and background noise n(t), the components received by the l'th sensor is $u_l(t) = r_l(t) + n_l(t)$ in the time domain. In the frequency domain the term is $u_l(f) = r_l(f) + n_l(f)$. The beamformer's output in the frequency domain is

$$x(f) = \sum_{l=1}^{M} b_l^*(f) u_l(f) = b^H(f) u(f) \tag{1}$$

where $b(f) = [b_1(f), \cdots, b_M(f)]^T$ is the beamforming weight vector composed of beamforming weights for each sensor, and $u(f) = [u_1(f), \cdots, u_M(f)]^T$ is the vector composed of outputs from each sensor, and $(\cdot)^H$ denotes conjugate transpose. The $b(f)$ depends on the array geometry and source directivity, as well as the array output optimization criterion such as a signal-to-noise ratio (SNR) gain criterion.

Suppose $r(f) = [r_1(f), \cdots, r_M(f)]^T$ is the source vector composed of the target source signals picked up by the sensors, and $n(f)$ is the noise vector composed of the spatially diffused noises also picked up by the sensors. Being a measure of improvement in signal-to-noise ratio, the array gain is defined as the ratio of the SNR at the output of the beamforming array to that at a single reference microphone. The reference SNR is defined, as in [9], to be the ratio of average signal power spectral densities over the microphone array, $\sigma_r^2(f) = E\{r^H(f) r(f)\}/M$, to the average noise power spectral density over the array, $\sigma_n^2(f) = E\{n^H(f) n(f)\}/M$. By derivation, the array gain at frequency f is expressed as

$$G(f) = \frac{b^H(f) R_{rr}(f) b(f)}{b^H(f) R_{nn}(f) b(f)} \tag{2}$$

where $R_{rr}(f) = r(f) r^H(f)/\sigma_r^2(f)$ is the normalized signal cross-power spectral density matrix, and $R_{nn}(f) = n(f) n^H(f)/\sigma_n^2(f)$ is the normalized noise cross-

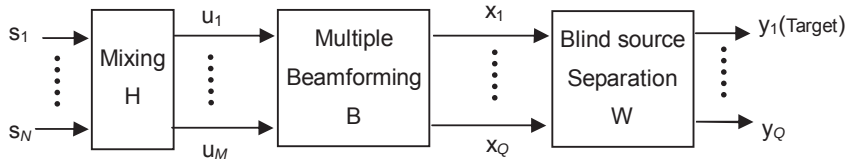

Figure 2. Signal flow of the proposed method combining beamforming and BSS

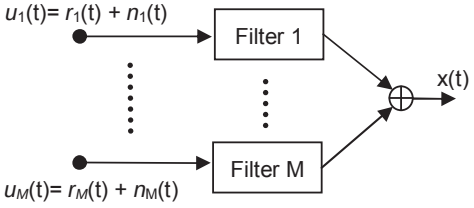

Figure 3. Principle of a filter-and-sum beamformer

power spectral density matrix. Provided that $R_{nn}(f)$ is nonsingular, equation (2) is maximized by the weight vector

$$b_{opt}(f) = R_{nn}^{-1}(f)r(f) \quad (3)$$

$R_{nn}(f)$ and $r(f)$ in equation (3) depend on the array geometry and the target source direction. Readers may refer to [8] for details on calculating $R_{nn}(f)$ and $r(f)$ for a circular array.

After calculating equation (3) at all frequency bins, the time-domain beamforming filter $b(n)$ is obtained by inverse Fourier transforming the $b_{opt}(f)$.

Blind source separation

Frequency-domain BSS is employed here due to its fast convergence and low computation. The mixed time-domain signals are converted into the time-frequency domain by short-time Fourier transform (STFT); then instantaneous independent component analysis (ICA) is applied to each frequency bin; after permutation alignment and scaling correction, the separated signals of all frequency bins are combined and inverse-transformed to the time domain.

For instantaneous ICA, we use a complex-valued Scaled Infomax algorithm, which is not sensitive to initial values, and is able to converge to the optimal solution within 100 iterations [10]. The scaling ambiguity problem is solved by using the Minimum Distortion Principle [11].

Permutation ambiguity inherent in frequency-domain BSS is a challenge problem. Generally, there are two approaches to solve it. One is to exploit the dependence of separated signals across frequencies [13, 12], and the other is to exploit the position information of sources: the directivity pattern of the mixing/unmixing matrix provides a good reference for permutation alignment [14]. However, in the proposed method, the directivity information contained in the mixing matrix does not exist any longer after beamforming. Even if the source positions are known, they are not much helpful to permutation alignment in the subsequent blind source separation. Consequently, what we can use for permutation is merely the first reference: the inter-frequency dependence of separated signals. Ref. [13] proposes a permutation alignment approach based on the power ratio measure. Bin-wise permutation alignment is applied first across all frequency bins, using the correlation of separated signal powers; then the full frequency band is partitioned into small regions based on the bin-wise permutation alignment result. Finally, region-wise permutation alignment is performed, which can prevent the spreading of the misalignment at isolated frequency bins to others and thus improves permutation. This permutation alignment algorithm is employed here.

EXPERIMENT AND ANALYSIS

We evaluate the performance of the proposed method in simulated conditions. A typical cocktail party environment with moving speakers and ambient noises is shown in Fig. 4. The room size is 7m × 5m × 3m, and all sources and microphones are 1.5m high. Four loudspeakers S1-S4 placed near the corners of the room play various interfering sources. Loudspeakers S5, S6 and S7 play speech signals concurrently. S5 and S6 remain in fixed positions, while S7 moves back and forth at a speed of 0.5 m/s. As the target, S5 is placed at either position P1 or P2. S5 simulates a female speaker, while S6 and S7 simulate male speakers. An 8-element circular microphone array with a radius of 0.1 m is placed as shown.

In blind source separation, the Tukey window is used in STFT, with a shift size of 1/4 window length, which is 2048 samples. The iteration number of instantaneous Scaled Infomax algorithm is 100. The permutation alignment algorithm in [13] is employed. In beamforming, a beamformer is designed with the algorithm presented in Section 2.1, using the circular array in Fig. 4. Three beams are formed towards S5, with the separation angle between two adjacent beams being 20˚. The room impulse responses are obtained by using the image method, with the reverberation time controlled by varying the absorption coefficient of walls [15]. The test signals last 8 seconds with a sampling rate of 8 kHz. The extraction performance is evaluated in terms of signal-to-interference ratio (SIR) for where the signal is the target speech.

With so many speakers in such a time-varying environment, BSS alone fails to work. Now we compare the performance of beamforming alone and the proposed method with reverberation RT_{60} of 130 ms and 300 ms respectively. The results are given in Table 1. As an example, for the close target case (P1) under RT_{60} = 300 ms, the input SIR is around -9 dB – the target is almost completely buried in noises and interference. The enhancement by beamforming alone is minimal. On the other hand, the proposed two-stage method improves the SIR by 15.1 dB. In the far target case (P2) of RT_{60} = 300 ms, the target signal received at the

microphones is much weaker with an input SIR around only -11 dB. The proposed method is still able to extract the target signal with an output SIR of 3.3 dB and a total SIR improvement of 13.5 dB.

For the close target (P1) with RT_{60} = 300 ms, Fig. 5 shows the waveforms at various processing stages: sources, microphone signals, beamformer outputs, and finally the BSS outputs. It can be seen that, the target signal S5 is totally buried in noises and interference in the mixture signals; it is enhanced to a certain degree after beamforming but is still difficult to tell from the background; and after blind source separation, the target signal is clearly exhibited at the channel Y2. In addition, an interference signal (S6) is observed at the output channel Y1, and the noise-like output Y3 is mainly composed of the interfering speech S7 and other noises. The extraction result also verifies that the validity of the proposed method in noisy cocktail-party environments.

The good performance of the proposed method in such time-varying environments is due to two reasons. First, fixed beamforming can enhance target signals even in time-varying environments. Second, the spectral components of the target and (moving or static) interfering signals are still independent after beamforming; besides, the target signal becomes dominant in the beamformed output. This helps the subsequent blind source separation.

The proposed method is under the assumption that the target source stays in a fixed position. For a moving target, it is possible that time-varying beamforming and sample-by-sample blind source separation algorithms are better choices. This can be a topic for future research.

CONCLUSIONS

It is challenging to extract a target speech in a time-varying, noisy, and reverberant environments. Emulating the human auditory system, the paper proposes a target speech extraction method for such a difficult condition by combining beamforming and blind source separation. The proposed method integrates the advantages of both technologies and complements their weakness. In addition, a special beamforming processing style is employed to deal with time-varying environments. Simulations verify that, the proposed method performs well in a time-varying cocktail-party-like situation where any of the two methods alone fails to work efficiently.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (60772161, 60372082) and the Specialized Research Fund for the Doctoral Program of Higher Education of China (200801410015). This work was also supported by NRC-MOE Research and Post-doctoral Fellowship Program from Ministry of Education of China and National Research Council of Canada. The authors gratefully acknowledge stimulating discussions with Dr. Michael R. Stinson and Dr. David I. Havelock from Institute for Microstructural Sciences, National Research Council Canada.

REFERENCES

[1] H. Sawada, S. Araki, R. Mukai and S. Makino, “Blind extraction of dominant target sources using ICA and time-frequency masking,” *IEEE Trans. Audio, Speech and Language Processing*, **16**(6), 2165-2173 (2006)

[2] H. N. Thi and C. Jutten, “Blind source separation for convolutive mixtures,” *Signal Processing*, **45**(2), 209-229 (1995)

[3] H. Sawada, S. Araki and S. Makino, “Frequency-domain blind source separation,” in *Blind Speech Separation*, Springer-Verlag, New York, 2007, pp. 47-78

[4] B. D. Van Veen and K. M. Buckley, “Beamforming: A versatile approach to spatial filtering,” *IEEE ASSP Magazine*, **5**, 4-24 (1988)

[5] W. Liu, S. Weiss, J. G. McWhirter and I. K. Proudler, “Frequency invariant beamforming for two-dimensional and three-dimensional arrays,” *Signal Processing*, **87**(11), 2535-2543 (2007)

[6] J. Chen, V. V. Barry and B. D. Hecox, “External ear transfer function modeling: a beamforming approach," *Journal of the Acoustical Society of America*, **92**(4), 1933-1944 (1992)

[7] Q. Pan and T. Aboulnasr, “Combined spatial/beamforming and time/frequency processing for blind source separation”, *Proceedings of the 13th European Signal Processing Conference*, Antalya, Turkey, 4-8 September 2005, pp. 1-4

[8] L. Wang, H. Ding and F. Yin, “Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals”, *EURASIP Journal on Audio, Speech, and Music Processing*, **2010**, Article ID 797962, 1-13 (2010)

[9] H. Cox, R. M. Zeskind and T. Kooij, “Practical supergain,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **ASSP-34**(3), 393-398 (1986)

[10] S. C. Douglas and M. Gupta, “Scaled natural gradient algorithms for instantaneous and convolutive blind source separation”, *Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, USA, April 2007, pp. 637-640

[11] K. Matsuoka and S. Nakashima, “Minimal distortion principle for blind source separation,” *Proceedings of the 3rd International Conference on Independent Component Analysis and Blind Source Separation* (ICA ’01), San Diego, USA, December 2001, pp. 722-727

[12] H. Sawada, S. Araki and S. Makino, “Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS,” *Proceedings of the International Symposium on Circuits and Systems* (ISCAS 2007), New Orleans, USA, May 2007, pp. 3247-3250

[13] L. Wang, H. Ding and F. Yin, “A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures”, *IEEE Transactions on Audio, Speech and Language Processing*, **19**(3), 549-557 (2011)

[14] H. Sawada, R. Mukai, S. Araki and S. Makino, “A robust and precise method for solving the permutation problem of frequency-domain blind source separation”, *IEEE Transactions on Speech and Audio Processing*, **12**(5), 530-538 (2004)

[15] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small room acoustics,” *Journal of the Acoustical Society of America*, **65**, 943-950 (1979)

Table 1. Comparison of beamforming and the proposed method in terms of signal-to-interference ratio (SIR)

Target S5	P1 (close)		P2 (far)	
RT_{60}	130 ms	300 ms	130 ms	300 ms
Input SIR	-8.2 dB	-9.1 dB	-10.7 dB	-10.8 dB
Beamforming	4.6 dB	0.6 dB	2.5 dB	-2.3 dB
Proposed method	11.9 dB	6.0 dB	9.1 dB	3.3 dB
SIR improvement	20.1 dB	15.1 dB	29.8 dB	13.5 dB

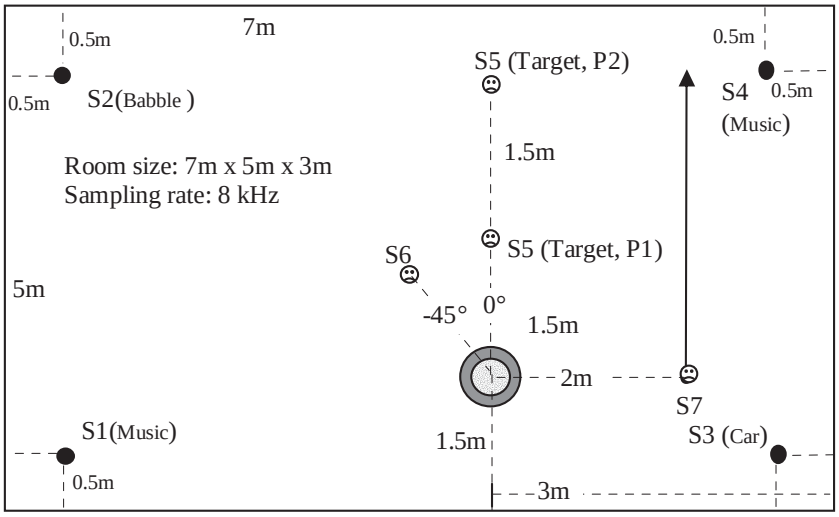

Figure 4. Simulated room environment

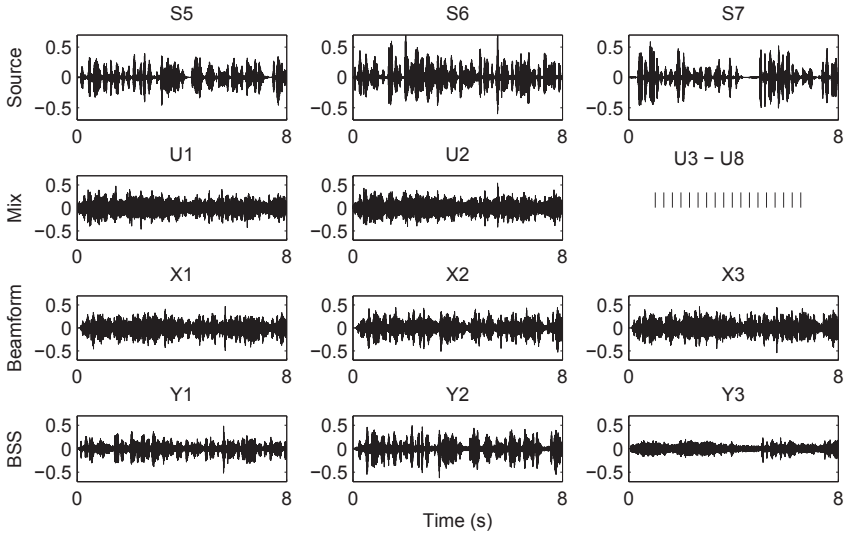

Figure 5. Waveforms at various processing stages