Nonstationary Source Separation for Underdetermined Speech Mixtures

Ryan M. Corey and Andrew C. Singer
University of Illinois at Urbana-Champaign

***Abstract*—We propose a multichannel source separation method for underdetermined mixtures of nonstationary signals, such as speech. Like other underdetermined algorithms, our method relies on the time-frequency sparsity of speech. However, our interference model allows more than one source to be active at the same time and frequency, providing better separation performance for mixtures of many sources. The system consists of several beamformers designed for different combinations of interference sources. A decision rule selects the beamformer that best suppresses the active interferers at each time-frequency point. Experiments on both simulated and real mixtures show improved interference suppression compared to conventional beamformers.**

***Keywords*—Source separation, beamforming, array processing, microphone arrays, speech enhancement, nonstationarity.**

I. Introduction

Multichannel audio source separation is a signal processing technique that isolates a single sound signal from a mixture of sources using signals from a set of spatially separated microphones. For example, source separation can be used to isolate a speech signal of interest in a noisy environment with multiple talkers, thereby improving speech intelligibility for both human listeners and machine audition algorithms. Source separation is particularly challenging in multitalker environments because the interfering speech signals are nonstationary with rapidly time-varying statistics. Most multichannel audio source separation methods [1], including adaptive techniques, are designed for stationary or slowly time-varying interference. While there have been several algorithms developed specifically for nonstationary speech sources [2]–[8], most are designed for two-microphone systems. Here, we present a multichannel source separation technique that takes advantage of the statistical properties of speech signals and is suitable for systems with many microphones and many interfering sources.

A typical source separation system is shown in Figure 1. Suppose we wish to recover one target signal from a mixture with N other sources using an array of M spatially separated sensors. The system can be modeled as an instantaneous linear mixture in the time-frequency domain:

$$\mathbf{X}(t,\omega) = \mathbf{H}(\omega)S(t,\omega) + \sum_{n=1}^{N} \mathbf{G}_n(\omega)V_n(t,\omega) + \mathbf{Z}_0(t,\omega), \quad (1)$$

This work was supported in part by Systems on Nanoscale Information fabriCs (SONIC), one of the six STARnet centers sponsored by MARCO and DARPA. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant Number DGE-1144245.

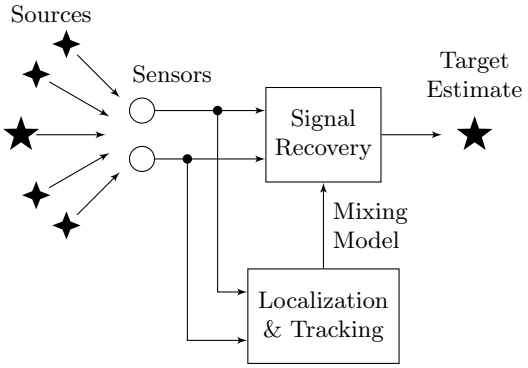

Figure 1. The source separation system attempts to recover the target source (★) from the noisy mixture of interference sources (✦).

where $\mathbf{X}$, S, V_n $(n = 1, \ldots, N)$, and $\mathbf{Z}_0$ are the short-time Fourier transforms of the M received signals at the microphones, the target source signal, the N interference signals, and diffuse additive noise, respectively, and $\mathbf{H}(\omega)$ and $\mathbf{G}_n(\omega)$ are the M-dimensional frequency-domain acoustic transfer function vectors for the target and interference sources. We wish to recover the unknown signal $S(t,\omega)$ from the observed signals $\mathbf{X}(t,\omega)$. If the transfer functions of the sources, which depend on the array geometry and the source locations, are not known, then they must first be estimated from the received data using a localization and tracking algorithm; this problem is known as blind source separation [9], [10]. In this work, we assume that the mixing parameters are known exactly and focus on the signal recovery problem. In multichannel source separation, the target signal is typically recovered using a linear estimator known as a beamformer [11]. In the time-frequency domain, the beamformer output $Y(t,\omega)$ is given by the complex weighted sum,

$$Y(t,\omega) = \mathbf{W}^H(\omega)\mathbf{X}(t,\omega), \quad (2)$$

where $\mathbf{W}^H(\omega)$ denotes the Hermitian transpose of the M-dimensional complex beamforming weight vector $\mathbf{W}(\omega)$. If the mixing system is overdetermined, meaning that $N < M$ and the diffuse noise is negligible, then we can choose weights such that $\mathbf{W}^H(\omega)\mathbf{H}(\omega) \approx 1$ and $\mathbf{W}^H(\omega)\mathbf{G}_n(\omega) \approx 0$ for all n, so that $Y(t,\omega) \approx S(t,\omega)$. If $N \geq M$, then the mixing problem is underdetermined and a linear estimator cannot simultaneously suppress all the interference sources.

In underdetermined source separation, spatial diversity alone does not provide enough information to separate the sources

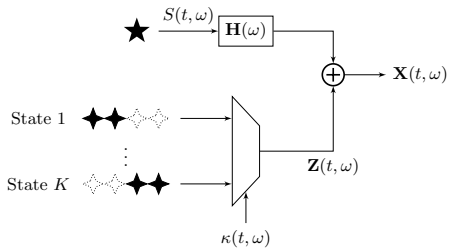

Figure 2. The nonstationary interference model. The interference is assumed to be in one of several states $\kappa(t,\omega)$ with different combinations of active (✦) and inactive (✧) sources.

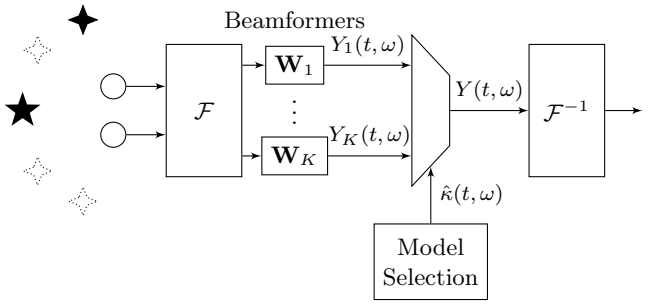

Figure 3. The input is transformed into the time-frequency domain and processed by several frequency-domain beamformers for different interference states. A classifier selects the best beamformer for each time-frequency point. The enhanced signal is then transformed back into the time domain.

and we must rely on additional assumptions about the source signal structure. Many underdetermined source separation methods take advantage of the time-frequency sparsity of speech signals. Because the frequency distribution of speech changes rapidly over time, the energy of a speech signal in the time-frequency domain is concentrated in a small fraction of (t,ω) points. Therefore, in a mixture of speech sources, there are likely only a few sources that contribute significant energy for a given (t,ω) point [12]. For small numbers of talkers, it is reasonable to assume that only one source has non-negligible energy at any given (t,ω). A number of recent algorithms, such as DUET [2], [3] and its variants [4]–[7], rely on this disjoint-sources assumption and classify each (t,ω) point as belonging to one source; the source signals are then recovered by applying a time-frequency mask to the mixture signal. These classification methods are effective for small-N mixtures, but they fail when many sources are present because the time-frequency signals are no longer disjoint. Furthermore, time-frequency masking does not take advantage of the spatial diversity available with large M.

Here, we propose a more general interference model that allows more than one source to be active simultaneously. Rather than assume only one source at each (t,ω), we assume that the interference is in one of several states at each (t,ω). The system estimates the state at each (t,ω) and the signals are recovered using a beamformer designed for that state. If the state represents an overdetermined mixture, then the beamformer can take advantage of the additional degrees of freedom to reduce residual interference and distortion. This interference model is advantageous for underdetermined mixtures because it transforms the underdetermined source separation problem into a set of overdetermined problems. Our approach is related to a recently proposed system that uses Gaussian mixture models [8] to represent interference sources. It is also conceptually similar to compressive sensing beamformers [13], which rely on spatial sparsity to achieve high spatial resolution; however, we use a different mathematical model.

II. Nonstationary Source Separation

A. Nonstationary mixing model

Many naturally occurring sounds are nonstationary. For example, the frequency distribution of speech signals changes over time as the mouth moves to form different sounds. Individual speech sounds are distinguished by their different frequency content. Because the frequency distribution of each source in a speech mixture varies over time, even if the sources themselves do not move, the spatial distribution of the received time-frequency signal vector is nonstationary. To better represent nonstationary sources, we propose the following stochastic model, shown in Figure 2, for the received mixtures in the time-frequency domain:

$$\mathbf{X}(t,\omega) = \mathbf{H}(\omega)S(t,\omega) + \mathbf{Z}(t,\omega), \quad (3)$$

$$\mathbf{Z}(t,\omega) \sim f_{\kappa(t,\omega)}(\mathbf{z};\omega), \quad (4)$$

where $\mathbf{H}(\omega)$ is known and fixed, $S(t,\omega)$ is the unknown signal of interest, and each $\mathbf{Z}(t,\omega)$ is a random variable that follows one of K known spatial distributions $f_1,\ldots,f_K$. Each distribution represents the combined effect of a set of active interferers and diffuse noise on the signals received by the microphone array. The true state at a given time and frequency is denoted by the latent variable $\kappa(t,\omega)\in\{1,\ldots,K\}$.

The choice of the states and corresponding distributions is critical to the success of the proposed method. In the implementation presented here, each state k corresponds to a different subset $\mathcal{V}_k$ of active sources, plus diffuse additive noise, so that the total interference and noise is given by

$$\mathbf{Z}(t,\omega) = \sum_{n\in\mathcal{V}_{\kappa(t,\omega)}} \mathbf{G}_n(\omega)V_n(t,\omega) + \mathbf{Z}_0(t,\omega). \quad (5)$$

Under this model, the sources not in the subset are assumed to contribute negligible energy to the mixture. If $|\mathcal{V}_k| < M$, then the mixing problem for state k is instantaneously overdetermined and the sources can be effectively separated. This assumption is reasonable for speech mixtures, in which a few sources are typically much stronger than the others at any given time-frequency point. The diffuse noise component $\mathbf{Z}_0$ represents noise sources with full-rank spatial covariance matrices, such as internal sensor noise. In a practical system the distributions would be chosen and updated online using a localization algorithm; for the purposes of this work, however, they are assumed to be known and fixed.

B. Time-varying beamformer

To remove the time-varying interference and recover the target signal, we propose a time-varying beamformer, shown in Figure 3. Each state k has a corresponding set of beamforming weights $\mathbf{W}_k(\omega)$. The beamformer outputs are given by

$$Y_k(t,\omega) = \mathbf{W}_k^H(\omega)\mathbf{X}(t,\omega), \tag{6}$$

for $k = 1,\ldots,K$. A decision rule selects the state index $\hat{\kappa}(t,\omega)$ to use at each time-frequency point. The recovered time-frequency signal is then $Y(t,\omega) = Y_{\hat{\kappa}(t,\omega)}(t,\omega)$.

To prevent audible distortion in the reconstructed time-domain signal due to the rapidly time-varying filtering process, it is important to impose constraints on the filter coefficients. For example, we can impose a distortionless constraint so that $\mathbf{W}_k^H(\omega)\mathbf{H}(\omega) = 1$ for all k. Then the output is

$$Y(t,\omega) = S(t,\omega) + \mathbf{W}_{\hat{\kappa}(t,\omega)}^H(\omega)\mathbf{Z}(t,\omega) \tag{7}$$

and only the residual interference is distorted.

C. Sample implementation

To assess the performance of the proposed source separation method, we use the nonstationary interference model of (5) with distributions corresponding to every set of up to p active interferers. The diffuse noise component is assumed to be white, i.e., to have a diagonal spatial covariance matrix.

The beamforming weights are computed according to the well-known minimum variance distortionless response (MVDR) formula [11],

$$\mathbf{W}_k(\omega) = \frac{\Sigma_k^{-1}(\omega)\mathbf{H}(\omega)}{\mathbf{H}^H(\omega)\Sigma_k^{-1}(\omega)\mathbf{H}(\omega)}, \tag{8}$$

where $\Sigma_k(\omega) = \mathrm{Cov}[\mathbf{Z}(t,\omega)|\kappa(t,\omega) = k]$. These weights satisfy the distortionless constraint to ensure that the target signal is not distorted.

Because the MVDR weights yield unbiased estimates of the target in zero-mean noise, if we wish to minimize the squared error of the reconstructed signal, we should select the beamformer output with the smallest variance. Since the variance cannot be measured directly, the selection rule chooses the beamformer output with the smallest energy:

$$\hat{\kappa}(t,\omega) = \arg\min_k |Y_k(t,\omega)|^2.$$

Because it fully computes the output of every beamformer, the computational complexity of this implementation is roughly K times larger than that of a single beamformer.

III. Experimental Results

The nonstationary source separation method was evaluated using both simulated and real-world mixtures. In both experiments, the speech signals were processed at 16 kHz with a frame size of 1024 samples, frame spacing of 512 samples (50% overlap), DFT length of 2048 samples (2x frequency oversampling), and highpass pre-emphasis filter. Figures 4, 5, and 6 compare results from four source separation methods: a data-independent beamformer, $\mathbf{W} = \mathbf{H}/\|\mathbf{H}\|^2$, which ignores the spatial distribution of the interference; a conventional MVDR beamformer, which uses a single stationary covariance model; the nonstationary source separation method described in Section II-C with $p = 1$; and the nonstationary method with $p = 2$. Simulations were also performed with $p = 3$, but the results are nearly identical to $p = 2$ and they are not shown in the plots. The signal-to-interference-plus-noise (SINR) ratio was computed in the time domain using the source image at the first microphone, $(h_1 \star s)(t)$, as the reference signal. To compute the average SINR, the source separation method was repeated with each of the $N+1$ sources as the target and the results were averaged over all sources.

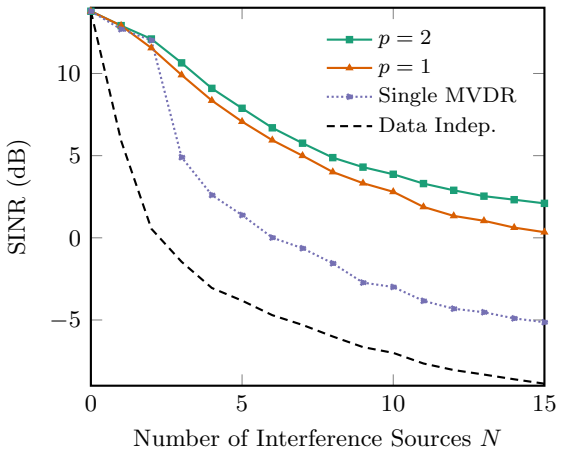

Figure 4. Mean SINR performance over all target sources of conventional and proposed source separation methods for simulated mixtures with $M = 4$.

A. Simulated mixtures

To evaluate the performance scaling of the proposed method with the number of sensors, M, and interferers, N, we used artificial mixtures generated from simulated room impulse responses [14]. The simulated microphones were linear isotropic sensors placed in a circular array of radius 10 cm. The sources were distributed in an approximately circular pattern several meters away. The source signals were sixty-second speech clips from the TSP speech dataset [15].

Figure 4 shows the SINR as a function of N for an array of four microphones. For $N = 0$ (noise only), all methods are identical. For $N = 1$ and 2, where the mixture is overdetermined, there is little difference between the stationary and nonstationary methods. However, for underdetermined mixtures, the performance of the conventional MVDR beamformer drops rapidly with increasing N, while the performance of proposed method degrades more gradually. In this scenario, the time-varying beamformer provides a 5–7 dB advantage over the fixed beamformer for underdetermined mixtures.

Figure 5 shows the SINR as a function of M for a set of five sources, one target source and four interference sources. Because the aperture of the circular array is fixed, there is a ceiling on the achievable SINR of the array even with many sensors. The proposed method reaches this ceiling

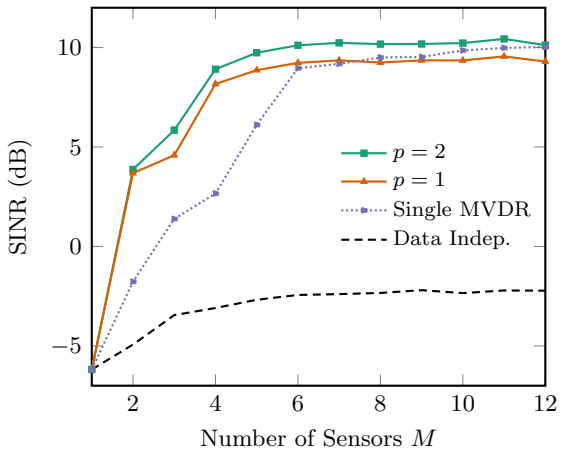

Figure 5. Mean SINR performance over all target sources of conventional and proposed source separation methods for simulated mixtures with $N = 4$.

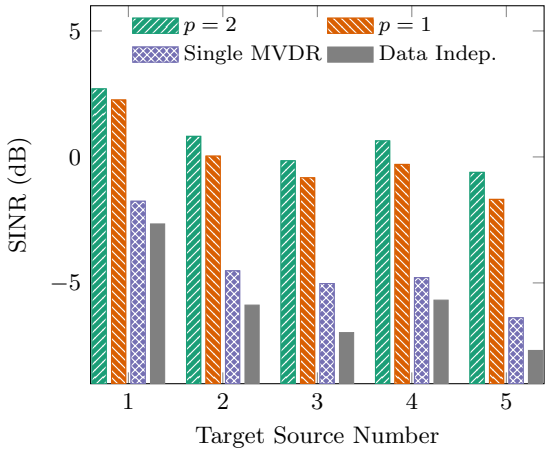

Figure 6. SINR performance of the conventional and proposed source separation methods for recorded speech mixtures. For each target source, the remaining four sources are treated as interference.

using fewer sensors than the conventional beamformer. The time-varying beamformer achieves a higher SINR than the conventional beamformer for underdetermined mixtures ($M < N$). For slightly overdetermined mixtures, the time-varying beamformer still outperforms the fixed beamformer because it can devote more degrees of freedom to suppressing side lobes. For strongly overdetermined mixtures, $M \gg N > p$, the MVDR beamformer performs slightly better because the tertiary interference sources are stronger than the diffuse noise.

B. Recorded mixtures

The proposed method was also evaluated using real-world data. The recordings were made in a conference room using the four-microphone array of a Microsoft Kinect sensor. The five talkers sat around a conference table and read aloud from newspapers and other written materials. The sources were recorded individually and later superimposed to form the mixtures. Figure 6 shows the SINR achieved by the conventional and proposed source separation methods. The SINR varies between sources because of their relative placement. The nonstationary source separation method improved performance by around 5 dB compared to the single beamformer. Figure 7 shows the spectrogram of one of the sources, the noisy mixture, and the outputs of the conventional and proposed source separation methods. The time-varying beamformer removes more of the interference than the conventional beamformer.

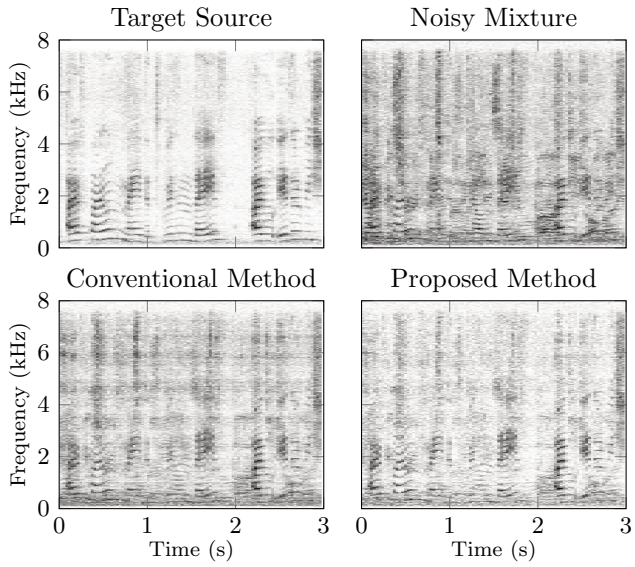

Figure 7. Time-frequency spectrograms of target, mixture, and separated speech signals from multichannel recordings. Darker points correspond to larger energy.

IV. Conclusions

The experimental results show that a nonstationary interference model and time-varying beamformer can achieve stronger separation performance than a conventional stationary model and fixed beamformer for mixtures of speech signals. By taking advantage of the sparsity of speech and other natural signals, the proposed method can better allocate its limited degrees of freedom to best suppress the interference sources that are most active at each time and frequency.

This performance improvement comes at a cost of increased computational complexity. It is also more sensitive to modeling errors: while in this work we assumed perfect knowledge of the acoustic transfer functions, a realistic implementation would have to infer those parameters and protect against estimation errors. Further work is required to develop such interference models, beamforming weights, and model selection rules that are computationally efficient and robust to modeling errors. With further refinement, this method can improve source separation performance in challenging acoustic environments with many nonstationary interference sources.

REFERENCES

[1] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer, 2013.

[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] S. Rickard, "The duet blind source separation algorithm," in *Blind Speech Separation*, pp. 217–241, Springer, 2007.

[4] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[5] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l 1-norm minimization," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 81–81, 2007.

[6] M. Kühne, R. Togneri, and S. Nordholm, "A novel fuzzy clustering algorithm using observation weighting and context information for reverberant blind speech separation," *Signal Processing*, vol. 90, no. 2, pp. 653–669, 2010.

[7] J. Traa and P. Smaragdis, "Multichannel source separation and tracking with ransac and directional statistics," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 22, no. 12, pp. 2233–2243, 2014.

[8] M. A. Dmour and M. Davies, "A new framework for underdetermined speech extraction using mixture of beamformers," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 3, pp. 445–457, 2011.

[9] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*. Springer, 2007.

[10] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "Convolutive blind source separation methods," in *Springer Handbook of Speech Processing*, pp. 1065–1094, Springer, 2008.

[11] H. L. Van Trees, *Optimum array processing*. Wiley, 2004.

[12] S. Rickard and Ö. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *IEEE Conf. on Acoustics, Speech, and Signal Process.*, vol. 1, pp. 529–532, 2002.

[13] A. C. Gürbüz, J. H. McClellan, and V. Cevher, "A compressive beamforming method," in *IEEE Conf. Acoustics, Speech and Signal Process.*, pp. 2617–2620, 2008.

[14] E. A. Lehmann and A. M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 6, pp. 1429–1439, 2010.

[15] P. Kabal, "Tsp speech database," Telecommunications and Signal Processing Lab, McGill University, 2002.