2016 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 13–16, 2016, SALERNO, ITALY

A HYPOTHESIS TESTING APPROACH FOR REAL-TIME MULTICHANNEL SPEECH SEPARATION USING TIME-FREQUENCY MASKS

Ryan M. Corey and Andrew C. Singer

University of Illinois at Urbana-Champaign

ABSTRACT

We propose a new approach to time-frequency mask generation for real-time multichannel speech separation. Whereas conventional approaches select the strongest source in each time-frequency bin, we perform a binary hypothesis test to determine whether a target source is present or not. We derive a generalized likelihood ratio test and extend it to underdetermined mixtures by aggregating the outputs of several tests with different interference models. This approach is justified by the nonstationarity and time-frequency disjointedness of speech signals. This computationally simple method is suitable for real-time source separation in resource-constrained and latency-critical applications.

1. INTRODUCTION

We consider the problem of separating a target speech source from a noisy mixture. High-quality source separation can improve intelligibility in noisy environments and would be beneficial in real-time audio enhancement applications, such as digital hearing aids. While there have been many recent advances in multichannel source separation, most modern algorithms are too computationally complex to run in real time on embedded devices [1]. Due to size, power, and latency constraints, most listening devices rely on simple and computationally inexpensive beamforming and filtering techniques [2]. In this paper, we seek a low-latency technique for embedded multichannel speech separation.

Speech signals from N sources received by an array of M microphones can be modeled as a convolutive mixture,

$$x_m(t) = \sum_{n=1}^{N} \left(h_{mn} \star s_n\right)(t) + z_m(t), \tag{1}$$

for $m = 1, \ldots, M$, where $x_m(t)$ is the signal at microphone m, $s_n(t)$ is a source signal, $h_{mn}(t)$ is the impulse response between source n and microphone m, and $z_m(t)$ is additive noise. We can write (1) as an instantaneous mixture by taking the short-time Fourier transform (STFT) of the signals. Then

$$\mathbf{X}(\tau,\omega) = H(\omega)\,\mathbf{S}(\tau,\omega) + \mathbf{Z}(\tau,\omega), \tag{2}$$

where $\mathbf{X}(\tau,\omega) \in \mathbb{C}^M$, $\mathbf{S}(\tau,\omega) \in \mathbb{C}^N$, and $\mathbf{Z}(\tau,\omega) \in \mathbb{C}^M$ are complex vectors containing the STFT coefficients at time index τ and frequency ω of the mixtures, sources, and noise, repectively, and $H(\omega) \in \mathbb{C}^{M\times N}$ is the mixing matrix. When the STFT is computed using the discrete Fourier transform, each sample index (τ,ω) is known as a time-frequency (T-F) bin. In the source separation problem, we wish to estimate one or more components of the unknown signal vector $\mathbf{S}(\tau,\omega)$ for each T-F bin. If the mixing parameters, such as H and the distribution of $\mathbf{Z}$, were unknown, they must be estimated from the observed data using blind source separation (BSS) methods [3]. Once $\mathbf{S}$ has been estimated, the time-domain signal can be reconstructed using the inverse STFT.

Blind source separation can be divided into two tasks: localization, in which the unknown mixing parameters are estimated, and signal recovery, in which the signal of interest is extracted from the mixture. There are many localization methods designed for different types of mixing problems. If the array configuration and room acoustics were kown, then H could be computed analytically as a function of the source locations. If the matrices were unknown but $M \geq N$ so that each $H(\omega)$ had full column rank, then the matrices could be estimated directly from the data using independent component analysis (ICA). If $M < N$ so that the mixing problem is underdetermined, then we cannot separate the signals using spatial diversity alone. Fortunately, speech signals are sparse in time-frequency. Thus, in any given T-F bin, there are often fewer than M *active* sources [4]. For small numbers of talkers, it is often reasonable to assume that only one source has non-negligible energy in each T-F bin. A number of recent algorithms [5–14] separate sources by clustering the T-F bins according to their active sources. These algorithms can be distinguished by the features they use for classification. The Degenerate Unmixing Estimation Technique (DUET) [7, 8] for closely spaced microphone pairs clusters sources based on interchannel phase differences (IPD) and interchannel level differences (ILD). It can be modified for widely spaced arrays by explicity modeling spatial aliasing [9, 10] and for more

This work was supported in part by Systems on Nanoscale Information fabriCs (SONIC), one of the six SRC 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.

978-1-5090-0746-2/16/$31.00

than two microphones by using subspace techniques [11] and pairwise IPD and ILD features [12, 13]. In reverberent environments, the clustering must be performed separately in each frequency band [14].

Once the mixing parameters have been estimated, they can be used to recover the signal of interest from the mixture. The classical recovery method is beamforming: the microphone signals are filtered and summed to form a linear estimate of the target. The commonly used minimum variance distortionless response (MVDR) beamformer [15], which has unity gain in the direction of the target and minimizes the output power elsewhere, is given by

$$\hat{S}_{\mathrm{MVDR}}(\tau,\omega) = \frac{\mathbf{h}_t^*(\omega)\Sigma^{-1}(\omega)\mathbf{X}(\tau,\omega)}{\mathbf{h}_t^*(\omega)\Sigma^{-1}(\omega)\mathbf{h}_t(\omega)}, \tag{3}$$

where $\hat{S}_{\mathrm{MVDR}}(\tau,\omega) \in \mathbb{C}$ is the estimate of the target source signal, $\mathbf{h}_t(\omega) \in \mathbb{C}^M$ is the steering vector (column of H) for the target source, and $\Sigma(\omega) \in \mathbb{C}^{M\times M}$ is the covariance matrix of the combined noise and interference. If the interference and noise are normally distributed, then (3) is the maximum likelihood estimate of the target signal [15]. A beamformer with $M \geq N$ can effectively align its nulls over the interfering sources to suppress them.

The MVDR and similar beamformers are designed for stationary signals. Speech singals, however, are highly nonstationary: the signal statistics change over time as the talker produces different speech sounds. To separate speech signals, we can take advantage of their time-frequency sparsity by applying a binary filter known as a T-F mask: the T-F bins in which the target source is considered active are retained and the rest are discarded. Applying a mask $\delta(\tau,\omega) \in \{0,1\}$ to the signal from the first microphone gives the estimate

$$\hat{S}_{\mathrm{mask}}(\tau,\omega) = X_1(\tau,\omega)\delta(\tau,\omega). \tag{4}$$

Because only a fraction of the bins contain useful speech information, and because of the perceptual properties of the human auditory system, a simple binary mask can be effective in improving intelligibility [16]. Masks are especially useful in underdetermined mixtures ($M < N$) and have typically been applied to one- or two-microphone systems, but they can also be beneficial in large-M systems, either alone or as a postprocessing stage [5, 6].

Localization methods are typically computationally demanding and require large blocks of samples for accurate performance, while signal recovery is computationally simple and has lower latency. In offline speech enhancement, localization and recovery are often performed jointly. In real-time applications, however, it is beneficial to separate the two tasks, as shown in Figure 1. Signal recovery, in this case accomplished using a mask, is applied immediately using parameters supplied by the localization block. The localization algorithm, unconstrained by latency, can use more data and computational resources; it may even run on a separate device. In this paper, we focus on the signal recovery task. We propose a masking method for low-latency recovery of speech signals given an accurate estimate of the mixing parameters.

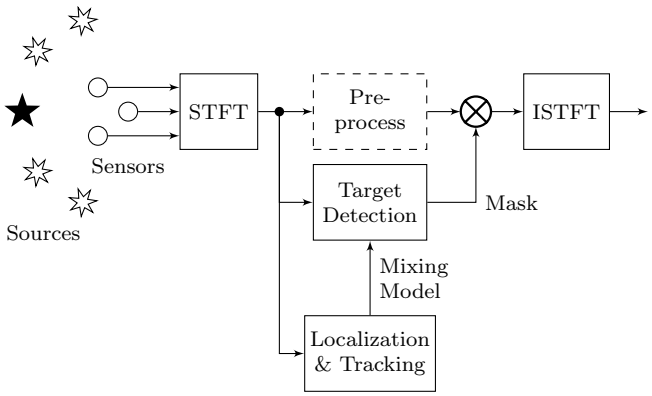

Fig. 1: The proposed system recovers the target source using a T-F mask. The mask is generated by a low-latency decision rule using model parameters from a higher-latency localization algorithm.

Source separation systems that recover signals using masks typically use clustering-based localization algorithms, such as DUET, and then classify each T-F bin as belonging to one source. These algorithms assume that exactly one source is active in each T-F bin. Here, we propose a novel mask generation strategy that uses hypothesis testing rather than classification. That is, instead of asking "Which source is strongest at (τ,ω)?", we ask "Is the target source active at (τ,ω)?" Because we explicitly model the presence of simultaneous interfering signals, our method can be applied to both over- and underdetermined mixtures with arbitrary numbers of sources, N, and sensors, M. In this paper, we first introduce the hypothesis testing framework for stationary and overdetermined mixtures. We relate the log-likelihood statistic to the output signal-to-noise ratio of the MVDR beamformer and show how it can be used to trade off interference for distortion. We then modify the method for nonstationary sparse signals and underdetermined mixtures using a multiple-model hypothesis test. Finally, we present experimental results from real recordings.

2. SIGNAL DETECTION FOR STATIONARY MIXING MODELS

To motivate the hypothesis testing approach, we first consider a stationary model. Let $S_t(\tau,\omega)$ be the unknown target signal and let $\mathbf{h}_t(\omega)$ be its known steering vector. The mixture is

$$\mathbf{X}(\tau,\omega) = \mathbf{h}_t(\omega)S_t(\tau,\omega) + \mathbf{Z}(\tau,\omega), \tag{5}$$

where $\mathbf{Z}(\tau,\omega)$ is a complex random vector with zero mean and nonsingular covariance matrix $\Sigma_{\mathbf{Z}}(\omega)$ that models the interference sources, diffuse noise, and sensor noise. The steering vectors and covariance matrices are different in each frequency band, but are assumed to be constant over the time interval of interest. In a practical system, these parameters

would be estimated by the source localization block and updated periodically as the sources or microphones move. For the remainder of the paper, we omit the (τ, ω) notation; each expression is applied separately to each T-F bin using the mixing parameters for the corresponding frequency band.

Our goal is to detect whether the signal is present ($S_t \neq 0$) or not present ($S_t = 0$). That is, we are testing between the two hypotheses:

$$\mathcal{H}_1 : \quad \mathbf{X} = \mathbf{h}_t S_t + \mathbf{Z} \tag{6}$$

$$\mathcal{H}_0 : \quad \mathbf{X} = \mathbf{Z}. \tag{7}$$

Problems of this form, known as noncoherent signal detection, are commonly solved with a generalized likelihood ratio test (GLRT), which treats S_t as a nonrandom parameter [15]. The test statistic, $T(\mathbf{X})$, is given by the log-likelihood ratio

$$T(\mathbf{X}) \quad = \quad \ln \frac{\sup_{S_t \neq 0} P\left(\mathbf{X} \mid S_t\right)}{P\left(\mathbf{X} \mid S_t = 0\right)}. \tag{8}$$

The binary decision rule is

$$\delta\left(\mathbf{X}\right) = \begin{cases} 1, & \text{if } T(\mathbf{X}) > \gamma \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

where γ is a tunable parameter that will be discussed later.

The test statistic can be computed by substituting the maximum likelihood estimate of S_t into the likelihood function. If $\mathbf{Z}$ is Gaussian, then the estimate is given by (3) and the ratio, after dropping a factor of $1/2$, reduces to

$$T(\mathbf{X}) = \frac{\left|\mathbf{h}_t^* \Sigma_{\mathbf{Z}}^{-1} \mathbf{X}\right|^2}{\mathbf{h}_t^* \Sigma_{\mathbf{Z}}^{-1} \mathbf{h}_t}. \tag{10}$$

Under the stochastic model (5) with Gaussian noise, the random variable $2T(\mathbf{X})$ has a noncentral chi-squared distribution with two degrees of freedom. The probability of correct detection is

$$\mathrm{P_D}(\delta) \quad = \quad P\left(T(\mathbf{X}) > \gamma \mid \mathcal{H}_1\right) \tag{11}$$

$$\quad = \quad F_2\left(2\gamma, 2\bar{T}(S_t)\right), \tag{12}$$

where $F_2\left(\cdot; v\right)$ is the complementary cumulative distribution function for the noncentral chi-squared distribution with two degrees of freedom and noncentrality parameter v, and

$$\bar{T}(S_t) \quad = \quad \left|S_t\right|^2 \mathbf{h}_t^* \Sigma_{\mathbf{Z}}^{-1} \mathbf{h}_t \tag{13}$$

$$\quad = \quad \frac{\left|S_t\right|^2}{\mathrm{Var}\left(\hat{S}_{\mathrm{MVDR}}\right)}. \tag{14}$$

The probability of correct detection increases monotonically with $\bar{T}(S_t)$ and decreases with γ. The probability of false alarm is

$$\mathrm{P_F}\left(\delta\right) \quad = \quad P\left(T(\mathbf{X}) > \gamma \mid \mathcal{H}_0\right) \tag{15}$$

$$\quad = \quad e^{-\gamma}. \tag{16}$$

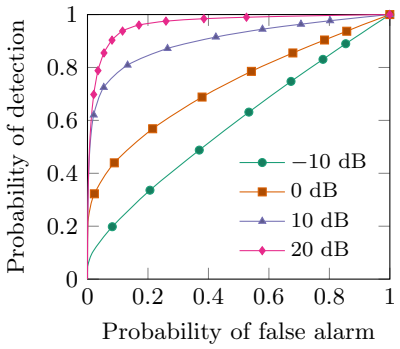

Fig. 2: Experimental ROC curves for detection of a speech signal in white noise with various overall SNRs.

In hypothesis testing problems, the tradeoff between $\mathrm{P_F}$ and $\mathrm{P_D}$ is expressed by a receiver operating characteristic (ROC) curve parametrized by γ. Figure 2 shows a set of experimental ROC curves for detecting speech in artifical white noise. Like all the experimental results in this paper, the signals were recorded at 16 kHz and the STFT used a window size of 1024 samples and a step size of 256 samples. The curves show the average detection and false alarm rates over all T-F bins for several additive noise levels. The ground truth mask is 1 for bins with instantaneous power greater than the average signal power at the same frequency. As expected, the performance of the detector improves with the overall SNR.

Because (16) depends only on γ and not on the data, we can select γ based on a desired false alarm rate. The probability of correct detection (12) is then determined by $\bar{T}(S_t)$. The test can also be interpreted in terms of the signal power: $T(\mathbf{X})$ is an estimate of the instantaneous signal-to-noise ratio (SNR) at the output of an MVDR beamformer and γ is an SNR threshold. If the system can fully suppress interference, then $T(\mathbf{X})$ is proportional to the target source power, *independent of the interfering signals*. Thus, γ determines a power cutoff. The rule resembles power-based voice activity detectors, e.g. [17], which are often used in speech enhancement. Smaller γ preserve more of the target signal energy, but may also preserve more components of interfering signals. Larger γ better isolate the target signal from noise and interference, but can harm intelligibility by removing speech features. Thus, the parameter can be tuned to trade off between interference and distortion. Figure 3 shows the fraction of T-F bins preserved and the energy remaining in those bins as a function of γ for recorded speech with an overall SNR of 30 dB, along with spectrograms of the masked signals. Based on informal listening tests, the speech quality is comparable to the original with 18% of the bins preserved (top right inset) and is degraded but still intelligible with about 1% of the bins (bottom right inset). In our experiments, we found that a reasonable starting value of γ is the average output SNR for the speech signal: a bin is labeled active if its instantaneous power is greater than the average power at that frequency.

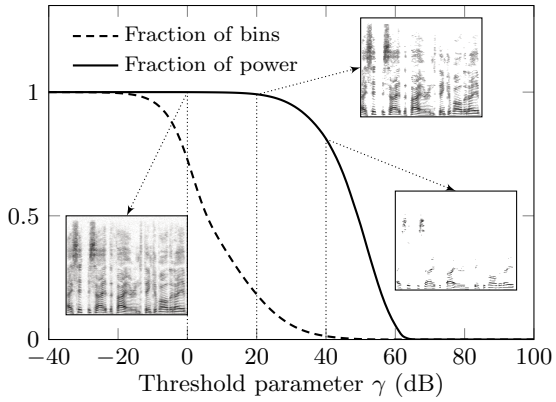

Fig. 3: The curves show the fraction of bins with instantaneous SNR greater than γ and the fraction of power in those bins for a recording with overall SNR 30 dB. The spectrograms show the masked signals for selected γ.

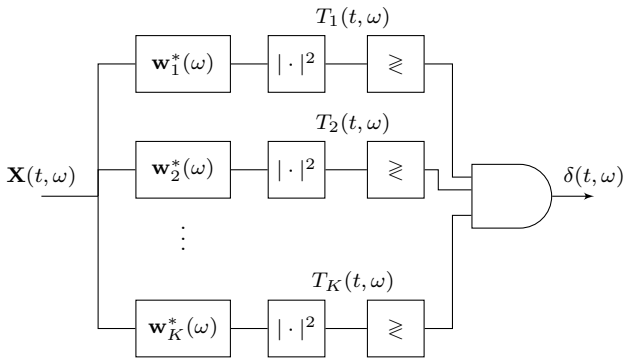

Fig. 4: The proposed signal detection rule aggregates the decisions of a set of likelihood ratio tests based on different interference models. The $\mathbf{w}_k$'s are weights that generate the test statistic (18).

3. SIGNAL DETECTION FOR SPARSE MIXTURES

The stationary model is appropriate when the noise and interference are stationary; speech signals, however, are nonstationary. If the interference consists primarily of speech or other sparse signals, then we can exploit that sparsity to improve detection performance in underdetermined mixing problems. Instead of a single stationary model, we assume that the system is described by one of K models,

$$\mathbf{X} = \mathbf{h}_t S_t + \mathbf{Z}^{(k)}, \tag{17}$$

for $k = 1, \ldots, K$, where $\mathbf{Z}^{(k)}$ has covariance matrix Σ_k. For the experiments in this paper, we assume at most one active interferer in each T-F bin, so that $K = N - 1$ and $\Sigma_k = \sigma_k^2 \mathbf{h}_k \mathbf{h}_k^* + \Sigma_0$, where $\mathbf{h}_k$ is the steering vector of interference source k, σ_k^2 is the interference power, and Σ_0 is the covariance of the stationary noise component. Thus, for each model, we are comparing the hypotheses:

$\mathcal{H}_{1,k}$: Both S_t and interference source k are active.

$\mathcal{H}_{0,k}$: Only interference source k is active.

More generally, the models might correspond to different interference subspaces rather than individual sources. It is straightforward to extend the analysis to these models.

The test statistic for each pair of hypotheses is analagous to (10) and is given by

$$T_k(\mathbf{X}) = \frac{\left|\mathbf{h}_t^* \Sigma_k^{-1} \mathbf{X}\right|^2}{\mathbf{h}_t^* \Sigma_k^{-1} \mathbf{h}_t} \tag{18}$$

for $k = 1, \ldots, K$. The noncentrality parameter is

$$\bar{T}_k(S_t) = |S_t|^2 \, \mathbf{h}_t^* \Sigma_k^{-1} \mathbf{h}_t, \tag{19}$$

which represents the output SNR of the beamformer when interference source k is present. The performance of the test depends on the relationships between the signal subspace, the assumed interference subspace, and the true interference. If $\mathbf{Z}$ is strongly correlated with the target steering vector but not with the assumed interference subspace, that is, if the interference is closer to the target than expected, then the test is likely to generate a false positive. To prevent excessive false positives, the aggregate decision rule uses the most conservative test statistic to make its decision:

$$\delta(\mathbf{X}) = \begin{cases} 1, & \text{if } \min_k T_k(\mathbf{X}) > \gamma \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

Equivalently, $\delta(\mathbf{X}) = 1$ only if $T_k(\mathbf{X}) > \gamma$ for all $k = 1, \ldots, K$. Thus, the rule is the product of the outputs of K parallel hypothesis tests, as shown in Figure 4.

The conservative decision rule helps to prevent false positives. However, if $\mathbf{h}_t$ is strongly correlated with $\mathbf{h}_k$, then $T_k(\mathbf{X})$ will be small and false negatives will be more likely. To achieve a high $\mathrm{P_D}$, the system should satisfy

$$\min_k |S_t|^2 \, \mathbf{h}_t^* \Sigma_k^{-1} \mathbf{h}_t \gg \gamma. \tag{21}$$

This condition shows how the performance of the hypothesis test relates to the parameters of the speech separation problem. We have already shown how γ can be used to trade off interference for distortion. For a fixed γ, the quality of the separation mask can be further improved by:

1. Increasing the source power (larger $|S_t|^2$),
2. Decreasing the interference power (smaller Σ_k),
3. Adding more microphones (larger $\|\mathbf{h}_t\|$),
4. Moving the interference farther from the target (smaller inner product of $\mathbf{h}_k$ and $\mathbf{h}_t$), or
5. Allowing sources close to the target to be included in the output (removing hypotheses with small $\bar{T}_k$'s).

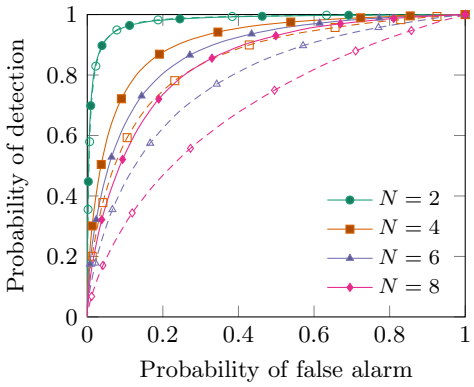

Fig. 5: ROC curves for widely spaced sources. The dashed and solid curves show the ROC for the stationary and multiple-model detectors, respectively.

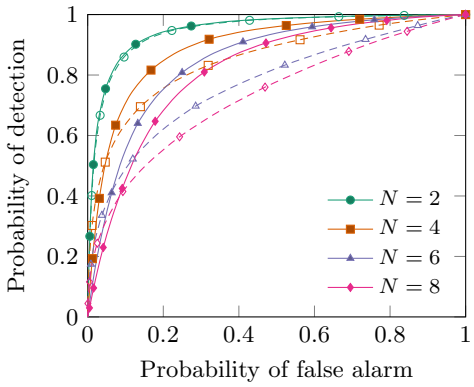

Fig. 6: ROC curves for closely spaced sources. The dashed and solid curves show the ROC for the stationary and multiple-model detectors, respectively.

Note that the total number of interference sources does not directly affect the separation performance as long as fewer than M are active within each time-frequency bin and the Σ_k's can be accurately estimated; however, more complex interference scenarios are more difficult to estimate.

4. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed detection strategy, we applied it to data recorded in a conference room ($T_{60} \approx 300$ ms) from eight talkers seated around a table. The audio was recorded by a Microsoft Kinect with an array of $M = 4$ microphones positioned at the head of the table. The speakers were recorded individually reading aloud from the Daily Illini newspaper and other sources. The separate source recordings were used to form a least squares estimate of the steering vector for each source, then combined to form the test mixtures. The background noise covariance was estimated from a recording with no speech. These measured steering vectors and noise covariances were supplied to the hypothesis test in place of the parameters that would be estimated by a localization algorithm in a real system. Artificial white noise was added to the mixed signals to give an overall SNR of about 16 dB. The ground truth mask used to calculate $\mathrm{P_F}$ and $\mathrm{P_D}$ is 1 for bins with power greater than the average source power. This mask retains about 10% of the time-frequency bins and 97% of the signal energy.

The separation masks were generated using the conventional GLRT of Section 2 and the multiple-model detector of Section 3. Figure 5 shows the ROC curves for detecting a female speaker sitting close to the array with a variable number of widely spaced interference sources. Figure 6 shows the ROC curves for detecting a male speaker sitting far from the array with a variable number of closely spaced interference sources. Both detectors perform slightly worse for the closely spaced interference sources and faraway target. The two rules are identical at $N = 2$ with only one interference source. For $N > 2$, the multiple-model detector has a clear advantage. It uses the signal's T-F sparsity to produce higher beamforming gain and more accurate detection results. Both detectors have decreasing performance with larger N, but the multiple-model detector's performance degrades more slowly.

Figure 7 shows the target, mixture, and masked spectrograms for the female target source and four widely separated interference sources. For comparison, two classification-based masks were also generated. The oracle classifier assigns each bin to the source with the largest power. The directional classifier assigns the bins based on the correlation of the microphone signals with the source steering vectors. As expected, the classifier masks are effective at removing interference but are more sensitive to noise. The rapid time variation of the classifier masks also produces distortion in the reconstructed audio signals. The hypothesis testing masks are less effective at removing interference but more closely match the shape of the clean target signal. The multiple-model detector produces a denser mask than the conventional GLRT detector at a given threshold since it more accurately estimates the instantaneous output SNR.

5. CONCLUSIONS

We have shown that a binary hypothesis test can be used to generate time-frequency masks for noisy speech mixtures. The hypothesis testing approach is fundamentally different from conventional classification methods: the masks show whether the target source is active or not, rather than which source is strongest in a particular T-F bin. A classifier mask would fail to include an important speech feature if there were a stronger overlapping interference signal. On the other hand, classifier masks are better at excluding strong interference. Thus, a hypothesis testing mask is best used in conjunction with another separation technique, such as a beamformer. Because hypothesis testing is based on the target source power, it does not require that the signals be strictly disjoint and is therefore effective for mixtures with large N. Since its performance depends on the achievable beamforming gain, the

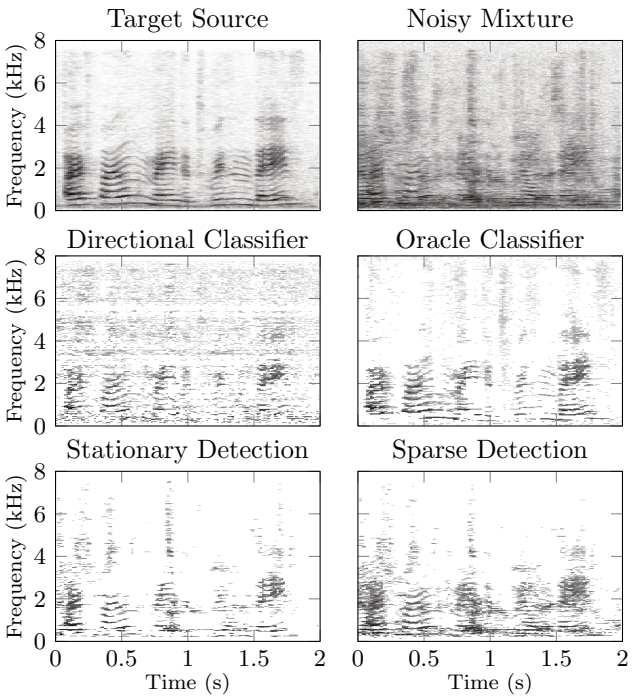

Fig. 7: Spectrograms for the target, mixture, and masked signals with $M = 4$, $N = 5$, and $\gamma = 16$ dB.

test also scales well with large M.

The likelihood ratio test presented here is well suited to the separation of speech mixtures. The tuning parameter, γ, controls the tradeoff between false negatives and false positives or, equivalently, between interference and distortion. Because most of the perceptually relevant information in a speech signal is concentrated in a few high-energy T-F bins, the detection threshold can be tuned to a high level in difficult separation environments and still produce an intelligible output signal. Furthermore, the multiple-model detection rule explicitly models the sparsity of speech to improve performance in underdetermined mixtures, providing significant benefit over stationary models. Further analysis is required to select the best set of models for a given interference scenario.

The computation is dominated by the inner product used to produce the test statistic. If the STFT uses 50% overlap between frames, then the number of T-F bins is equal to the number of samples and the detection rule requires MK complex multiply-accumulate operations per sample period. As shown in Figure 4, the test has a highly parallel structure; furthermore, both the detection rule and the mask are applied independently in each frequency band. Thus, the system can be implemented in a low-latency parallel architecture. The latency of the T-F mask is determined by the STFT frame length. The low latency and modest computation of the proposed method make it suitable for real-time embedded speech enhancement systems.

The detection rule proposed here is not a standalone speech separation system; it requires an accurate estimate of the steering vectors and noise statistics and must be used in combination with a source localization algorithm. In future work, we will analyze the sensitivity of the proposed technique to model mismatch and will consider blind localization techniques that are well suited to the hybrid architecture. The hypothesis testing approach, which has long been used for signal detection in communication and radar arrays, provides a new perspective on time-frequency masks in multichannel speech signal processing.

6. REFERENCES

[1] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, 2015.

[2] J. M. Kates, *Digital hearing aids*. Plural publishing, 2008.

[3] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," *Multichannel Speech Processing Handbook*, pp. 1065–1084, 2007.

[4] 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. I–529, IEEE, 2002.

[5] 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 Process.*, vol. 14, no. 6, pp. 2165–2173, 2006.

[6] D. Kolossa and R. Orglmeister, "Nonlinear postprocessing for blind speech separation," in *Independent Component Analysis and Blind Signal Separation*, pp. 832–839, Springer, 2004.

[7] 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.

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

[9] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 2, pp. 382–394, 2010.

[10] 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.

[11] T. Melia and S. Rickard, "Underdetermined blind source separation in echoic environments using desprit," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 1–19, 2006.

[12] 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.

[13] 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.

[14] 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.

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

[16] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, pp. 181–197, Springer, 2005.

[17] H.-G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Intl. Conf. on Acoustics, Speech, and Signal Process.*, vol. 1, pp. 153–156, IEEE, 1995.