SPEECH SOURCE SEPARATION BY COMBINING LOCALIZATION CUES WITH MIXTURE MODELS OF SPEECH SPECTRA

*Kevin Wilson**

Mitsubishi Electric Research Lab, Cambridge, MA
wilson@merl.com

ABSTRACT

We present a method for simultaneous speech source separation in reverberant environments using both localization cues and a speech model. Previous source separation work has focused primarily on one or the other of these approaches; we use a novel localization cue observation noise model to allow for a natural combination of the approaches. We model speech as a Gaussian mixture model (GMM) of short-time spectral magnitudes and model localization cue noise using a time-varying noise model learned from labeled training data. We show that our technique outperforms competing techniques as measured by segmental signal-to-noise ratio (SNR) and segmental log-spectral distortion (LSD) and also show that our technique is robust to typical levels of audio localization error.

Index Terms— Speech enhancement, Array signal processing, Acoustic arrays, Speech processing

1. THE SOURCE SEPARATION PROBLEM

We address the problem of simultaneous speech source separation in reverberant environments. In this work, we assume that all of the mixed signals are speech-like, and we assume that localization cues derived from the mixed signal are available. This is consistent with the real-life "cocktail party" scenario in which a (presumably binaural) listener is able to attend to a single chosen voice among the many simultaneously active voices at a cocktail party.

Cherry [1] lists several factors that could contribute to successful cocktail party performance, including differing localization cues among the speakers, knowledge of the spectrotemporal dynamics of speech, differences among speakers' voices, and speech-related visual cues. A successful source separation solution will likely need to exploit several of these factors, although to this point most have focused on only one. This paper combines two of them, localization cues and spectrotemporal dynamics, and empirically demonstrates the utility of this combined approach.

1.1. Previous work

This work combines two previous approaches to speech source separation. The first approach is to learn generative models of individual speech sources and to use these in combination to "decode" a mixture of speech signals. One example of this approach is Roweis [2], which trains speaker-specific hidden Markov models (HMMs) on a spectrogram representation of speech and develops an efficient algorithm for decoding the factorial HMM resulting from the simultaneous activity of two individual speakers. From the states of the factorial HMM, [2] infers a binary mask on the spectrogram to separate the two speakers. Another example of this approach is Hershey and Casey [3], which trains speaker-specific narrow-band and wide-band HMM speech models and then uses simulated annealing to find the marginal probabilities of simultaneous states for a pair of speakers' HMMs given the observed mixed waveform. Based on these state probabilities, [3] creates a time-varying Wiener filter (implemented as a continuous-valued mask on the spectrogram representation) to separate the two simultaneous speakers. This paper builds most directly on the work in [4], which is itself a follow-up to the work in [2] and which found that modeling speech spectra with a Gaussian mixture model (an HMM model without a transition matrix) works just as well for the source separation problem. By using a model of the dynamics of speech, these techniques are able to achieve some speech separation given only a single audio channel as input.

The second approach that we combine is that of using localization cues to generate a binary spectrogram mask for source separation as in Yilmaz and Rickard [5]. [5] assumes that source locations are unknown and clusters localization cues across time and frequency to determine likely source locations according to an anechoic propagation model. Once these source location estimates are obtained, a spectrogram mask is generated by assigning each time-frequency bin to the source with which the observed intermicrophone phase and level differences at that time-frequency are most consistent. This approach does not require a spectrotemporal model of speech, but it does require a source of localization cues, most typically derived from a microphone array.

Our technique, described below, combines the above two approaches. It is conceptually similar to Nix et al. [6], although because of our implementation choices (and to a small extent because of the faster processors available today), our technique runs over ten-thousand times faster than that of [6]. Because of this, we are able to test our technique more extensively, and we report results for more strongly reverberant environments. In addition, we use a novel time-varying localization cue observation noise model, described in detail in [7], to facilitate our combined approach.

2. OUR APPROACH: COMBINING LOCALIZATION CUES WITH SPEECH MODELS

We combine localization cues with a spectrotemporal model of speech in a probabilistic framework and use this model to find a binary spectrogram mask for separating mixed speech signals. We focus on the case of a two-microphone array with two simultaneous speakers, and we begin by defining a Gaussian mixture state-space model for the problem (similar to that defined in [4]):

*The author performed the work while at MIT CSAIL.

$$p(z_i(u)=k) = \pi_k \qquad k \in \{1,\ldots,K\} \quad (1)$$

$$a(u,f) = \arg\max_i(\mu_{z_i(u)}(f)) \quad (2)$$

$$p(y(u,f)|z_i(u)) \sim \mathcal{N}(\mu_{z_{a(u,f)}}(f), \sigma^2_{spec_{z_{a(u,f)}}}(f)) \quad (3)$$

$$p(\theta(u,f)|z_i(u)) \sim \mathcal{N}(\theta_{direct_{a(u,f)}}, \sigma^2_{loc}(u,f)) \quad (4)$$

In the above model, $z_i(u)$ is the hidden state index (in our GMM with K states) for speaker $i \in \{1,2\}$ in the spectrogram frame indexed by u. Because we are assuming a GMM with no dynamics, the distribution over states is independent of time and is given by the probability mass function π_k. Each state z is associated with a log-spectral output distribution specified by mean $\mu_z(f)$ and diagonal covariance $\sigma^2_{spec_z}(f)$, where f is an index over frequency bins. To determine the observed log-spectrogram of the mixed signal, we make the assumption (as in [4]) that $\log(a+b) \approx \max(\log a, \log b)$, so the log-spectrogram of the mixed signal, $y(u,f)$, will be approximately the bin-wise maximum of the two individual speakers' spectrograms. Note that, incontrast to [4, 2, 3], we use the same speaker-independent spectral model for each of the two speakers. (We train this model on speech from several different speakers, none of which are in our test set.) The advantage of the speaker-independent model is that we do not need to know in advance whose voices we are separating. The disadvantage is that much of the separating ability of, for example [4], is due to the differences in the marginal spectral distributions of the speaker-dependent models. As such, our speaker-independent model is not sufficient to separate voices in a monaural mixed signal, and this is why we also incorporate localization cues from the array data.

We assume the locations of the two speakers relative to the microphone array are known, either by simultaneously using the microphone array to localize them or by tracking them in some other way, for example by a vision-based tracker. Given the known speaker locations, we can compute the intermicrophone phase difference due to the direct path propagation as a function of frequency, which we denote as $\theta_{direct}(f)$. We assume that the observed intermicrophone phases, $\theta(u,f)$, will have the $\theta_{direct}(f)$ of the bin-wise loudest source as their means and will have variances $\sigma^2_{loc}(u,f)$ determined by a linear function of the log-spectrogram values learned from a labeled training corpus. The details of computing $\sigma^2_{loc}(u,f)$ are in [7], where we showed that these variance estimates can be used to improve localization performance and are related to the psychoacoustics of the precedence effect. The important thing for this work is that these variance estimates provide a natural way to incorporate localization cues into this state-space model.

Given the above generative model for log-spectrum and phase difference observations of simultaneous speech, we must now compute these observations from our input signals. We begin with a time-domain signal $x_i(t)$ from each of our two microphones. We then compute a complex spectrogram representation of each signal, $s_i(u,f)$. From these spectrograms, we compute the intermicrophone phase difference

$$\theta(u,f) = \angle \frac{s_1(u,f)}{s_2(u,f)} \quad (5)$$

and we take the log-magnitude of the first channel as our log-spectrogram observation $y(u,f) = \log s_1(u,f)$.

We chose $K = 40$ states for our speech GMM, which means that the factorial GMM resulting from the mixture of two speakers has $40^2 = 1600$ states, which is still a manageable number of states to evaluate on current computer hardware. To infer the posterior probability of the factorial states, we simply compute the likelihood of each state and normalize:

Technique	Segmental SNR (dB)	Segmental LSD (dB)	Listener Pref. (%)
Wiener filter (oracle)	11.1	6.2	97
Ideal mask (oracle)	9.7	4.0	83
GMM + loc.	5.2	6.4	34
DUET	-0.6	6.6	19
Delay-and-sum	1.8	8.2	44
Convolutive BSS	3.8	9.2	39
Original mixture	0.3	8.4	33

Table 1. Average separation performance in synthetic rooms. "Listener pref." is the percentage of the time that each technique was preferred in paired comparisons with other techniques.

$$L_{m_{tot}}(u) = L^{\alpha}_{m_{shape}}(u) L^{\beta}_{m_{loc}}(u) \quad (6)$$

$$L_{m_{spec}}(u) = \prod_f \mathcal{N}(y(u,f); \mu_m(f), \sigma^2_{spec_m}(f)) \quad (7)$$

$$L_{m_{loc}}(u) = \prod_f p(\theta(u,f); \theta_{direct_{a(u,f)}}(u,f), \sigma^2_{loc}(u,f)) \quad (8)$$

$$p_{m_{tot}}(u) = \frac{L_{m_{tot}}(u)}{\sum_{m=1}^{K^2} L_{m_{tot}}(u)} \quad (9)$$

Here, m is an index over the factorial state, $L_{m_{tot}}$ is the overall likelihood, which is a product of $L_{m_{spec}}$, the spectral shape likelihood, and $L_{m_{loc}}$, the localization cue likelihood, each of which is computed according to the state-space model defined above. (We have abused notation slightly by using μ_m and $\sigma^2_{spec_m}$ to now represent the mean and covariance of a combined factorial state, whereas in Equation 3, they represented the mean and covariance of an individual speaker state.) Likelihood weighting terms α and β (with values chosen to maximize performance on a validation data set) are used to adjust the dynamic ranges of the two terms to improve performance.

Each factorial state implies a single-frame binary spectrogram mask (denoted as $mask_m(f)$) based on which of the two individual speaker states has the higher expected log-magnitude. Given the posterior state probabilities, we can compute the expected value of the binary mask (yielding a continuous-valued mask on $[0,1]$) which we then threshold to generate a final time-frequency binary mask

$$M(u,f) = \begin{cases} 1 & \text{for } (u,f) \text{ s.t.} \sum_m (p_{m_{tot}}(u) * mask_m(f)) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

which can then be applied to the spectrogram as in [4, 5] to separate the speakers. Chapter 5 of [8] contains a more detailed description of our technique and its relationship to other approaches.

3. RESULTS

We have tested our technique on data from real and synthetic reverberant environments. Table 1 summarizes our results on synthetic data, and Table 2 summarizes our results on real data. (Real data was collected from individual speakers in real rooms and additively combined to generate simultaneous speaker data for our tests.) We evaluate by calculating segmental signal-to-noise ratio (SNR) and seg-

Technique	Segmental SNR (dB)	Segmental LSD (dB)
Wiener filter (oracle)	7.6	4.0
Ideal mask (oracle)	5.3	4.8
GMM + loc.	2.7	7.0
DUET	0.6	7.9
Delay-and-sum	1.5	8.0
Convolutive BSS	1.1	8.6
Original mixture	0.6	8.4

Table 2. Average separation performance in real rooms.

mental log-spectral distortion (LSD) as described in [9]. We calculate segmental SNR and segmental LSD on over an hour of recorded speech with synthetically added reverberation (for the synthetic data case) consisting of ten male and six female speakers and on roughly thirty minutes of audio of live humans recorded in three different real rooms for the real data case. The intermicrophone spacing was 37.5 cm for both the real and synthetic data. The synthetic data environments ranged in reverberation time from 200 ms to 1600 ms, and the real data environments ranged in reverberation time from 400 ms to 800 ms. The synthetic data tested speaker separations from 8° to 61°, and the real data tested speaker separations from 16° to 61°. Due to space constraints, we present only average performance numbers here. Chapter 5 of [8] presents detailed results as a function of room reverberation time and speaker separation angle. For the purposes of this evaluation, we used the measured positions of the speakers with respect to the microphones for the real data and the known simulated locations for the synthetic data to determine θ_{direct} for each scenario.

Results in the summary tables are average performance over all of these conditions. In addition, for the synthetic data case, we conducted a small scale (15 subject) listener study on a subset of the data to determine whether our automated objective measures were consistent with subjective human evaluation. In the listener study, subjects were presented with separation results from pairs of different techniques and asked to pick which one of the two techniques separated the speakers better. For each technique, we report the fraction of the trials in which that technique appeared in which it was favored.

We evaluate our technique, "GMM + loc." (described in Section 2), in comparison to several others. "Wiener filter" is an oracle-based technique in which a time-varying Wiener filter was calculated from knowledge of the isolated speech spectra. "Ideal mask" is the ideal binary mask, again calculated from knowledge of the isolated speech spectra. "DUET" corresponds to using our technique without any speech GMM, in which we make independent mask decisions in each time-frequency bin based only on localization cues. This is nearly identical to the DUET technique described in [5], although we use our Gaussian noise model as opposed to the histogram-based noise model in [5], and we assume that the source locations are known, while [5] determined the source locations in an initial unsupervised clustering phase. Assuming known source locations should only help the DUET technique, and in preliminary tests, we did not find significant differences in performance between our noise model and that of [5]. ([5] notes that their technique does not perform well in reverberant environments and that it could benefit from a speech model. We feel that our formulation is a natural extension of theirs.) We also compare to two beamformer-based source separation techniques. "Convolutive BSS" is the blind source separation (BSS) technique described in [10] which finds an unmixing filter to decorrelate its outputs at multiple points in time. "Delay-and-sum" is a delay-and-sum beamformer steered to the known source locations. Finally, "Original mixture" is the signal from one microphone with no processing.

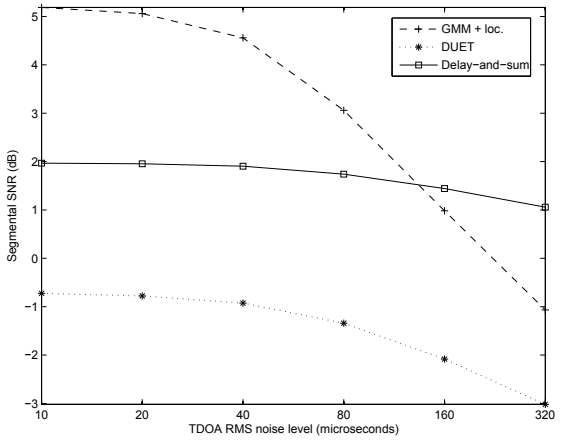
Both of these oracle-based techniques out-perform all of the non-oracle techniques, and the continuous-valued Wiener filter outperforms the binary mask. The ideal binary mask can be thought of as a thresholded Wiener filter, so this is not surprising. Our technique has the best performance on the automated objective metrics for both real and synthetic data of all of the non oracle techniques (highest segmental SNR and lowest segmental LSD). Of the other techniques, "DUET" achieved the next best segmental LSD for both real and synthetic data, and "Convolutive BSS" had the next best segmental SNR on synthetic data while "Delay-and-sum" had the next best segmental SNR on real data. The fact that our "GMM + loc." outperforms "DUET" shows the utility of including a spectral model. The fact that our technique outperforms delay-and-sum and convolutive BSS shows that binary spectrogram mask techniques are a promising alternative to beamformer-based techniques for incorporating both spectral models and localization information. The fact that the overall ordering of the results is similar for real and synthetic data shows that none of the techniques are unfairly exploiting any potential peculiarities of the synthetic data.

The human listener tests provide a somewhat different perspective on the performance of the separation algorithms. Again, the two oracle-based techniques outperform the non-oracle techniques, and the Wiener filter outperforms the ideal binary mask. However, human listeners prefered the two beamformer-based techniques over the binary mask-based techniques, and in fact the simple delay-and-sum algorithm was the best-performing non-oracle method. In informal post-listener-test interviews, several subjects mentioned that they were annoyed by the "artifacts" generated by some of the techniques, and that they preferred techniques that generated fewer artifacts. (Subjects were simply instructed to choose the technique in each trial that "separated best." Nothing was explicitly mentioned about the perceptibility of artifacts.) The binary mask-based techniques tend to generate more noticeable artifacts because the binary masks themselves change every frame, and any errors in these rapid changes are easily perceived. The beamformer-based techniques simply compute a single separating filter for the entire utterance, and even if this single filter does not separate as well (as measured, for example, by segmental SNR), the lack of abrupt changes leads to fewer artifacts. Still, our listener test shows that "GMM + loc." outperforms "DUET," again showing the utility of including a spectral model.

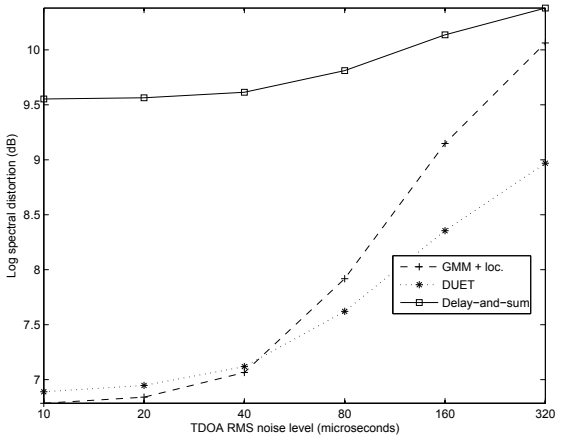
Finally, we note that our algorithm description in Section 2 was slightly simplified for expository purposes. The results here were obtained using a slight modification to Equation 8, in which the $p(\theta(u,f); \theta_{direct_{a(u,f)}}(u,f), \sigma^2_{loc}(u,f))$ term was smoothed across frames using a first-order autoregressive filter on the log probability. This was done to reduce unnecessarily rapid fluctuations in the localization likelihood (which would lead to unnecessarily rapid fluctuations in the binary mask itself). More details can be found in [8].

3.1. Sensitivity to localization errors

To this point, we have assumed that the precise locations of the speakers are known. This is unrealistic, so we now examine the sensitivity of our technique to localization errors. Figure 1 shows the separation performance on a randomly selected subset of the synthetic data after adding random time-delay errors at varying root-mean-square (RMS) levels to the assumed source positions. We do this only for the techniques that require knowledge of the source



(a) Segmental SNR



(b) Log spectral distortion

Fig. 1. Source separation performance as a function of TDOA estimation error. The horizontal axis shows the RMS level of the synthetically generated time delay noise on a log scale. These results are average performance across all tested reverberation times and source separations.

locations. (The two oracle-based techniques and convolutive BSS do not require this.) Our results show that error levels up to 40 μs RMS have only negligible effects on segmental SNR and segmental LSD performance. In [8], we show that time delay estimation errors are typically smaller than this in a wide range of moderately reverberant (below 400 ms reverberation time) and moderately noisy (above 12 dB SNR) acoustic environments. This shows that our results obtained above using perfect knowledge should in fact apply to realistically noisy localization estimates.

4. CONCLUSION

We have presented a simple and computationally efficient state-space model of speech that combines a speaker-independent GMM spectral model with a time- and frequency-varying observation noise model of intermicrophone phase to separate mixtures of two speakers. We have demonstrated source separation in real and synthetic reverberant environments, and have shown that our algorithm is robust to typical levels of localization error. Thanks to Trevor Darrell, Michael Brandstein, John Fisher, and Michael Siracusa for their help.

5. REFERENCES

[1] E. Colin Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] Sam T. Roweis, "One microphone source separation.," in *NIPS*, 2000, pp. 793–799.

[3] J. Hershey and M. Casey, "Audiovisual sound separation via hidden markov models," in *NIPS*, 2002.

[4] S. Roweis, "Automatic speech processing by inference in generative models," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed., pp. 97–134. Springer, 2004.

[5] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[6] J. Nix, M. Kleinschmidt, and V. Hohmann, "Computational scene analysis of cocktail-party situations based on sequential monte carlo methods," in *37th Asilomar Conference on Signals, Systems, and Computers*, 2003, vol. 1, pp. 735–739.

[7] Kevin Wilson and Trevor Darrell, "Learning a precedence effect-like weighting function for the generalized cross-correlation framework," *IEEE Transactions on Audio, Speech, and Language Processing*, Nov 2006 (to appear).

[8] Kevin W. Wilson, *Estimating Uncertainty Models for Speech Source Localization in Real-World Environments*, Ph.D. thesis, Massachusetts Institute of Technology, 2006.

[9] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective measures of speech quality*, Prentice Hall, 1988.

[10] L. Parra and C. Spence, "Convolutive blind source separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.