Combining Monaural and Binaural Evidence for Reverberant Speech Segregation

John Woodruff[1], *Rohit Prabhavalkar*[1], *Eric Fosler-Lussier*[1], *DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, United States
[2]Center for Cognitive Science, The Ohio State University, United States

{woodrufj,prabhava,fosler,dwang}@cse.ohio-state.edu

Abstract

Most existing binaural approaches to speech segregation rely on spatial filtering. In environments with minimal reverberation and when sources are well separated in space, spatial filtering can achieve excellent results. However, in everyday environments performance degrades substantially. To address these limitations, we incorporate monaural analysis within a binaural segregation system. We use monaural cues to perform both local and across frequency grouping of mixture components, allowing for a more robust application of spatial filtering. We propose a novel framework in which we combine monaural grouping evidence and binaural localization evidence in a linear model for the estimation of the ideal binary mask. Results indicate that with appropriately designed features that capture both monaural and binaural evidence, an extremely simple model achieves a signal-to-noise ratio improvement of up to 3.6 dB relative to using spatial filtering alone.

Index Terms: Speech segregation, binaural localization, monaural grouping, linear model

1. Introduction

Binaural speech segregation systems typically use spatial filtering to enhance the signal from a specific direction of arrival [1, 2]. Beamforming is a ubiquitous approach to spatial filtering but has well known limitations, such as substantial performance degradation in reverberant environments. Recent approaches to spatial filtering have incorporated time-frequency (T-F) masking. Although some effort has been made to overcome the performance degradation caused by reverberation [3, 4, 5], such systems are fundamentally limited by the decreased discriminative capacity of directional cues in reverberant environments.

To address such limitations, we complement spatial filtering with monaural analysis. Monaural cues are potentially more robust to reverberation than binaural cues, and provide a powerful mechanism for both local and across frequency grouping of speech energy. If a partial grouping of signal components can be obtained using monaural cues, then binaural cues can be integrated over the grouped components and can be used more effectively in reverberant conditions.

Prior work exploring the integration of monaural and binaural cues for speech segregation or enhancement is limited. In [6], localization cues are used to perform initial segregation in reverberant conditions. Initial segregation provides a favorable starting point for estimating the pitch track of the target voice, which is then used to further enhance the target signal. Our prior work showed that pitch-based monaural grouping can be used to improve segregation of voiced speech over binaural analysis alone [7].

The segregation framework presented here can be considered a *computational auditory scene analysis* (CASA) approach [2]. However, in CASA-based speech segregation systems, one typically performs segmentation followed by grouping, where fixed T-F segments are formed using one or more features, and then each segment is labeled as target or interference dominant using a set of grouping features. As an alternative to such an approach we previously argued for the use of a random field formalism for incorporating multiple cues for binary T-F mask estimation [8]. Here we use a computationally simpler approach based on the perceptron algorithm [9] while maintaining the structure of a random field. An attractive aspect of the proposed framework is its ability to integrate multiple sources of evidence by capturing the interaction between different types of features jointly in a linear model. Further, it allows the use of multiple approaches to segmentation or feature generation and avoids the need for heuristic parameter tuning.

In the following section we present a linear model for binary time-frequency mask estimation. Section 3 describes the monaural and binaural processing used, and the specifics of how we generate features that combine binaural and monaural cues. We present segregation results using different variations of the proposed system and a comparison system in Section 4, and conclude with a discussion in Section 5.

2. Time-frequency mask estimation using linear models

In this study, we assume a binaural recording of two speech sources. We convert the binaural mixture to a T-F representation using a bank of 128 gammatone filters with center frequencies from 50 to 8000 Hz spaced on the equivalent rectangular bandwidth scale. Each filtered signal is divided into 20 ms time frames with a frame shift of 10 ms to create a *cochleagram* [2] of T-F units. For notational convenience, we index T-F units by a single variable, i, although it is important to keep in mind that each T-F unit has an associated frequency channel, c, and time frame, m.

We perform segregation by estimating a binary T-F mask, $\mathbf{y}$, using the observed mixture data, $\mathbf{x}$. We seek to estimate the *ideal binary mask* (IBM), which has been proposed as a main computational goal of CASA systems [2]. We consider two approaches for calculating a binary mask. In the first approach, we use a perceptron to independently assign labels, y_i, to individual T-F units based on a vector of L local features,

$\psi(\mathbf{x}, i) = [\psi_1(\mathbf{x}, i), \psi_2(\mathbf{x}, i), \cdots, \psi_L(\mathbf{x}, i)]^T$. Explicitly,

$$y_i = \begin{cases} 1 & \text{if } \varphi(\mathbf{w}^T \psi(\mathbf{x}, i)) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\mathbf{w}$ is a set of linear weights and the function $\varphi(\cdot)$ is the hyperbolic tangent sigmoid transfer function. Given a set of training examples, the parameters $\mathbf{w}$ are determined using the standard perceptron training algorithm.

As an alternative, we consider a model that attempts to estimate the T-F mask as a whole by treating T-F units as vertices in a graph G,

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \sum_i \mathbf{w_1}^T \phi_{\mathbf{1}}(\mathbf{x}, y_i) + \sum_{\substack{i \\ j \in \mathcal{N}(i)}} \mathbf{w_2}^T \phi_{\mathbf{2}}(\mathbf{x}, y_i, y_j) \quad (2)$$

where, $\mathcal{N}(i)$ is the set of neighbors of T-F unit i in G, $\phi_{\mathbf{1}}(\mathbf{x}, y_i)$ and $\phi_{\mathbf{2}}(\mathbf{x}, y_i, y_j)$ are vectors of *association* and *interaction* feature functions, respectively, and $\mathbf{w_1}$ and $\mathbf{w_2}$ are linear weights learned through training.

The association feature vector associates the L local features for T-F unit i, $\psi(\mathbf{x}, i)$, with the two possible label assignments,

$$\phi_{\mathbf{1}}(\mathbf{x}, y_i) = \left[\cdots, \psi_l(\mathbf{x}, i)\delta_{(y_i=0)}, \psi_l(\mathbf{x}, i)\delta_{(y_i=1)}, \cdots\right]^T \quad (3)$$

where, $\delta_{(\cdot)}$ is an indicator function that is 1 if the associated condition is true and 0 otherwise.

Similarly, given a set of K features corresponding to an edge (i, j) between T-F units, $\xi(\mathbf{x}, i, j)$, the interaction feature vector is constructed using,

$$\phi_{\mathbf{2}}(\mathbf{x}, y_i, y_j) = \left[\cdots, \xi_k(\mathbf{x}, i, j)\delta_{(y_i=y_j)}, \xi_k(\mathbf{x}, i, j)\delta_{(y_i \neq y_j)}, \cdots\right]^T \quad (4)$$

Intuitively, the weights $\mathbf{w_1}$, associated with the association features are intended to model local dependencies between the features and the label assigned to a particular T-F unit; the weights $\mathbf{w_2}$, associated with the interaction features are intended to model whether pairs of connected T-F units should be assigned same or different labels. Note that the form of the interaction feature functions used in this work are essentially the same as those used in our previous work [8]. The parameters $\mathbf{w_1}$ and $\mathbf{w_2}$ in Equation (2) are learned using the averaged perceptron algorithm for structured inputs proposed by Collins [9].

2.1. Determining the optimal T-F mask

At test time or during the execution of the perceptron algorithm, given a data example $\mathbf{x}$, we need to compute the most likely T-F mask, $\hat{\mathbf{y}}$, according to Equation (2). The optimal T-F mask can be determined using the graph-cut algorithms that have been previously used for inference in Markov random fields [10]. We use the graph-cut solution as a seed-labeling for a number of iterations the QPBO-Improve algorithm proposed by Rother et. al. [11] to find the minimum energy configuration. In our experiments, we used the software implementations of the graph cut algorithms described in [12] and the implementation of QPBOI described in [11].

3. Feature functions for combining monaural and binaural cues

To estimate the IBM using the methods discussed above, we must design a set of data-dependent feature functions. We first describe the nature of the binaural and monaural cues themselves, before discussing how these cues are used to generate both association features and interaction features.

3.1. Azimuth-dependent binaural cues

As in existing spatial filtering systems, we use binaural analysis as a means of indicating whether sound energy is more likely due to the target source or interference source, working under the assumption that sources impinge on the microphones from distinct azimuth angles. We calculate the interaural time difference (ITD), τ_i, and interaural level difference (ILD), λ_i, between the left and right mixture signals, and map ITD-ILD cues to azimuth-dependent cues using non-parametric ITD-ILD likelihood functions, $P_c(\tau_i, \lambda_i|\theta)$, where θ denotes azimuth. Note that we include the frequency channel subscript on the likelihood function, as we train a separate function for each frequency channel and azimuth considered. The likelihood functions are described in more detail in [7].

3.2. Monaural grouping cues

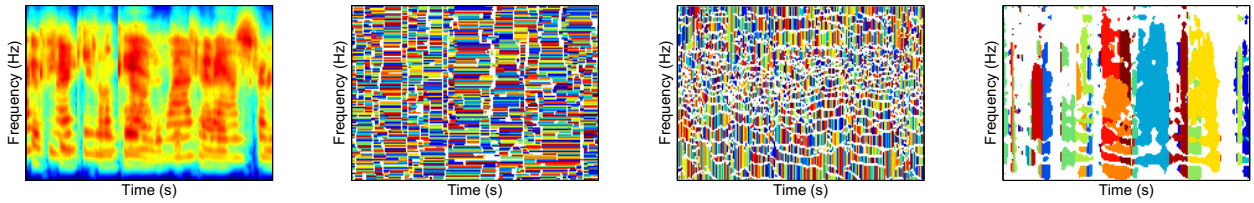
In existing CASA systems, monaural analysis has been used for *segmentation* (local grouping of T-F units) and *simultaneous grouping* (across frequency grouping within a continuous time interval) [2]. We use monaural cues in the same functional role, but whereas existing CASA systems generate a fixed set of T-F regions that are then labeled, we generate multiple types of T-F regions using different methods, and encode these regions as features in the model used for mask estimation. The linear weights learned through training can then be interpreted as balancing the evidence provided by different *region hypotheses*.

We focus on three types of monaural processing: local grouping in time, local grouping in frequency, and non-local grouping across frequency. To capture local grouping in time we use both an energy-based analysis and a correlation analysis of correlogram responses between T-F units in neighboring time frames. To capture local grouping across frequency, we again incorporate an energy-based analysis and a correlation analysis of correlogram responses between T-F units in neighboring frequency channels. To generate features that capture non-local relationships across frequency, we use pitch-based cues generated by the systems presented in [13, 14]. The pitch-based cues can be used to relate T-F units at an arbitrary distance in frequency.

In Figure 1, we show three examples of T-F regions formed using monaural analysis, one for each type of grouping described. In total, we incorporate 11 different methods to generate T-F regions. It is important to note that these regions are used to generate features for our model rather than treated as fixed.

3.3. Association features

The association features used in the model are fundamentally binaural, as we use source azimuth as the mechanism to determine whether the target or interference source is dominant. In this study, we assume the azimuth of each source is known, and we use the azimuth-dependent likelihood functions described in Section 3.1 to generate an exclusively binaural association fea-



(a) Cochleagram (b) Local grouping in time (c) Local grouping in frequency (d) Non-local grouping

Figure 1: We show the cochleagram (a) of a two-talker mixture with reverberation time of 0.4 s, and three different sets of hypothesized T-F regions. We show hypothesized local grouping in time (b), local grouping in frequency (c) and non-local grouping (d).

ture for each T-F unit,

$$\psi_1(\mathbf{x}, i) = \log\left(\frac{P_c(\tau_i, \lambda_i|\theta_0)}{P_c(\tau_i, \lambda_i|\theta_1)}\right). \quad (5)$$

where θ_0 and θ_1 denote the azimuth of the target and interference signals, respectively. In Section 4.3, we include a subscript 'B' for systems that use this association feature.

We can incorporate the monaural grouping cues into additional association features by generating *context-sensitive* binaural features. For example, given a set of T-F regions, S_l, we generate a feature for each T-F unit using,

$$\psi_{S_l}(\mathbf{x}, i) = \frac{1}{|s_k|}\sum_{j \in s_k} \log\left(\frac{P_c(\tau_j, \lambda_j|\theta_0)}{P_c(\tau_j, \lambda_j|\theta_1)}\right), \quad (6)$$

where i is a member of T-F region s_k, which is a member of the set S_l and $|s_k|$ denotes the number of T-F units contained in T-F region s_k. The feature $\psi_{S_l}(\mathbf{x}, i)$ thus integrates information from a region hypothesis and is shared by all T-F units contained in the region. Given a number of different hypothesized sets of T-F regions, $\{S_1, S_2, \ldots, S_L\}$, where each S_l is itself a collection of T-F regions s_k, we can generate the set of features $[1, \psi_1(\mathbf{x}, i), \psi_{S_1}(\mathbf{x}, i), \cdots, \psi_{S_L}(\mathbf{x}, i)]$ for each T-F unit i. In this study $L = 11$. When combined with $\psi_1(\mathbf{x}, i)$ and a bias feature we have a total of 13 association features. In Section 4.3, we include a subscript 'A' for systems that use this entire vector of association features.

3.4. Interaction features

Using interaction features to encode the monaural cues is a natural fit since the monaural cues capture relatedness between different T-F units. In generating interaction features, we treat edges between two neighboring T-F units in time within the same frequency channel (*time edges*) and between T-F units in different frequency channels but the same time frame (*frequency edges*) separately, by using different interaction features for each edge type. We generate binary interaction features for time edges using the T-F regions for local grouping in time. Similarly, we generate binary interaction features for frequency edges using the T-F regions for both local and non-local frequency grouping. The binary features take a value of 1 when two connected T-F units are both contained in the same T-F region.

In addition to these binary features, we generate two real-valued features for both time and frequency edges using the pitch-related cues discussed in Section 3.2. We also include a separate bias feature for time and frequency edges, giving us 9 features over time edges and 6 features over frequency edges, for a total of 15 interaction features. In Section 4.3, we include a subscript 'I' for systems that incorporate the interaction features.

4. Evaluation

4.1. Database

We use the ROOMSIM package [15] to generate impulse responses that simulate binaural input at human ears. We generate a library of binaural impulse responses for direct sound azimuth angles between -90° and 90° spaced by 5°, and 3 reverberation conditions: T_{60} = 0.2, 0.4, 0.6 s. We generate 200 two-talker mixtures where sources are set to have equal energy using the monaural utterances prior to mixing. Sources are spaced randomly, but are constrained to be between 10° and 120° apart. We use 100 of the mixtures with T_{60} = 0.4 s for training and 25 for development. To analyze generalization of the models, we test on the remaining 75 mixtures and on 75 mixtures for T_{60} = 0.2 and 0.6 s.

4.2. Graph structure

For the systems in which we incorporate interaction features and decode the mask as a whole, we use a neighborhood structure that connects each T-F unit to its neighbors in time, and to T-F units at logarithmic jumps in frequency. Specifically, each T-F unit is connected to T-F units in the same time frame at jumps of 1, 2, 4, 8, 16 and 32 frequency channels, both up and down. Thus, each T-F unit can have up to 14 neighbors.

4.3. Segregation performance

To assess segregation performance of the proposed systems, we measure the signal-to-noise (SNR) of the estimated signals relative to the signal generated by the IBM. We compare our proposed framework to the recent, binaural only "MESSL" system [5]. We also compare a number of alternative systems using the models described in Section 2. We consider two *local* systems that estimate a T-F mask independently in each T-F unit using Equation (1). For one local system, Local_B, we use only a bias and the binaural association feature. The second local system, Local_A, uses the full set of association features and thus captures the monaural grouping through the context-sensitive binaural features. We also consider two *global* systems, which generate an entire T-F mask using Equation (2). The neighborhood structure of these graph-based systems is described above in Section 4.2. The first global system, Global_BI, uses only the binaural feature and a bias for the association features, and incorporates the monaural cues by using the full set of interaction features. The second global system, Global_AI, uses the full set of association and interaction features, thereby capturing monaural grouping in both ways.

Table 1: Average SNR (in dB) for five alternative segregation systems as a function of T_{60} in s. Subscript 'B' denotes use of the single, binaural association feature. Subscript 'A' denotes use of the full vector of association features. Subscript 'I' denotes use of the full vector of interaction features.

T_{60}	MESSL	Local_{B}	Local_{A}	$\text{Global}_{\text{BI}}$	$\text{Global}_{\text{AI}}$
0.2	8.9	11.3	13.6	11.7	12.8
0.4	5.6	6.6	10.2	7.7	9.2
0.6	4.4	4.9	7.7	5.5	6.8

In Table 1 we show the SNR averaged over 75 mixtures in each of 3 T_{60} times for each of the 5 alternative systems. Our first observation is that each system that incorporates monaural analysis, either in the association or interaction features, outperforms both exclusively binaural systems in all 3 T_{60} times. We should note that while the MESSL system is a recent binaural system, Mandel et al. define the segregation goal as estimating the anechoic target signal, whereas we desire an estimate of the reverberant target signal. With our goal in mind, we see the comparison to both MESSL and the Local_{B} system as informative because one can see the ability to discriminate between target or interference dominant T-F units is severely degraded by reverberation.

Comparing the two local systems, we can see that encoding monaural cues in a set of association features improves SNR by between 2.3 and 3.6 dB, depending on the reverberation time. Comparing the Local_{A} system to the $\text{Global}_{\text{BI}}$ system, we can see that encoding the monaural information in the interaction features does not appear to be as successful in terms of SNR. In listening to the output signals, this appears to be due to less effective suppression of the interference source. The performance of the globally estimated masks is improved when encoding the monaural cues in both association and interaction features, as in the $\text{Global}_{\text{AI}}$ system.

Visual inspection of the masks generated by the global systems shows that the estimated masks are much smoother (neighboring T-F units are more likely to have the same label) than either of the local systems. In highly reverberant conditions, or for mixtures in which sources are closely spaced, the global systems have a tendency to group too many T-F units across frequency together. This is potentially due to the fact that we treat all frequency neighbors the same in the interaction features, so the weights learned over the features are shared for frequency edges whether the T-F units are in neighboring channels or are 32 channels away.

5. Concluding remarks

Our results indicate that integrating monaural and binaural cues improves segregation performance relative to using binaural cues alone. We have shown an increase of up to 3.6 dB in terms of SNR. We have proposed a novel method for learning to weight multiple sources of monaural grouping evidence, and shown that using a simple linear combination in the context-sensitive feature space can achieve good performance.

The global systems which seek to estimate the T-F mask as a whole produce very different time-frequency masks. These masks are much smoother in that neighboring T-F units across frequency are much more likely to receive the same label. This seems to reduce the amount of artifacts present in the signal, and in some cases produces a more natural sounding output. In future work we will consider treating across frequency edges differently, so that weights are not tied for each type of frequency connection. Future work should also consider alternative graph structures and explore how across frequency connectivity affects performance. One could also consider a more sophisticated training approach or alternative methods for monaural or binaural analysis.

6. Acknowledgements

The authors would like to thank M. Mandel for providing his implementation of the MESSL algorithm. This research was supported by an AFOSR grant (FA9550-08-1-0155), and NSF grants (IIS-0534707) and (IIS-0905420).

7. References

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[2] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley/IEEE Press, 2006.

[3] N. Roman, S. Srinivasan, and D. L. Wang, "Binaural segregation in multisource reverberant environments," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 4040–4051, 2006.

[4] H. Sawada, S. Araki, and S. Makino, "A two-state frequency-domain blind source separation method for underdetermined convolutive mixtures," in *Proc. WASPAA*, Oct. 2007, pp. 139–142.

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

[6] A. Shamsoddini and P. N. Denbigh, "A sound segregation algorithm for reverberant conditions," *Speech Commun.*, vol. 33, pp. 179–196, 2001.

[7] J. Woodruff and D. L. Wang, "Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization," *IEEE Trans. Acoust., Speech, Signal Proc.*, 2010, in press.

[8] R. Prabhavalkar, Z. Jin, and E. Fosler-Lussier, "Monaural segregation of voiced speech using discriminative random fields," in *Proc. Interspeech*, 2009.

[9] M. Collins, "Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms," in *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*. Morristown, NJ, USA: Association for Computational Linguistics, 2002, pp. 1–8.

[10] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 147–159, 2004.

[11] C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer, "Optimizing binary MRFs via extended roof duality," in *Prod. CVPR*, 2007.

[12] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1068–1080, 2008.

[13] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 17, pp. 625–638, 2009.

[14] ——, "A multipitch tracking algorithm for noisy and reverberant speech," in *Proc. ICASSP*, 2010.

[15] D. R. Campbell. (2004) The ROOMSIM user guide (v3.3). [Online]. Available: http://media.paisley.ac.uk/~campbell/Roomsim/