INTEGRATING MONAURAL AND BINAURAL ANALYSIS FOR LOCALIZING MULTIPLE REVERBERANT SOUND SOURCES

John Woodruff and DeLiang Wang

Department of Computer Science and Engineering
& Center for Cognitive Science
The Ohio State University
Columbus, OH, 43210-1277, USA
{woodrufj, dwang}@cse.ohio-state.edu

ABSTRACT

Localization of simultaneous sound sources in natural environments with only two microphones is a challenging problem. Reverberation degrades performance of localization based exclusively on directional cues. We present an approach that integrates monaural and binaural analysis to improve localization of multiple speech sources in noisy and reverberant environments. Our approach incorporates pitch-based monaural processing to perform simultaneous organization of voiced speech. We propose a probabilistic framework to jointly perform localization and sequential organization using binaural cues. We evaluate our system on multi-source speech mixtures in the presence of reverberation and diffuse noise and compare it to two localization approaches that do not incorporate monaural cues. Results indicate that our system can accurately localize multiple sources in very challenging conditions.

Index Terms— Binaural sound localization, sequential organization, monaural grouping, computational auditory scene analysis

1. INTRODUCTION

Localization of one or more sound sources is fundamental to auditory perception and signal processing strategies that seek to enhance a source signal by spatial filtering. Numerous approaches have relied on the cross-correlation framework [1, 2], but there are well known limitations to these approaches in reverberant environments.

Inspired by human sound localization, systems have been developed that seek to localize sources with only two microphones [3]. The approach proposed in [4], termed the "stencil" filter, performs coincidence detection between left and right mixtures signals in individual frequency channels. Source azimuths are estimated by integrating coincidence cues along azimuth-dependent primary traces, due to the *interaural time difference* (ITD) for that azimuth, and secondary traces, due to spatial aliasing present at frequencies where the wavelengths of the signal are shorter than the distance between microphones. The method proposed in [5] computes a "skeleton" cross-correlogram of the mixture signal, a running cross-correlation computed in individual frequency bands in which cross-correlation peaks are replaced by Gaussian functions with narrower width. The time-lag dimension is warped to azimuth using a learned set of monotonic functions. Like the stencil filter, cues are integrated across time and frequency into an azimuth response function where peaks can be selected as the underlying source angles.

These systems are representative of approaches to binaural localization, which rely exclusively on directional cues. Performance of these approaches degrades substantially in reverberant and noisy conditions. To combat the effect of reverberation on localization performance, the system proposed in [6] explicitly models ITD variability due to convolutive noise as a mixture of Gaussians within an expectation-maximization framework to cluster time-frequency (T-F) units with similar ITD cues. In [7, 8], monaural grouping is suggested as a mechanism to increase robustness to reverberation. Christensen *et al.* use pitch as a cue for the generation of contiguous T-F regions, or segments. Cross-correlation cues are then integrated over each segment to estimate the azimuth of the dominant source in each time frame [7]. Our prior work analyzes the impact of monaural grouping on localization and segregation of speech in reverberant environments [8], showing the benefit of monaural grouping for localizing reverberant sources.

In this work we propose a probabilistic approach to binaural localization of multiple sound sources that integrates monaural and binaural analysis. Our proposed system achieves across frequency grouping, or *simultaneous organization*, of voiced speech using pitch-based monaural cues. This allows locally extracted, unreliable binaural cues to be integrated within time-frequency regions called *simultaneous streams*. Localization and *sequential organization*, or grouping across time, of simultaneous streams are achieved jointly within a maximum likelihood framework. In contrast to [7], our approach utilizes both ITD and *interaural level difference* (ILD) cues and integrates cues across disparate regions of time through sequential organization.

In Section 2, we describe our approach to simultaneous organization, the methods used for generation of localization cues, and a mechanism for weighting binaural cues based on their expected reliability. We present a maximum likelihood approach to localization and sequential organization in Section 3. We evaluate our proposed framework in terms of multiple sound source localization in reverberant and noisy conditions in Section 4 and provide concluding remarks in Section 5.

2. BACKGROUND

2.1. Simultaneous organization

Simultaneous organization in computational auditory scene analysis (CASA) forms simultaneous streams, each of which may contain disconnected T-F segments across a continuous time interval. We use the tandem algorithm [9, 10] to generate simultaneous streams

This research was supported by an AFOSR grant (FA9550-08-1-0155), an NSF grant (IIS-0534707) and a grant from the Oticon Foundation.

978-1-4244-4296-6/10/$25.00 ©2010 IEEE

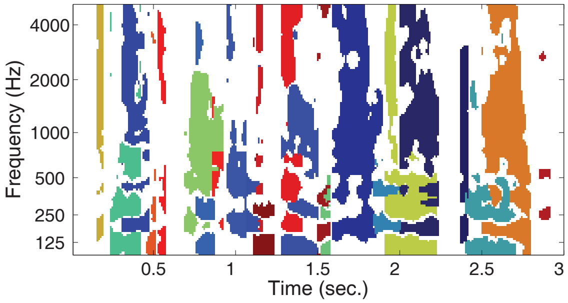

Fig. 1. (Color online) Example of simultaneous organization using the tandem algorithm. Simultaneous streams corresponding to different pitch contours are shown with different colors.

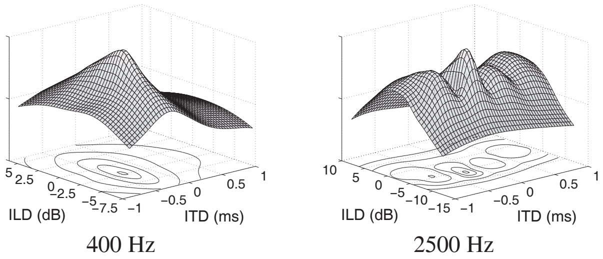

Fig. 2. Examples of ITD-ILD likelihood functions for azimuth 25° at frequencies of 400 and 2500 Hz. Each example shows the log-likelihood as a surface with projected contour plots that show cross sections of the function at equally spaced intervals.

for the summation of the left and right ear mixtures. The tandem algorithm iteratively estimates a set of pitch contours and associated simultaneous streams. In a first pass, T-F segments that contain voiced speech are identified using cross-channel correlation of correlogram responses. Up to two pitch points per time frame are estimated by finding peaks in the summary correlogram created from only the selected, voiced T-F segments. For each pitch point found, T-F units that are consistent with that pitch are identified using a set of trained multi-layer perceptrons (one for each frequency channel). Pitch points and associated sets of T-F units are linked across time to form pitch contours and simultaneous streams using a continuity criterion that measures pitch deviation and spectral overlap. Pitch contours and simultaneous streams are then iteratively refined until convergence.

We focus on mixtures in reverberant environments, and find that in this case the continuity criterion used in the tandem algorithm for connecting pitch points and simultaneous streams across time is too liberal. We find that performance improves if we break pitch contours and simultaneous streams when the pitch deviation between time frames is large. Specifically, let τ_1 and τ_2 be pitch periods from the same contour in neighboring time frames. If $|\log_2(\tau_1/\tau_2)| > 0.08$, the contour and associated simultaneous streams are broken into two contours and two simultaneous streams. The value of 0.08 was selected on the basis of informal analysis, and was not specifically tuned for optimal performance on the data set discussed in Section 4.

An example set of simultaneous streams is shown in Figure 1 for a mixture of two talkers in a reverberant environment with 0.4 sec. reverberation time (T_{60}). There are a total of 27 simultaneous streams shown, where each color corresponds to a separate simultaneous stream. One can see that simultaneous streams may contain multiple segments across frequency but are continuous in time.

2.2. Localization cues

For each T-F unit, we calculate the ITD as the time delay that produces the maximum peak in the cross-correlogram, and ILD as the energy ratio in dB between the left and right signals. We denote the ITD and ILD of T-F unit $u_{c,m}$ as $\tau_{c,m}$ and $\lambda_{c,m}$, respectively, where c and m index frequency channel and time frame, respectively. To translate from ITD-ILD cues to azimuth-dependent cues, we train a joint ITD-ILD likelihood function, $P_c(\tau_{c,m}, \lambda_{c,m}|\phi)$, for each of 37 azimuth angles, indexed by ϕ, and 128 frequency channels with center frequencies from 50 to 8000 Hz. The likelihood functions are trained on ITDs and ILDs generated from single-source speech in various simulated room environments using kernel density estimation [11]. Our approach is adapted from the one proposed in [5].

The likelihood distributions capture the frequency dependent pattern of ITDs and ILDs for a specific azimuth and the multi-peak ambiguities due to spatial aliasing at higher frequencies. The distributions also capture common deviations from the free-field cues due to reverberation. We show two distributions in Figure 2 for azimuth 25°. Note that, in addition to the above points, the azimuth-dependent distributions capture the complementary nature of localization cues [12] in that ITD provides more discrimination between angles at lower frequencies (note the large ILD variation in the 400 Hz example) and ILD provides more discrimination between angles at higher frequencies (note the large ITD variation in the 2500 Hz example).

2.3. Cue weighting

In reverberant recordings, many T-F units will contain cues that differ significantly from free-field cues. Including a weighting function or cue selection mechanism that indicates when an azimuth cue should be trusted can improve localization performance [13, 7]. Motivated by the *precedence effect* [14], we incorporate a simple cue weighting mechanism that identifies strong onsets in the mixture signal. We generate a real-valued weight, $w_{c,m}$, that measures the energy ratio between unit $u_{c,m}$ and $u_{c,m-1}$.

We have found better performance by keeping only those weights above a specified threshold. The difficulty with a fixed threshold however, is that one may end up with a simultaneous stream with no unit above the threshold. To avoid this we set a threshold for each simultaneous stream so that the T-F units exceeding the threshold retain 25% of the signal energy in the simultaneous stream.

3. LOCALIZATION AND SEQUENTIAL ORGANIZATION

Once simultaneous streams have been formed and azimuth-dependent cues have been generated, we localize the underlying source signals and perform sequential organization. We take a probabilistic approach similar to the one presented for model-based sequential organization in [15].

Let N be the number of sources in the mixture, and I be the number of simultaneous streams formed using monaural analysis. Denote the set of all possible azimuths as Φ and the set of simultaneous streams as $S = \{s_1, s_2, ..., s_I\}$. Let Y be the set of all N^I

sequential organizations, or labelings, of the set S and y be a specific organization. We seek to maximize the joint probability of a set of angles and a sequential organization given the observed data, D. This can be expressed as,

$$\hat{\phi}_0, \ldots, \hat{\phi}_{N-1}, \hat{y} = \underset{\phi_0,\ldots,\phi_{N-1}\in\Phi, y\in Y}{\arg\max} P(\phi_0, \ldots, \phi_{N-1}, y|D). \tag{1}$$

For simplicity, assume that $N = 2$ and denote the two angles as ϕ_0 and ϕ_1 for target and interference signals, respectively. Assuming that all angles and sequential organizations are equally likely we have,

$$\hat{\phi}_0, \hat{\phi}_1, \hat{y} = \underset{\phi_0,\phi_1\in\Phi, y\in Y}{\arg\max} P(D|\phi_0, \phi_1, y). \tag{2}$$

Now, let S_0 be the set of simultaneous streams labeled as target and S_1 be the set of simultaneous streams labeled as interference by the sequential organization y. Using ITD and ILD as the observed data, and assuming independence between simultaneous streams and T-F units of the same simultaneous stream, we can express Equation (2) as,

$$\hat{\phi}_0, \hat{\phi}_1, \hat{y} = \underset{\phi_0,\phi_1\in\Phi, y\in Y}{\arg\max} \Big(\prod_{s_i\in S_0} \prod_{u_{c,m}\in s_i} P_c(\tau_{c,m}, \lambda_{c,m}|\phi_0) \cdot \prod_{s_j\in S_1} \prod_{u_{c,m}\in s_j} P_c(\tau_{c,m}, \lambda_{c,m}|\phi_1) \Big). \tag{3}$$

Due to the assumption of independence between simultaneous streams, the above equation can be expressed as two separate equations using,

$$\hat{\phi}_0, \hat{\phi}_1 = \underset{\phi_0,\phi_1\in\Phi}{\arg\max} \Big(\sum_{i=1}^{I} \max_{k\in\{0,1\}} \Big(\sum_{u_{c,m}\in s_i} w_{c,m} \log(P_c(\tau_{c,m}, \lambda_{c,m}|\phi_k)) \Big) \Big), \tag{4}$$

$$\hat{y}_i = \underset{k\in\{0,1\}}{\arg\max} \Big(\sum_{u_{c,m}\in s_i} w_{c,m} \log(P_c(\tau_{c,m}, \lambda_{c,m}|\phi_k)) \Big). \tag{5}$$

In Equations (4) and (5) we have also incorporated the cue weighting parameter, $w_{c,m}$. For the case with $N > 2$, use $k \in \{0, \ldots, N-1\}$ rather than $k \in \{0, 1\}$ in both (4) and (5). The complexity of the search space is $I|\Phi|^N$, which is reasonable when the number of sources of interest is relatively small and the size of the azimuth space is moderate. In our experiments in Section 4, $|\Phi| = 37$ and $N \leq 3$.

4. EVALUATION

4.1. Database

We use the ROOMSIM package [16] to generate impulse responses that simulate binaural input at human ears. We generate a training and a testing library of binaural impulse responses for direct sound azimuth angles between -90° and 90° spaced by 5°, and 5 reverberation conditions: $T_{60} = 0$, 0.2, 0.4, 0.6, 0.8 seconds. Numerous room sizes, microphone positions, and source distances from the microphones are represented in both training and testing impulse response libraries. The ITD-ILD likelihood distributions are trained using impulse responses drawn from the training library. For all testing mixtures we select utterances from the TIMIT database at random and use impulse responses from the testing library. Mixture lengths are set using the first randomly selected speech source, where all subsequent speech sources are either truncated or concatenated with themselves to match the first source's length.

Table 1. Average azimuth error (in °) in reverberant conditions

Two talkers

$\mathbf{T_{60}}$ **(sec.)**	**0**	**0.2**	**0.4**	**0.6**	**0.8**
Skeleton CC	0.31	2.18	12.62	19.1	23.6
Stencil	0.36	1.9	3.21	4.69	5.53
Proposed	0.63	0.86	1.26	2.54	3.91

Three talkers

$\mathbf{T_{60}}$ **(sec.)**	**0**	**0.2**	**0.4**	**0.6**	**0.8**
Skeleton CC	1.3	8.29	20.11	24.09	25.98
Stencil	1.82	4.69	8.7	11.48	13.68
Proposed	0.63	0.87	2.41	4.48	7.01

Localization performance is evaluated on two and three-talker mixtures in all 5 T_{60} times. Performance is also shown for one, two and three-talker mixtures in the 0.4 sec. T_{60} condition with diffuse noise added. Diffuse noise is generated by passing white noise through a speech-shaped filter, where the left and right white noise signals are uncorrelated. For each testing condition, 200 mixtures are generated. For all mixtures, each speech source is set so that the summation of the left and right signals has equal energy. In the cases where diffuse noise is added, the level of the noise is set to achieve a specific SNR, using the summation of left and right signals, relative to *one* of the speech signals. For the one-talker mixtures with diffuse noise, the SNR reflects the level of the speech signal as compared to the noise. In the two and three-talker mixtures with diffuse noise, the SNR reflects the level of an individual speech source as compared to the other speech source(s) plus noise. Since each speech source is mixed to have equal energy, SNR is the same for each speech source taken individually. In the tests with diffuse noise, note that in the two-talker conditions, 0 dB SNR is the condition in which no diffuse noise is added, and in the three-talker conditions, -3 dB SNR is the condition in which no diffuse noise is added.

4.2. Localization performance

We compare the proposed system to the stencil filter approach proposed in [4] and the skeleton cross-correlogram approach proposed in [5]. For comparison on the database described, some alterations to the stencil filter method were necessary to account for the (somewhat) frequency-dependent nature of ITDs as detected by a binaural system and the discrete azimuth space. Further, because angles are assumed constant over the length of the mixture, azimuth responses from the stencil filter were integrated over all time frames for added accuracy and the N most prominent peaks were selected as the underlying source angles.

The average azimuth error (in °) is shown for all three methods as a function of T_{60} for two and three-talker mixtures in Table 1. We denote the skeleton cross-correlogram method as 'Skeleton CC', and the stencil filter method as 'Stencil'. We can see that the proposed approach estimates source angles within 5°, on average, in all but the three-talker case in 0.8 sec. T_{60}. In moderate reverberation (0.2 - 0.6 sec. T_{60}), the probability that the error was less than or equal to 5° is over 0.97 for the two-talker mixtures and nearly 0.95 for the three-talker mixtures using the proposed approach. Our approach is

Table 2. Average azimuth error (in °) in diffuse noise, 0.4 sec. T_{60}

One talker

SNR (dB)	**6**	**0**	**-3**	**-6**
Skeleton CC	11.98	19.8	25.83	30.85
Stencil	0.4	2.65	4.75	8.68
Proposed	0.95	2.43	4.38	7.9

Two talkers

SNR (dB)	**0**	**-1.5**	**-3**	**-6**
Skeleton CC	12.63	18.59	21.51	26.26
Stencil	3.21	6.71	8.64	15.36
Proposed	1.26	2.88	4.14	11.28

Three talkers

SNR (dB)	**-3**	**-4**	**-5**	**-6**
Skeleton CC	20.11	20.72	22.95	24.57
Stencil	8.7	11.38	14.52	14.81
Proposed	2.41	4.71	6.68	9.82

more accurate than the existing approaches in all but the two-talker, anechoic case, where all three methods achieve less than 1° average azimuth error. On average, the proposed method improves localization accuracy relative to the stencil filter approach by nearly 40% on the two-talker mixtures and over 60% on the three-talker mixtures.

We show the localization performance on one, two and three-talker mixtures in diffuse noise and 0.4 sec. T_{60} in Table 2. On average, the proposed method improves localization accuracy relative to the stencil filter approach by 42% on the two-talker mixtures in diffuse noise and 52% on the three-talker mixtures in diffuse noise. The performance between the two approaches is comparable on the one-talker mixtures. The localization error averaged less then 5° in all conditions in which the SNR for the speech sources was higher than -5 dB.

5. CONCLUDING REMARKS

We have presented a system for localization of multiple sound sources in noisy and reverberant conditions. The proposed approach utilizes monaural analysis to achieve simultaneous grouping, and estimates source azimuths and sequential organization in a maximum likelihood framework. We have shown that the incorporation of monaural grouping allows for robust localization in environments with diffuse noise and considerable reverberation. The proposed method outperforms two existing localization methods that exclusively use binaural cues.

The primary advantage of the proposed system is that binaural cues are not integrated over the entire mixture, as they are in the two existing systems used for comparison. The combination of monaural grouping and localization within the sequential organization framework integrates binaural cues over a subset of the mixture in which a single source is considered dominant. In this way, voiced speech segregation and localization are jointly achieved.

The results are made even more encouraging by noting that we do not utilize any unvoiced speech in the proposed framework. While this may be advantageous in the diffuse noise case, our proposed framework currently ignores much of the mixture. In general, one could expect localization performance to improve through inclusion of unvoiced speech regions simply because of the larger streams resulting from such inclusion.

Future work will need to allow for sources to change positions. Although research on localization of moving sources has been little [17], the assumption that sources remain stationary is a limitation of the current approach. We plan to extend our probabilistic framework to account for head movements and changes in source position.

6. REFERENCES

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

[2] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 24, no. 4, pp. 320–327, 1976.

[3] R. M. Stern, G. J. Brown, and D. L. Wang, "Binaural sound localization," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D. L. Wang and G. J. Brown, Eds., pp. 147–185. Wiley, 2006.

[4] C. Liu, B. C. Wheeler, W. D. O'Brien, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," *J. Acoust. Soc. Am.*, vol. 108, no. 4, pp. 1888–1905, 2000.

[5] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[6] M. Mandel, D. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Adv. Neural Info. Proc. Sys.*, 2007.

[7] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker, "A speech fragment approach to localising multiple speakers in reverberant environments," in *Proc. ICASSP*, 2009.

[8] J. Woodruff and D. L. Wang, "On the role of localization cues in binaural segregation of reverberant speech," in *Proc. ICASSP*, 2009.

[9] Guoning Hu, *Monaural Speech Organization and Segregation*, Ph.D. thesis, The Ohio State University, 2006.

[10] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Proc.*, 2010, in press.

[11] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, 1986.

[12] J. Blauert, *Spatial Hearing - The Psychophysics of Human Sound Localization*, MIT Press, 1997.

[13] K. W. Wilson and T. Darrell, "Learning a precedence effect-like weighting function for the generalized cross-correlation framework," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 14, pp. 2156–2164, 2006.

[14] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Am.*, vol. 106, pp. 1633–1654, 1999.

[15] Y. Shao and D. L. Wang, "Sequential organization of speech in computational auditory scene analysis," *Speech Commun.*, vol. 51, pp. 657–667, 2009.

[16] D. R. Campbell, "The ROOMSIM user guide (v3.3)," 2004.

[17] N. Roman and D. L. Wang, "Binaural tracking of multiple moving sources," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 16, pp. 728–739, 2008.