BINAURAL SPEECH SEGREGATION BASED ON PITCH AND AZIMUTH TRACKING

John Woodruff and DeLiang Wang

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

ABSTRACT

We propose an approach to binaural speech segregation in reverberation based on pitch and azimuth cues. These cues are integrated within a statistical tracking framework to estimate up to two concurrent pitch frequencies and three concurrent azimuth angles. The tracking framework implicitly estimates binary time-frequency masks by solving a data association problem, thereby performing speech segregation. Experimental results show that the proposed approach compares favorably to existing two-microphone systems in spite of less prior information. The benefit of the proposed approach is most pronounced in conditions with substantial reverberation or for closely spaced sources.

Index Terms— Computational auditory scene analysis, speech segregation, binaural localization, multipitch tracking

1. INTRODUCTION

Speech segregation is a challenging problem that has received considerable attention due to its potential application in hearing prostheses, robust speech recognition or audio information retrieval. When multiple microphones are available, the most ubiquitous approach to signal enhancement is beamforming, which filters and sums the received signals in order to create a spatially-dependent attenuation pattern [1]. Alternatively, blind source separation methods have been developed to segregate a known number of sources (see [2–4] as recent examples). While such methods have been widely utilized, there are shortcomings. Segregation performance degrades in reverberant environments and the assumption that sources are sufficiently well separated in space is not always met. Depending on the approach, there may also be constraints imposed on the number of sources, whether the number of sources or source positions can change across time, the bandwidth that can be effectively segregated (due to spatial aliasing), or the amount of interference attenuation that is possible.

In previous work we have developed a method that addresses some of these limitations by integrating monaural and binaural cues [5]. Building on recent advances in monaural computational auditory scene analysis (CASA) [6], this system performs segregation on the basis of both pitch and azimuth and we have demonstrated that this approach can improve localization of concurrent speech sources and segregation of voiced speech relative to exclusively binaural systems [5]. Related work has explored joint pitch and time delay estimation for a single source [7] and segregation based jointly on monaural and binaural cues [8–10].

In the current study we propose a binaural system for segregation of an unknown and time-varying number of sources. We incorporate an existing multipitch tracking system [11] such that localization and segregation are influenced by pitch cues. Whereas in our previous systems we assume a known number of sources [5], the framework proposed here estimates the number of active sources across time, the azimuth of each actives source, groups pitch and azimuth estimates, and generates a binary time-frequency (T-F) mask for each source. These problems are handled jointly using a novel hidden Markov model (HMM) framework.

In the following section we provide an overview of the proposed system. We describe the monaural and binaural features used in Section 3. In Section 4 we present the components of the HMM tracking and segregation framework. We perform an evaluation and comparison to existing two-microphone algorithms in Section 5 and conclude with a discussion in Section 6.

2. OVERVIEW

To perform segregation we seek to estimate the ideal binary mask (IBM), which has been proposed as a main goal of CASA [6]. The proposed approach to IBM estimation integrates pitch and azimuth evidence, tracked across time using an HMM. Segregation is performed within the tracking framework by solving a data association problem, which we accomplish with a set of trained multi-layer perceptrons (MLPs). By finding the optimal path through the multi-source pitch and azimuth states of the HMM, we estimate the pitch and azimuth of up to three sources and simultaneously generate a T-F mask for each source.

The cardinality of the HMM state space is prohibitively large, and thus it is necessary to constrain the number of states considered by the model in each frame. In this study we choose to do so be ignoring any HMM states that are inconsistent with the pitch estimates generated by a recent multipitch tracker [11]. Since there are numerous alternative ways in which one could constrain the state space of the HMM, we present the framework in its most general form in Section 4, but describe how the current instantiation incorporates estimates from the multipitch tracker in Section 4.4.

3. FEATURE EXTRACTION

We assume a binaural input signal sampled at a rate of 44.1 kHz. The binaural signal is analyzed using a bank of 64 gammatone filters with center frequencies from 80 to 5000 Hz spaced on the equivalent rectangular bandwidth scale. Each bandpass filtered signal is divided into 20 ms time frames with a frame shift of 10 ms to create a *cochleagram* [6] of time-frequency (T-F) units. We denote a T-F

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

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

unit as $u^E_{c,m}$, where m and c index time frames and filter channels, respectively, and $E \in \{L, R\}$ indicates the left or right ear signal.

To generate pitch-related features we first compute the correlogram and envelope correlogram [6] for both the left and right signals after downsampling to 16 kHz. The correlogram, denoted $A^E(c, m, \gamma)$, is a normalized running auto-correlation performed in individual frequency channels for each time frame and is thus a three-dimensional function of frequency channel (c), time frame (m) and lag time (γ). The envelope correlogram, denoted $\bar{A}^E(c, m, \gamma)$, is the same, but envelope extraction is performed prior to computation of the auto-correlation. We use a low-pass filter with 500 Hz cutoff frequency and a Kaiser window to extract signal envelopes. The four-dimensional pitch feature for each T-F unit and lag time is then: $\chi_{c,m}(\gamma) = \{A^L(c, m, \gamma), A^R(c, m, \gamma), \bar{A}^L(c, m, \gamma), \bar{A}^R(c, m, \gamma)\}$.

The binaural features calculated are the interaural time difference (ITD) and the interaural level difference (ILD). We calculate ITD, denoted $\tau_{c,m}$, as the maximum peak in a running cross-correlation between T-F units $u^L_{c,m}$ and $u^R_{c,m}$, where we consider time lags between -1 and 1 ms. ILD, denoted $\lambda_{c,m}$, corresponds to the energy ratio in dB between $u^L_{c,m}$ and $u^R_{c,m}$. Both values are calculated as described in [5].

4. TRACKING AND SEGREGATION

We assume a discrete grid of possible pitch and azimuth states. We consider pitch lags from 32 to 200 samples (16 kHz sample rate), which correspond to frequencies between 80 and 500 Hz. As sources may also be unvoiced, the pitch state space for a single source is $\gamma_k \in \{\emptyset, 32, ..., 200\}$. We consider azimuths in steps of 5° from $-90°$ to $90°$ and allow sources to be inactive such that the azimuth state space for a single source is $\theta \in \{\emptyset, -90°, -85°, ..., 90°\}$. Each multisource state contains three azimuth and pitch states, denoted $S = \{\theta_1, \theta_2, \theta_3, \gamma_1, \gamma_2, \gamma_3\}$.

The HMM framework is used to model the posterior probability of a multisource state given all observed monaural and binaural data,

$$p(S_m|T_{1:m}, \Lambda_{1:m}, X_{1:m}), \tag{1}$$

where we use T_m and Λ_m to denote the full set of ITD and ILD features for frame m, respectively, and use X_m to denote the full set of 4-dimensional pitch features for frame m. The subscript $_{1:m}$ denotes a collection of features from frame 1 through frame m.

In order to calculate the posterior, we must compute the observation likelihood, denoted $p(T_m, \Lambda_m, X_m|S_m)$, and the state transition probabilities, denoted $p(S_m|S_{m-1})$. Further, we perform data association between T-F units and individual sources to facilitate computation of the likelihood and generation of T-F masks. These three components are described in the following subsections.

4.1. Observation likelihood

To construct the multisource observation likelihood we first assume conditional independence between pitch and azimuth features such that,

$$p(T_m, \Lambda_m, X_m|S_m) = (p(T_m, \Lambda_m|S_m)p(X_m|S_m))^{\xi(S_m)}, \tag{2}$$

where $\xi(S)$ is used to adjust the likelihoods based on the number of active sources contained in the multisource state. Multisource states with more sources will produce higher likelihoods due to an increased flexibility in the data association described in Section 4.3. We set $\xi(S)$ to minimize any systematic bias towards overestimating of the number of sources. Based on a validation set, we set $\xi(S)$ to 1, 1, 1.05 and 1.08 for the cases with 0, 1, 2 and 3 active sources contained in S, respectively.

Table 1. Single source state transition probabilities. Rows 1, 2 and 3 list transitions out of voiced, unvoiced and inactive states, respectively. Columns 1, 2 and 3 list transitions into voiced, unvoiced and inactive states, respectively.

	$p(\theta, \gamma|\cdot)$	$p(\theta, \emptyset|\cdot)$	$p(\emptyset, \emptyset|\cdot)$
$p(\cdot|\theta', \gamma')$	$P_{\sim d}P_{vv}g(\gamma|\gamma')f(\theta|\theta')$	$P_{\sim d}P_{vu}f(\theta|\theta')$	P_d
$p(\cdot|\theta', \emptyset)$	$P_{\sim d}P_{uv}p(\gamma)f(\theta|\theta')$	$P_{\sim d}P_{uu}f(\theta|\theta')$	P_d
$p(\cdot|\emptyset, \emptyset)$	$P_bP_vp(\gamma)p(\theta)$	$P_bP_{\sim v}p(\theta)$	$P_{\sim b}$

In keeping with the assumption made in the formulation of the IBM, we assume that each frequency channel is dominated by a single source and associate T-F units with one of the underlying sources in each hypothesized multisource state. This allows us to decompose frame-based observation likelihoods conditioned on the properties of multiple sources into channel-based likelihoods conditioned on the properties of a single source, which are easier to model. We describe the data association method used in Section 4.3, but for now, let $y_c(S_m) \in \{1, 2, 3\}$ denote an assignment of channel c to one of the three sources contained in S_m. We then calculate the azimuth and pitch likelihoods using,

$$p(T_m, \Lambda_m|S_m) = \prod_c p_c(\tau_{c,m}, \lambda_{c,m}|\theta_{y_c(S_m)}), \tag{3}$$

$$p(X_m|S_m) = \prod_c p_c(\chi_{c,m}(\gamma_{y_c(S_m)})|\gamma_{y_c(S_m)}). \tag{4}$$

Note that only voiced sources are allowed to contribute to the pitch likelihood in Equation (4). When all sources are inactive, we set $p(T_m, \Lambda_m|S_m)$ and $p(X_m|S_m)$ to 0.03 based on a validation set. The channel-based likelihood functions, $p_c(\tau, \lambda|\theta)$ and $p_c(\chi(\gamma)|\gamma)$, are modeled using multi-layer perceptrons (MLPs). The MLP output can be interpreted as a posterior probability, and assuming a uniform prior over source states, is proportional to the likelihood. The data and procedures used to train MLPs are described in Section 5.2.

4.2. State transition probabilities

We define the multisource state transition probabilities assuming independence between sources, or,

$$p(S_m|S_{m-1}) = \prod_k p(\theta_{m,k}, \gamma_{m,k}|\theta_{m-1,k}, \gamma_{m-1,k}). \tag{5}$$

We list individual state transition probabilities in Table 1 where P_b and P_d are birth and death probabilities, respectively, $f(\theta|\theta')$ denotes the azimuth transition probability, $g(\gamma|\gamma')$ denotes the pitch transition probability, P_v is the prior probability of a source being voiced, and P_{vv} and P_{uu} are the voiced-voiced and unvoiced-unvoiced transition probabilities, respectively. P_b, P_d, $f(\theta|\theta')$ and $p(\theta)$ are related to source activity, source motion and listener movements, whereas P_v, P_{vv}, P_{uu}, $g(\gamma|\gamma')$ and $p(\gamma)$ capture general properties of speech. In the current study we do not consider source movement and thus set $f(\theta|\theta') = 1$ if $\theta = \theta'$ and $f(\theta|\theta') = 0$ otherwise, and $p(\theta) = \frac{1}{|\theta|-1}$. Based on a small validation set we set $P_b = 0.01$ and $P_d = 0.03$. Based on a small set of clean utterances from the TIMIT corpus [12], we set $P_v = 0.71$, $P_{vv} = 0.97$, $P_{uu} = 0.91$. Following [11] we use a Laplacian distribution with mean 0.4 and standard deviation 2.4 for $g(\gamma|\gamma')$, and set $p(\gamma) = \frac{1}{|\gamma|-1}$. Finally, $P_{\sim b} = 1 - P_b$, $P_{\sim d} = 1 - P_d$, $P_{\sim v} = 1 - P_v$, $P_{vu} = 1 - P_{vv}$ and $P_{uv} = 1 - P_{uu}$.

4.3. Data association

The data association stage has two main functions. First, by assigning T-F units to a single source for each multisource state, identifying a path through the multisource states across time allows us to generate a binary T-F mask for each source. Second, as state above, it allows us to utilize single-source models in the observation likelihoods (see Equations (3) and (4)). One of the simplest approaches to data association for multitarget tracking is to assign measurements based on the posterior probability that a given source generated the measurement. For a given multisource state hypothesis, S_m, we let $p_c(k|\tau_{c,m}, \lambda_{c,m}, \chi_{c,m}(\gamma_k))$ denote the posterior probability that a source with azimuth θ_k and pitch γ_k generated the monaural and binaural observations calculated from T-F units $u^L_{c,m}$ and $u^R_{c,m}$. We then perform data association according to,

$$y_c(S_m) = \underset{k \in \{1,2,3\}}{\arg\max} \left[p_c(k|\tau_{c,m}, \lambda_{c,m}, \chi_{c,m}(\gamma_k)) \right]. \qquad (6)$$

We train a set of multi-layer perceptrons (MLPs) to model $p_c(k|\tau_{c,m}, \lambda_{c,m}, \chi_{c,m}(\gamma_k))$. For unvoiced states (i.e. $\theta \neq \emptyset$ and $\gamma = \emptyset$), the models ignore the correlogram features and consider only ITD and ILD. For voiced states (i.e. $\theta \neq \emptyset$ and $\gamma \neq \emptyset$), the models consider both monaural and binaural features. The data and procedures used to train these models are described in Section 5.2.

4.4. Implementation details

As stated in Section 2, we incorporate pitch estimates generated by the system proposed in [11] to limit the multisource state space. This system generates up to two pitch estimates per frame. Each estimated pitch can be assigned to any of the three possible sources such that each set of pitch estimates yields up to six pitch configurations. We run the pitch tracker on both left and right signals such that the total number of pitch configurations is constrained to be less than or equal to twelve. Any multisource state that does not contain one of the identified pitch configurations is then ignored by the model. To further reduce complexity, we also incorporate beam search where the beam width is set to 500 states.

We use the Viterbi algorithm to determine the optimal path through the multisource state space. As discussed in Section 4.3, segregation is naturally performed within the tracking system through the data association process. However, one still must identify which (if any) of the active sources contained in the selected multisource states correspond to the target source. Resolving this problem is highly application dependant and proposing a robust solution is beyond the scope of the current study. For the evaluation presented here we associate azimuth and pitch estimates with the target source when the azimuth estimate is within a predetermined bound around the known target azimuth. This essentially gives the system a predetermined "look direction" with an error tolerance. In the tests below we set the error tolerance to be 10° around the known target direction (except for the test cases below when sources are only 5° apart, where we require azimuth estimates to match the target azimuth).

5. EVALUATION

5.1. Binaural simulation

For both the training and evaluation databases, we generate binaural mixtures that simulate pickup of multiple speech sources in a reverberant space. In all cases we use the ROOMSIM package [13] to generate binaural impulse responses (BIRs). Monaural speech signals are drawn from the TIMIT database [12] and passed through a BIR for a specified angle and room condition. Room size, microphone position and microphone orientation are selected randomly and the reflection coefficients of wall surfaces are set to be equal and to be the same across frequency. The dependent parameters for each mixture are then: number of sources, source azimuths, source distances and room reverberation time (T_{60}).

5.2. Model training

To train the MLPs described in Sections 4.1 and 4.3, we generate a set of 25 mixtures for each of the azimuths considered by the system. The number of interfering talkers, interference azimuths, source distances and mixture T_{60} are selected randomly. For each training mixture we extract features as described in Section 3. We also calculate the IBM as proposed in [14] and generate the ground truth pitch by running the pitch estimation method proposed in [15] on the clean target signal.

The pitch-based MLPs, used in Equation (4), are trained on the 4-dimensional correlogram features corresponding to the ground truth pitch period of the target source, where the IBM is used to provide the ground truth classification label for each T-F unit. We pool data across all azimuths and train a single MLP for each frequency channel. The azimuth-based MLPs, used in Equation (3), are trained on the ITD and ILD data, where again the IBM provides the classification label for each T-F unit. In this case we train a separate MLP for each frequency channel and azimuth. Finally, one set of MLPs for unvoiced speech and one set of MLPs for voiced speech are used for data association, described in Section 4.3. MLPs for unvoiced speech are the same as those used for the binaural likelihood. MLPs for voiced speech are trained on the ITD, ILD and 4-dimensional correlogram features corresponding to the ground truth pitch period and again the IBM is used to provide the classification label.

For simplicity each MLP has the same network topology consisting of a hidden layer with 30 nodes, and sigmoid transfer functions for both hidden and output nodes. Training is accomplished using a generalized Levenberg-Marquardt backpropagation algorithm.

5.3. Experimental design

We generate 25 binaural mixtures for 12 different evaluation conditions. In all cases a target source is randomly selected from TIMIT and placed at 0° azimuth. We generate a set of two-talker mixtures where the interference source is placed at 5°, 15°, 30° or 45°, and where T_{60} is set to either 0.4 s or 0.6 s. We generate a set of three-talker mixtures where interfering sources flank the target source at a distance of 15° or 30°, and where T_{60} is set to either 0.4 s or 0.6 s. All sources are placed 2 m from the microphone array. Sources are set to have equal power prior to spatialization.

We measure segregation performance in terms of change in signal-to-noise ratio (ΔSNR) relative to the mixture signal, where SNR is averaged across left and right signals. We compare to three existing two-channel methods. The first is an idealized minimum variance distortionless beamformer (MVDR) [1]. In our implementation we calculate covariance matrices from the clean target and

Table 2. Avg. ΔSNR (in dB) for all systems and test conditions.

	Two talkers				Three talkers	
$T_{60} = 0.4$ s	5°	15°	30°	45°	15°	30°
IBM	9.5	9.6	10	11.3	8.3	8.4
Ideal MVDR	3.5	4.8	7.0	7.6	3.8	4.5
Duong et al.	2.9	4.0	3.9	4.0	3.2	3.7
MESSL	1.7	6.1	9.0	9.9	4.9	6.5
Proposed	**5.7**	**8.3**	**9.3**	**10.4**	**6.5**	**7.4**
$T_{60} = 0.6$ s	5°	15°	30°	45°	15°	30°
IBM	8.7	8.8	9.6	9.9	8.3	8.3
Ideal MVDR	3.2	4.2	4.9	5.3	3.5	4.0
Duong et al.	2.7	3.6	3.4	3.5	3.3	3.6
MESSL	2.8	5.7	7.0	7.9	5.0	6.2
Proposed	**4.9**	**7.1**	**7.8**	**8.5**	**6.1**	**6.8**

residual signals, and thus this method represents the upper bound performance obtainable by a MVDR. We process 16 kHz mixture signals (downsampled from 44.1 kHz) through a 256 channel linear filterbank with a decimation factor of 64 samples. We also compare our method to recent segregation methods presented in [3,4]. These methods assume the number of sources are known *a priori* and that sources are spatially stationary.

5.4. Results

In Table 2 we show the average ΔSNR for the proposed and comparison systems for each of the 12 evaluation conditions. We also show the ΔSNR achieved by the IBM as a point of reference. As one would expect, the ideal MVDR is able to achieve much larger SNR gains for mixtures with two talkers that are well separated in space because it is able to create a deeper null in the interference direction. As reverberation increases, sources are spaced more closely or the number of talkers is increased, the beamformer is less effective. The Duong *et al.* system is an iterative implementation of the multichannel Wiener filter [4]. This approach has the potential to yield larger SNR gains due to the combination of a beamformer and post-filter, but the system does not perform well on our evaluation set because it is unable to effectively resolve the across-frequency permutation ambiguity with such a large distance between microphones (roughly 18 cm). The MESSL system, proposed in [3], clearly outperforms the other comparison methods and is capable of achieving large gains in SNR when sources are well separated in space. This is notable particularly because MESSL requires very little prior training and is still capable of handling substantial spatial aliasing.

We can see that the proposed system outperforms the comparison methods in all conditions. The improvement is largest in cases where sources are not spaced sufficiently far apart given the reverberation time and thus considering pitch and azimuth together allows one to more robustly detect, localize and therefore segregate sources. The benefit of the pitch information is most pronounced in lower frequency channels where, because wavelengths are larger than the microphone distance, spatial information is a weak grouping cue.

6. CONCLUDING REMARKS

The evaluation results show that the proposed system is capable of improving speech segregation relative to existing binaural systems. Improvement is due to jointly considering pitch and azimuth for segregation. As one might expect, improvements are largest in conditions with long reverberation times or with closely spaced sources.

The improvement of the proposed system relative to the comparison methods of [3,4] is particularly notable given that these methods both assume the number of sources is known, while sources are detected by the proposed method. Further, while [3,4] assume sources are spatially stationary, it is possible to extend the proposed method to deal with moving sources by including a motion model.

7. REFERENCES

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

[2] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech Audio Proc.*, vol. 13, no. 1, pp. 120–134, January 2005.

[3] 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.

[4] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 18, pp. 1830–1840, 2010.

[5] 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.*, vol. 18, pp. 1856–1866, 2010.

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

[7] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Joint DOA and fundamental frequency estimation methods based on 2-d filtering," in *Proc. EUSIPCO*, 2010.

[8] W. S. Woods, M. Hansen, T. Wittkop, and B. Kollmeier, "A simple architecture for using multiple cues in sound separation," in *Proc. ICSLP*, 1996.

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

[10] S. N. Wrigley and G. J. Brown, "Binaural speech separation using recurrent timing neural networks for joint F0-localisation estimation," in *Machine Learning for Multimodel Interaction*, pp. 271–282. Springer Berlin / Heidelberg, 2008.

[11] Z. Jin and D. L. Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 19, pp. 1091–1102, 2011.

[12] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic phonetic continuous speech corpus," 1993.

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

[14] N. Roman and J. Woodruff, "Intelligibility of reverberant noisy speech with ideal binary masking," *J. Acoust. Soc. Am.*, vol. 130, pp. 2153–2161, 2011.

[15] P. Boersma, "Accurate short-time analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Inst. of Phonetic Sci.*, 1993, vol. 17, pp. 97–110.