EM LOCALIZATION AND SEPARATION USING INTERAURAL LEVEL AND PHASE CUES

Michael I. Mandel and Daniel P. W. Ellis[*]

LabROSA, Dept. of Electrical Engineering
Columbia University
{mim,dpwe}@ee.columbia.edu

ABSTRACT

We describe a system for localizing and separating multiple sound sources from a reverberant stereo (two-channel) recording. It consists of a probabilistic model of interaural level difference and interaural phase difference and an EM algorithm for finding the maximum likelihood parameters of the model. By assigning points in the interaural spectrogram to the source with the best-fitting parameters and then estimating the parameters of the sources from the points assigned to them, the system is able to both separate and localize more sound sources than available channels. It is able to estimate frequency-dependent level differences from an synthetic mixture that correspond well to the synthesis parameters. In experiments in simulated anechoic and reverberant environments, the proposed system was better able to enhance the signal-to-noise ratio of target sources than two comparable algorithms.

1. INTRODUCTION

A person can understand the speech of a conversational partner even in reverberant environments and even in the presence of noise distractions such as other speakers. In order for machines to approach the abilities of humans in this task – for instance to recognize spoken commands – they must be able to attend to a single speaker in such adverse conditions. This paper describes a source separation system addressing this problem, based on the binaural cues that humans use in this task.

The cues, interaural level difference (ILD) and interaural time or phase difference (ITD or IPD), are computed in our system from spectrograms of speech mixtures at the left and right ears of a KEMAR mannequin head. A key assumption is that each point in the interaural spectrogram is dominated by a single source. But unlike binary masking systems [1], our system creates true probabilistic masks. Unlike cross-correlation-based systems [2], our system makes no assumptions about the statistics of the sources. Unlike independent component analysis [3] and microphone arrays [4], our system can separate more sources than it has microphones. Unlike DUET [5], our system can use the entire available frequency range to estimate source parameters. Furthermore, our system does not require any training or pre-defined head-related impulse responses (HRIRs), and it works successfully in reverberation.

This paper adds the interaural level difference cue to the work presented in [6]. Not only does the addition of this cue greatly improve separation results, but it also shows the emergence of certain properties similar to human hearing. For example, humans are known to use ITD for low frequencies but rely mainly on ILD for high frequencies, a characteristic that this model shares not by design but as a result of its probabilistic nature and the distribution of information in the interaural spectrogram. We use the IPD to initialize the separation, which we then use to begin estimating the ILD. We can also use the ILD to separate sources in frequency bands for which IPD is ambiguous, thanks to the complementary nature of the cues.

Our system has no model of ILD, as might come from training on HRIRs. Instead, it simply relies on differences in ILD between sources to separate them and better infer the differences. ILD varies with frequency in HRIRs, a fact that we are able to exploit in our system by estimating a separate ILD at every frequency for each source. In order to avoid permuting the sources at each band, we progressively refine the ILD estimates starting with wide frequency bands which are slowly broken up into smaller bands over the course of the EM iterations.

We evaluate our system on the localization and separation of two and three simultaneous speakers in simulated anechoic and reverberant environments. In total, we used 300 different mixtures of speakers. The speech comes from the TIMIT acoustic-phonetic continuous speech corpus, the anechoic simulations use the head related transfer functions described in [7], and the reverberant simulations use the binaural classroom impulse responses described in [8]. For each separated mixture, we measure the signal-to-noise improvement effected by the resulting time-frequency mask, and the root mean squareerror of localization. Our EM approach outperformed Yilmaz and Rickard's DUET algorithm [5] and Aarabi's PHAT-histogram [2].

2. MODEL

We now describe the probabilistic model and an EM algorithm for estimating its parameters. For a more complete description of the model only using the IPD cue, please see [6].

Denote the sound source as $s(t)$, and the signals received at the left and right ears as $\ell(t)$ and $r(t)$, respectively. For a sufficiently narrowband source, the two received signals can be described simply by some delay and gain relative to the source, in addition to a disruption due to noise; we further expect the ITD to depend only weakly on frequency even for broader-band signals. In this model, we assume a convolutive noise process, because it fits our empirical observations, it is easy to analyze, and in general is it is very similar to the additive noise processes that other authors assume

[*] This work was supported by the Fu Foundation School of Engineering and Applied Science via a Presidential Fellowship, and by the National Science Foundation (NSF) under Grants No. IIS-0238301 and IIS-0535168. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

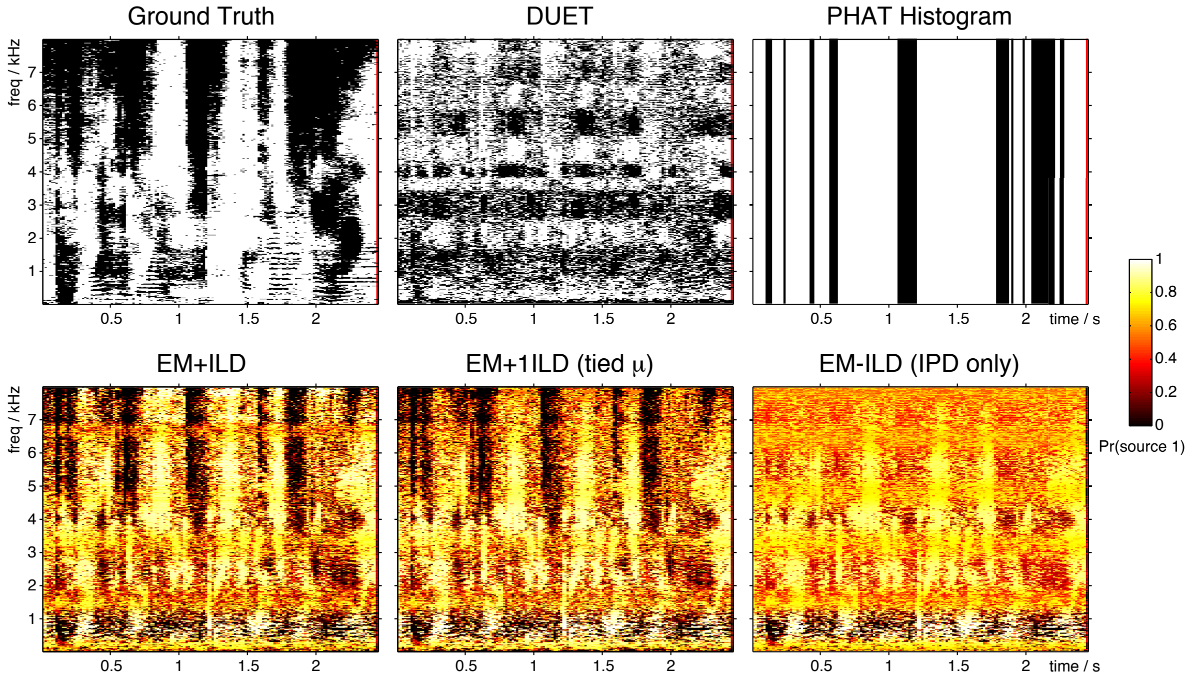

Figure 1: Example mask created by each algorithm for the same source in the same mixture of two sources separated by 75° in reverberation.

[9]. The various signals are then related by:

$$\ell(t) = a_\ell s(t - \tau_\ell) * n_\ell(t) \quad (1)$$

$$r(t) = a_r s(t - \tau_r) * n_r(t). \quad (2)$$

The ratio of the short-time Fourier transforms, $\mathcal{F}\{\cdot\}$, of both equations is the interaural spectrogram,

$$X_{IS}(\omega, t) \equiv \frac{L(\omega, t)}{R(\omega, t)} = 10^{\alpha(\omega,t)/20} e^{\phi(\omega,t)} \quad (3)$$

$$\approx 10^{a/20} e^{-j\omega\tau} N(\omega, t), \quad (4)$$

where $\tau = \tau_\ell - \tau_r$, $N(\omega, t) = \frac{N_\ell(\omega,t)}{N_r(\omega,t)} = \frac{\mathcal{F}\{n_\ell(t)\}}{\mathcal{F}\{n_r(t)\}}$, and $a = 20 \log_{10} \frac{a_\ell}{a_r}$. This equivalence assumes that τ is much smaller than the length of the window over which the Fourier transform is taken, a condition easily met for dummy head recordings with moderately sized Fourier transform windows. In our experiments the maximum delay was 0.75ms, and the window length was 64ms.

As observed in [9], $N(\omega, t)$, the noise in the interaural spectrogram of a single source is unimodal and approximately identically distributed for all frequencies and times. Using the standard rectangular-to-polar change of coordinates, the noise can be separated into independent magnitude and phase components. The magnitude noise is approximately log-normal, while the phase noise has a circular distribution with tails heavier than the von Mises distribution. We approximate the phase noise as a single Gaussian, ignoring its heavy tails and circularity because of its concentration around 0. We also model the ILD measured in dB as Gaussian, with a different mean in different frequency bands.

From the mixed signals, we observe $\alpha(\omega, t)$ and $\phi(\omega, t)$, the level and phase differences between the left and right channels at frequency ω and time t. To sidestep issues of 2π ambiguity in the phase, we define a transformed phase variable $\hat{\phi}$ as the the difference between observed IPD and that expected for a delay of τ seconds:

$$\hat{\phi}(\omega, t; \tau) = \arg\left(e^{\phi(\omega,t)} e^{-j\omega\tau}\right) \quad (5)$$

which is always in the interval $(-\pi, \pi]$. For conciseness we drop the (ω, t) from these expressions and just use α and $\hat{\phi}(\tau)$.

Our model of IPD and ILD is a mixture over sources and delays. In particular, we have I sources, indexed by i, each of which has a distribution over delays, τ. The delays are discretized to a grid and probabilities over them are computed as a multinomial. For a particular source, the probability of an observed delay is modeled as:

$$p(\phi, \alpha \,|\, i, \tau) = \mathcal{N}(\hat{\phi}(\tau) \,|\, 0, \sigma_i^2) \mathcal{N}(\alpha \,|\, \mu_{i\omega}, \eta_i^2). \quad (6)$$

In order to allow parameter estimation, we define hidden indicator variables $z_{i\tau}^{\omega t}$ such that $z_{i\tau}^{\omega t} = 1$ if cell (ω, t) comes from source i at delay τ, and 0 otherwise. There is one indicator for each observation, so $\sum_{i,\tau} z_{i\tau}^{\omega t} = 1$ and $z_{i\tau}^{\omega t} \geq 0$. The estimated parameters of our model for each source i are thus the mean ILD $\mu_{i\omega}$ for frequency ω, the variance of the ILD noise η_i^2, the variances of the phase noise σ_i^2, and $\psi_{i\tau} \equiv p(i, \tau)$, a vector of probabilities for each ITD under consideration. For convenience, we define $\theta \equiv \{\mu_{i\omega}, \eta_i, \sigma_i, \psi_{i\tau} \quad \forall i, \tau, \omega\}$.

Thus, the total log-likelihood of our data, including marginalization over the hidden variables, is:

$$\log p(\phi, \alpha \,|\, \theta) = \sum_{\omega,t} \log \sum_{i,\tau} \psi_{i\tau} \mathcal{N}(\hat{\phi}(\tau) \,|\, 0, \sigma_i^2) \mathcal{N}(\alpha \,|\, \mu_{i\omega}, \eta_i^2). \quad (7)$$

This log likelihood allows us to derive the E and M steps of our algorithm. For the E step, we compute the expected value of

$z_{i\tau}^{\omega t}$ given the data and our current parameter estimates,

$$\nu_{i\tau}(\omega,t) \equiv E\{z_{i\tau}^{\omega t} \mid \phi, \alpha, \theta\} = \frac{p(z_{i\tau}^{\omega t}=1, \phi, \alpha \mid \theta)}{p(\phi, \alpha \mid \theta)}$$
$$= \frac{\psi_{i\tau}\mathcal{N}(\hat{\phi}(\tau) \mid 0, \sigma_i^2)\mathcal{N}(\alpha \mid \mu_{i\omega}, \eta_i^2)}{\sum_{i,\tau}\psi_{i\tau}\mathcal{N}(\hat{\phi}(\tau) \mid 0, \sigma_i^2)\mathcal{N}(\alpha \mid \mu_{i\omega}, \eta_i^2)} \quad (8)$$

The M step updates the model parameters using the above expected membership

$$\mu_{i\omega} = \frac{\sum_t \sum_\tau \nu_{i\tau}(\omega,t)\alpha}{\sum_t \sum_\tau \nu_{i\tau}(\omega,t)} \quad (9)$$

$$\eta_i^2 = \frac{\sum_{\omega,t}\sum_\tau \nu_{i\tau}(\omega,t)(\alpha-\mu_{i\omega})^2}{\sum_{\omega,t}\sum_\tau \nu_{i\tau}(\omega,t)} \quad (10)$$

$$\sigma_i^2 = \frac{\sum_{\omega,t}\sum_\tau \nu_{i\tau}(\omega,t)\hat{\phi}(\tau)^2}{\sum_{\omega,t}\sum_\tau \nu_{i\tau}(\omega,t)} \quad (11)$$

$$\psi_{i\tau} = \frac{1}{\Omega T}\sum_{\omega,t}\nu_{i\tau}(\omega,t). \quad (12)$$

where Ω is the number of frequency bands and T is the number of time steps.

We factor $\psi_{i\tau} = p(i,\tau) = p(\tau \mid i)p(i)$ into the marginal probability of a point's coming from source i, and the distribution over delays for each source. From the indicator variables, we can also derive the probability of each time-frequency point's coming from each source, $M_i(\omega,t)$. To calculate these masks, we marginalize $p(z_{i\tau}^{\omega t} \mid \phi, \alpha, \theta)$ over τ to get

$$M_i(\omega,t) = \sum_\tau \nu_{i\tau}(\omega,t). \quad (13)$$

3. EXPERIMENTS

In order to evaluate our system, we simulated speech in anechoic and reverberant situations by convolving anechoic speech samples with binaural impulse responses. We used 15 utterances chosen at random from the TIMIT corpus. To balance the speakers in each mixture, we normalized all of the signals by their average energies before convolving them with the binaural impulse responses.

The anechoic binaural impulse responses came from Algazi et al. [7], a large effort to record head-related transfer functions for many different individuals. Impulse response measurements were taken over the sphere surrounding subjects' heads at 25 different azimuths and 50 different elevations. The measurements we used were for the KEMAR dummy head with small ears, although the dataset contains impulse responses for around 50 individuals.

We used reverberant binaural impulse responses recorded by Shinn-Cunningham et al. in a real classroom [8]. These measurements were also made with a KEMAR dummy head, although a different actual unit was used. Measurements were taken from four different positions in the classroom, at three distances from the subject, seven directions, and with three repetitions of each measurement. We used the measurements taken in the middle of the classroom with the sources at a distance of 1 m from the subject.

In the experiments, the target speaker was located directly in front of the listener, with distractor speakers located off to the sides. The angle between the target and distractors was systematically varied and the results combined for each direction. In the anechoic setting, there were 12 different angles at which we placed the distractors. In the reverberant setting, there were 6 different angles, but 3 different impulse responses for each angle, for a total of 18 angles. Each setup was tested with 5 different randomly-chosen sets of speakers and with one and two distractors, for a total of 300 different mixtures.

We measure the performance of localization with the root-mean-square error from the ground truth time delay, ignoring any ILD cues in its calculation except through their coupling with IPD. The ground truth was determined by passing white noise through each impulse response pair separately and measuring the cross correlation.

We measure the performance of separation with the signal-to-noise ratio (SNR) improvement due to the mask. Before masking, the SNR is the ratio of energy in the target signal to energy in the distractor(s). After masking, it is measured as ratio of the energy in the target signal that is passed through the mask to the difference between the true target signal and the total signal passed through the mask. This measurement of noise includes both noise that is passed through the mask and signal that is rejected by the mask. Both SNRs are measured in dB, and the SNR improvement is the difference of these two numbers, also in dB. We use the SNR improvement instead of the final SNR becase even after normalizing the anechoic sources there are level variations in the spatialized utterances; the SNR improvement shows less variation as a result of these level differences than the final SNR.

4. RESULTS

The results of these experiments are shown in Figure 2. In an anechoic setting, our system is able to separate sources almost as well as the ground truth 0 dB hard mask. On average, the proposed system achieves an SNR improvement 1.7 dB greater than the system from [6], 2.7 dB greater than that of [2], and 3.4 dB greater than that of [5]. The proposed algorithm localized sources as well as PHAT-histogram, which was much better than DUET, especially for multiple sources and reverberant environments. The performance of all of the systems, even the ground truth mask, improved with separation of the sources presumably because of the reduced cues overlap between sources.

Comparing the three versions of our system to each other, using a different ILD at each frequency slightly outperforms using just a single ILD, and both outperform the EM system using only IPD. While this is true on average, using a single ILD performs better in reverberation, while using different ILDs for each frequency works better in anechoic situations. Initialization of the algorithm is very important to the success of all of the variants. Even using both cues, performance suffers greatly when the initial positions of the sources are not well determined.

Figure 3 shows an example of the various ILD estimates from our algorithm. For two sources 75° apart, in a reverberant environment, the full ILD model was able to recover a remarkably accurate ILD for both sources. In certain frequency bands both estimates tend toward the overall mean – this generally happens where the IPD is ambiguous and so no other cue is available to break the symmetry between sources. This could probably be improved by running the algorithm for more iterations or tying the $\mu_{i\omega}$ parameters across frequency bands for more iterations before letting them separate from each other. Notice also that when just a single ILD is used for all frequencies, it settles to the mean of the ground truth ILD.

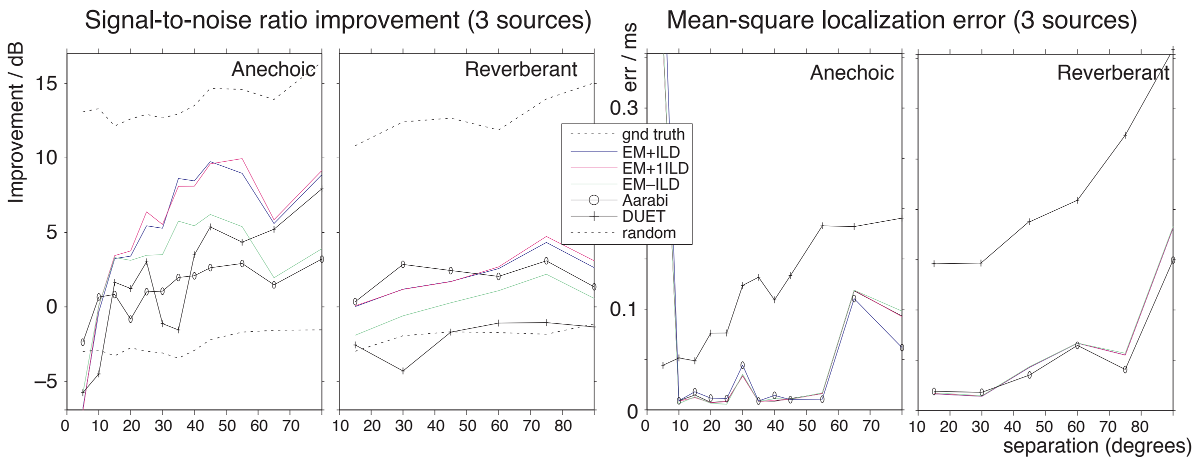

Figure 2: SNR improvement and MSE localization error comparison. "EM+ILD" is the full ILD/IPD based system; "EM+1LD" assumes that ILD is constant across frequency; "EM-ILD" is the IPD-only system; "Aarabi" is the approach from [2]; "DUET" is the approach from [5]; "gnd truth" is the optimal SNR from the oracle time-frequency mask; "random" uses a random mask.

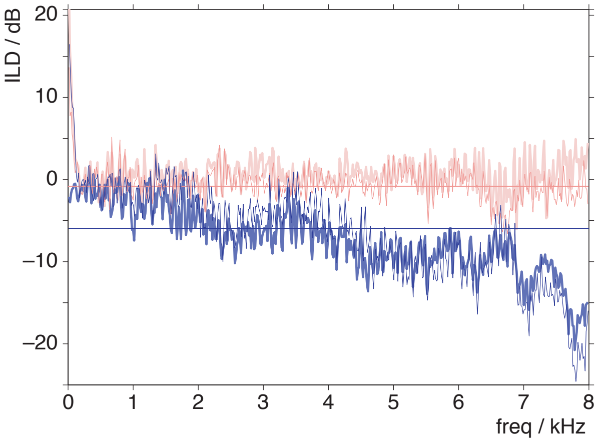

Figure 3: ILD as a function of frequency for two speakers separated by 75° in reverberation. The thin lines are our best ILD estimates for each source, the horizontal lines are ILD estimates when tied across frequency, and the thick lines are the ground truth ILDs. Light lines are speaker 1, dark for speaker 2.

5. CONCLUSIONS

We have presented a system for source separation and localization that avoids many of the limitations of other algorithms by building a fully probabilistic model of interaural spectrograms. It is able to separate more sources than it has available microphones.

The system takes advantage of interaural phase difference cues at low frequencies and interaural level difference cues at high frequencies. Because of the ambiguity of IPD, distinct sources will have the same IPD at certain frequencies that depend on the azimuth difference between them. For sources well separated, this ambiguity can exist at frequencies lower than those where ILD is effective at differentiating the two signals. We predict that such a phenomenon might also affect human hearing, i.e. narrowband sources at such an ambiguous, but low frequency, might not be separable from one another.

In the future, we would like to model ILD and IPD more closely together. This could include linking them both to azimuthal angle of arrival, as well as using measured head impulse responses. In our model, these constraints would take the form of a joint prior distribution over these parameters.

6. REFERENCES

[1] N. Roman, D. Wang, and G. J. Brown, "A classification-based cocktail party processor," in *Proceedings of Neural Information Processing Systems*, 2003.

[2] P. Aarabi, "Self-localizing dynamic microphone arrays," *IEEE transactions on systems, man, and cybernetics*, vol. 32, no. 4, November 2002.

[3] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, no. 94-128, pp. 3–1, 1999.

[4] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proceedings of IEEE Conference on Audio, Speech, and Signal Processing*, 1997, pp. 375–378.

[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, July 2004.

[6] M. Mandel, D. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007.

[7] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*, Oct 2001, pp. 99–102.

[8] B. Shinn-Cunningham, N. Kopco, and T. J. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *Journal of the Acoustical Society of America*, vol. 117, pp. 3100–3115, 2005.

[9] M. I. Mandel and D. P. W. Ellis, "A probability model for interaural phase difference," *Workshop on Statistical and Perceptual Audio Processing (SAPA)*, 2006.