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 two-channel recording. It consists of a probabilistic model of interaural level and phase differences and an EM algorithm for finding the maximum likelihood parameters of this model. By assigning points in the interaural spectrogram probabilistically to sources with the best-fitting parameters and then estimating the parameters of the sources from the points assigned to them, the system is able to separate and localize more sound sources than there are available channels. It is also able to estimate frequency-dependent level differences of sources in a mixture that correspond well to those measured in isolation. In experiments in simulated anechoic and reverberant environments, the proposed system improved the signal-to-noise ratio of target sources by 2.7 and $3.4dB$ more than two comparable algorithms on average.

1. INTRODUCTION

A person can understand the speech of a conversational partner even in reverberant environments and even in the presence of 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. It concludes with a prediction about human hearing derived from observations of this model.

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 in this model is that each point in the interaural spectrogram is dominated by a single source. This assumption allows our system to avoid many of the limitations of comparable systems. 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 estimate source parameters from all available frequencies. Furthermore, our system does not require any training 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 similarities to human hearing. For example, humans are known to use ITD for localization at low frequencies but rely mainly on ILD at high frequencies [7], a characteristic that this model shares not by design but as a result of the distribution of information in the interaural spectrogram.

Instead of modeling the relationship between ILD and location, our system relies on differences in sources' ILDs to separate them. We use the IPD to initialize the separation, and then estimate the ILD from the partially separated sources. The ILD can then be used to separate sources in frequency bands for which phase information is ambiguous. In head-related impulse responses, ILD varies with frequency and our model is able to estimate a separate ILD at every frequency for each source.

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 [8], and the reverberant simulations use the binaural classroom impulse responses described in [9]. For each separated mixture, we measure the signal-to-noise improvement affected by the resulting time-frequency mask, and the root mean square error of localization. Our EM approach outperformed Yilmaz and Rickard's DUET algorithm [5] and Aarabi's PHAT-histogram [2]. See Figure 1 for the masks extracted by each algorithm on an example mixture.

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, which, when combined with the single dominant delay, amounts to an unknown transfer function with a large impulse in it at a certain delay and small coefficients at all other delays. It also behaves very similarly to the additive noise processes that other authors assume [10]. The vari-

[*] This work was supported by the Fu Foundation School of Engineering and Applied Science via a Presidential Fellowship, and by the Columbia Academic Quality Fund, 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.

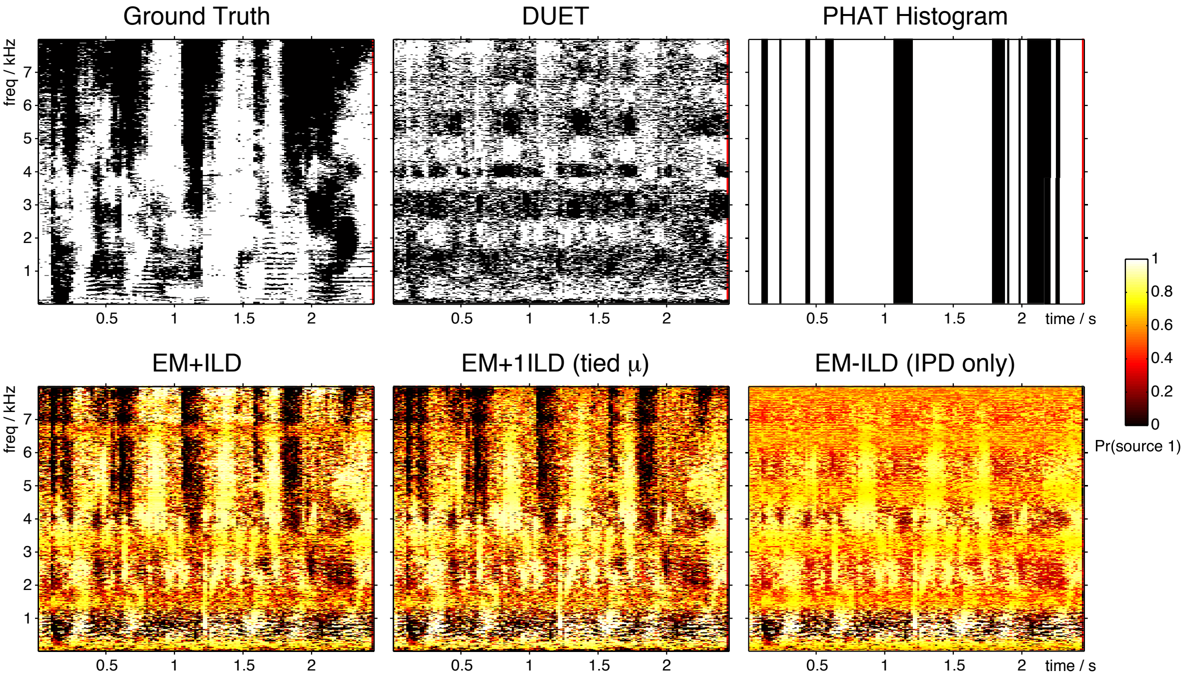

Figure 1: Masks extracted by each algorithm compared in the experiments for an example mixture containing two sources separated by 75° in reverberation. Color indicates the probability that a given time-frequency cell belongs to source 1 (white) or source 2 (black).

ous signals are then related by:

$$\ell(t) = a_\ell s(t-\tau_\ell) * n_\ell(t)$$
$$r(t) = a_r s(t-\tau_r) * n_r(t). \quad (1)$$

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

$$X_{IS}(\omega,t) \equiv \frac{L(\omega,t)}{R(\omega,t)} = 10^{\alpha(\omega,t)/20} e^{\phi(\omega,t)} \quad (2)$$
$$\approx 10^{a/20} e^{-j\omega\tau} N(\omega,t), \quad (3)$$

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 $a = 20\log_{10}\frac{a_\ell}{a_r}$. The observation in (2) is equal to the model prediction in (3) when τ is much smaller than the window over which the Fourier transform is taken. In our experiments the maximum delay was 0.75ms, and the window length was 64ms (1024 samples).

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 the phase residual $\hat{\phi}$ as the the difference between observed IPD and that expected for a delay of τ seconds:

$$\hat{\phi}(\omega,t;\tau) = \arg\left(e^{\phi(\omega,t)} e^{-j\omega\tau}\right) \quad (4)$$

which is always in the interval $(-\pi,\pi]$. By modeling this residual, we are free to evaluate the probability of a phase difference at any set of τs we choose. For conciseness we drop the (ω,t) from these expressions and just use α, ϕ, and $\hat{\phi}(\tau)$.

The characteristics of α and ϕ were studied in [10], which suggests the use of a Gaussian model of the ILD measured in dB and an independent Gaussian mixture model of the phase residual. We have found that in practice a single Gaussian can be used to model $\hat{\phi}(\tau)$. Combining these models, the likelihood of a delay of τ in the ith source is given by

$$p(\phi,\alpha\,|\,i,\tau) = \mathcal{N}(\hat{\phi}(\tau)\,|\,0,\sigma_i^2)\mathcal{N}(\alpha\,|\,\mu_i(\omega),\eta_i^2), \quad (5)$$

where $\mu_i(\omega)$ is the mean ILD for the ith source. Note that the total number of sources must be specified *a priori*. In the experiments we compare frequency-dependent and -independent means.

The 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. 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 i and τ, is:

$$\log p(\phi,\alpha\,|\,\theta) = \sum_{\omega,t}\log\sum_{i,\tau}\psi_{i\tau}\mathcal{N}(\hat{\phi}(\tau)\,|\,0,\sigma_i^2)\mathcal{N}(\alpha\,|\,\mu_i(\omega),\eta_i^2). \quad (6)$$

This log likelihood allows us to derive the E and M steps of our algorithm. To do so, we define hidden indicator variables $z_{i\tau}^{\omega t}$ such that $z_{i\tau}^{\omega t} = 1$ if point (ω,t) comes from source i at delay τ, and 0 otherwise. The E step computes the expected value of $z_{i\tau}^{\omega t}$ given the data and our current parameter estimates:

$$\nu_{i\tau}(\omega,t) \equiv E\{z_{i\tau}^{\omega t}\,|\,\phi,\alpha,\theta\} = \frac{p(z_{i\tau}^{\omega t}=1,\phi,\alpha\,|\,\theta)}{p(\phi,\alpha\,|\,\theta)}$$
$$= \frac{\psi_{i\tau}\mathcal{N}(\hat{\phi}(\tau)\,|\,0,\sigma_i^2)\mathcal{N}(\alpha\,|\,\mu_i(\omega),\eta_i^2)}{\sum_{i,\tau}\psi_{i\tau}\mathcal{N}(\hat{\phi}(\tau)\,|\,0,\sigma_i^2)\mathcal{N}(\alpha\,|\,\mu_i(\omega),\eta_i^2)} \quad (7)$$

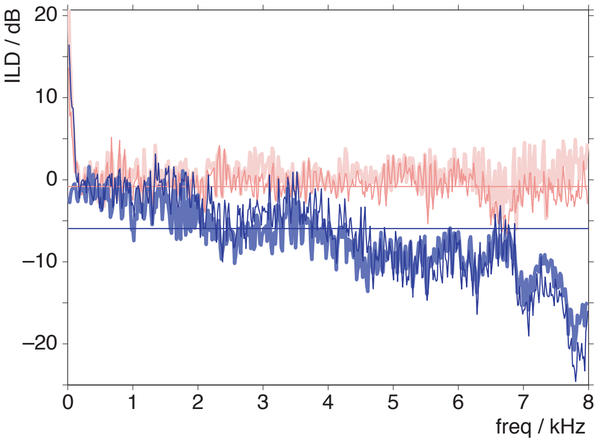

Figure 2: ILD as a function of frequency for two speakers separated by 75° in reverberation. The thin lines are our best ILD estimates for each source, the horizontal lines are ILD estimates when tied across frequency, and the thick lines are the ground truth ILDs. Light lines are speaker 1, dark for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

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

$$\mu_i(\omega) = \frac{\sum_t \sum_\tau \nu_{i\tau}(\omega,t)\alpha}{\sum_t \sum_\tau \nu_{i\tau}(\omega,t)} \tag{8}$$

$$\eta_i^2 = \frac{\sum_{\omega,t} \sum_\tau \nu_{i\tau}(\omega,t)(\alpha - \mu_i(\omega))^2}{\sum_{\omega,t} \sum_\tau \nu_{i\tau}(\omega,t)} \tag{9}$$

$$\sigma_i^2 = \frac{\sum_{\omega,t} \sum_\tau \nu_{i\tau}(\omega,t)\hat{\phi}(\tau)^2}{\sum_{\omega,t} \sum_\tau \nu_{i\tau}(\omega,t)} \tag{10}$$

$$\psi_{i\tau} = \frac{1}{\Omega T} \sum_{\omega,t} \nu_{i\tau}(\omega,t). \tag{11}$$

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

Factoring $\psi_{i\tau} = p(i,\tau) = p(\tau \mid i)p(i)$ gives the marginal probability of any point's coming from source i, and the distribution over delays for each source. Marginalizing $p(z_{i\tau}^{\omega t} \mid \phi, \alpha, \theta)$ over τ gives the probabilistic mask for source i.

$$M_i(\omega,t) = \sum_\tau \nu_{i\tau}(\omega,t). \tag{12}$$

3. EXPERIMENTS

We compare seven different source separators in the experiments. Three of them are based on the proposed EM algorithm, two of them are other comparable algorithms, and the last two are controls. The EM system with the most parameters, having a different mean ILD for each frequency, is denoted "EM+ILD" in the figures. The EM system that uses a single, frequency independent ILD mean for each source is "EM+1ILD." The EM system describe in [6] that does not use ILD is "EM-ILD." The comparable algorithms are "Aarabi" from [2] and "DUET" from [5]. These methods are all compared again separation by ground truth 0 dB masks, denoted "gnd truth" and random masks, denoted "random."

In order to evaluate these systems, we simulated speech in anechoic and reverberant conditions by convolving anechoic speech samples with binaural impulse responses. We used 15 utterances chosen at random from the TIMIT corpus. Before convolving them with the binaural impulse responses, all of the signals were normalized by their average energies and slightly high-pass filtered to whiten their spectra and avoid overemphasizing the low frequencies in our measurements of SNR improvement.

The anechoic binaural impulse responses came from Algazi et al. [8], a large effort to record head-related transfer functions for many different individuals. We use the measurements for a KEMAR dummy head with small ears, taken at 25 different azimuths at 0° elevation. We used reverberant binaural impulse responses recorded by Shinn-Cunningham et al. in a real classroom [9]. These measurements were also made with a KEMAR dummy head, although a different actual unit was used. The measurements we used were taken in the center of the classroom, with the source 1 m from the head at 7 different azimuths, each repeated 3 times.

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 response pairs for each angle, for a total of 18 conditions. 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 separation with signal-to-noise ratio improvement. The final and initial SNRs are

$$\mathcal{R}_f(i) = \frac{||M_i S_i||^2}{||S_i - M_i \sum_j S_j||^2} \quad \mathcal{R}_i(i) = \frac{||S_i||^2}{||\sum_{j \neq i} S_j||^2}. \tag{13}$$

Where M_i is the ith mask, S_i is the ith source spectrogram, and $||X||^2$ indicates summing the squared magnitude of X over ω and t. Both SNRs are converted to dB, and the SNR improvement is the difference of these two numbers, also in dB. This measure penalizes both noise that is passed through the mask and signal that is rejected by the mask. We use the SNR improvement instead of the final SNR because 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 alone.

We measure the performance of localization with the root-mean-square error from the ground truth time delay. Note that we ignore any ILD cues in this calculation except through their coupling with IPD. The ground truth was determined by passing white noise through each impulse response pair individually and measuring the cross correlation of these monophonic signals.

4. RESULTS

The results of these experiments are shown in Figure 3 and summarized numerically in Table 1. On average, EM+ILD achieves an SNR improvement 1.7 dB greater than EM-ILD, 2.7 dB greater than PHAT-histogram, and 3.4 dB greater than DUET. All of the EM systems localize sources as well as PHAT-histogram and much

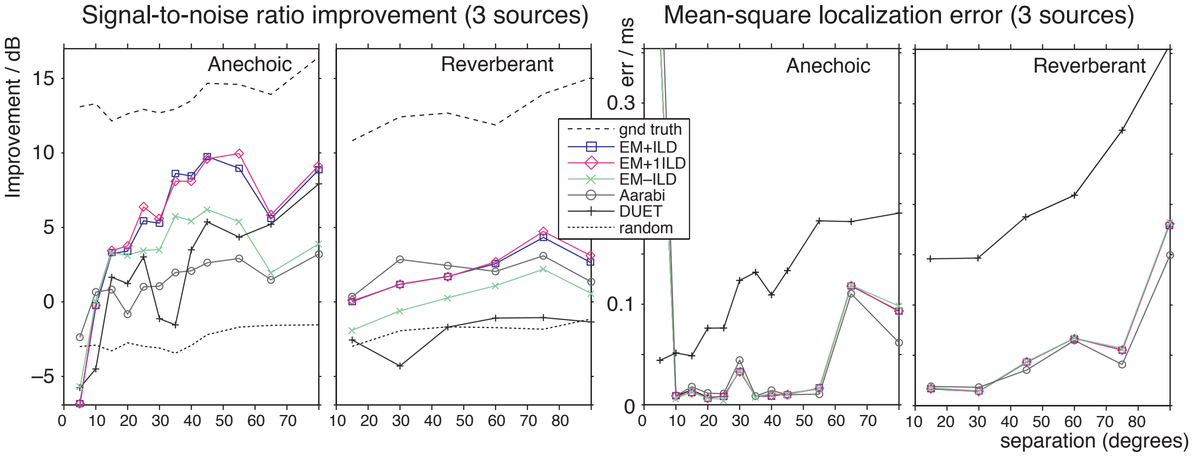

Figure 3: SNR improvement and MSE localization error comparison. Abbreviations described in Section 3

Table 1: Average SNR improvement in dB for each algorithm in each test case. The test cases are described by the number of simultaneous sources (2 or 3) and whether the impulse responses were anechoic or reverberant (A or R).

	2A	2R	3A	3R	Tot
Gnd Truth	12.97	12.10	13.25	12.45	12.69
EM+ILD	8.00	4.05	3.74	2.08	4.47
EM+1ILD	7.46	3.86	4.00	1.94	4.31
EM-ILD	5.58	1.97	2.87	0.70	2.78
Aarabi	1.07	2.14	1.34	2.46	1.75
DUET	4.53	-0.39	1.83	-1.73	1.06
Random	-2.72	-2.01	-2.63	-2.04	-2.35

better than DUET, especially for three 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 overlap of cues between sources. Note that the random mask has negative SNR improvement because it eliminates some power from the target signal.

Comparing the three versions of our system to each other, on average EM+ILD slightly outperforms EM+1ILD, and both outperform EM-ILD. Looking at particular cases, however, EM+1ILD performs better in reverberation, while EM+ILD works better in anechoic conditions. Initialization of the algorithm is very important to the success of all of the variants. For this reason, we initialized the source locations in the EM systems from the localization of PHAT-histogram, and begin with the E step.

Figure 2 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. Around 3.3 and 6.6 kHz, both estimates tend toward the overall mean, a remnant of the ambiguity of IPD at those frequencies that has generally been resolved by ILD at neighboring frequencies. Notice also that when just a single ILD is used for all frequencies, it settles to the mean of the ground truth ILD for each source.

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. The system takes advantage of interaural phase difference cues at low frequencies and interaural level difference cues at high frequencies. Distinct sources have the same IPD at certain frequencies, depending on their azimuthal separation, and for large separations this ambiguity occurs at frequencies lower than those where ILD is useful. We predict that humans might not be able to distinguish narrowband sources at such ambiguous, low frequencies either.

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 timedelay 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 timefrequency 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] J. Blauert, *Spatial Hearing: Psychophysics of Human Sound Localization*. MIT Press, 1997.

[8] 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.

[9] 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.

[10] M. I. Mandel and D. P. W. Ellis, “A probability model for interaural phase difference,” *Workshop on Statistical and Perceptual Audio Processing*, 2006.