Building a Binaural Source Separator

Michael I. Mandel, Daniel P. W. Ellis
Department of Electrical Engineering
Columbia University
New York, NY 10027
{mim,dpwe}@ee.columbia.edu

Tony Jebara
Department of Computer Science
Columbia University
New York, NY 10027
jebara@cs.columbia.edu

Abstract

We propose a number of cues and a strategy for combining them that could be used by a binaural machine to perform source separation. Our previous work has used the single cue of interaural phase difference (IPD) to segment the time-frequency plane using an EM algorithm. We see this as a first step towards a larger and more complete system that takes advantage of more of the cues available to a listener from the stereo mixture such as interaural level difference (ILD), monaural cues, and reliability cues. Additionally, these cues could be integrated with one another by extending the existing probabilistic framework.

1 Introduction

The human auditory system can discern many simultaneous sound sources in a single auditory scene using the input from only two ears. Whereas much effort of late has been focused on source separation from a single microphone and from arrays of many microphones, few investigations have attempted to wring as much information out of a stereo recording as might be possible. While much as been written about the various cues mentioned in this paper that can be used for source separation, they have yet to be integrated in a single system.

We are interested in stereo recordings for a number of reasons. While stereo recordings are less common than monophonic recordings, they are much more common than recordings from large microphone arrays and the hardware for producing new stereo recordings is ubiquitous. Microphone arrays, in addition to requiring more complicated hardware, also present calibration problems that grow exponentially in the number of microphones. Adding a second microphone may provide the greatest amount extra information for the least amount of extra calibration and other overhead. And finally, we are interested in stereo recordings because we would like to gain some insight into the methods that humans might use for separating sources.

2 Cues

There are many cues useful for source separation that can be extracted from binaural recordings. In addition to strictly binaural cues, a binaural strategy can take advantage of monaural source separation cues operating on both channels independently.

2.1 Binaural cues

Standard binaural cues for localization compare the signals received at the two ears to one another. These include the difference in arrival time of a signal at the two ears, known as interaural time difference (ITD), and the difference in sound energy between the two signals, known as interaural level difference (ILD). While the ILD can be computed at every point in a spectrogram, the ITD can only be computed unambiguously on wideband signals. A related cue, however, interaural phase difference (IPD), can be computed at every point in a spectrogram and used to calculate a probability distribution over ITDs at each point [1]. The aforementioned paper describes a system in which only the IPD is used to create a probabilistic mask using an EM algorithm. This mask could be used as cue for another algorithm or these other cues could be incorporated into the probabilistic framework and used directly in creating the mask.

2.2 Monaural cues

Many systems have been developed recently either to separate a target source like speech from background noise [2] or to separate two sources from one another [3, 4]. Systems that work in the spectral domain (as opposed to the cepstral domain) typically create masks which indicate which regions of the spectrogram might be associated with which sources. Such a mask for each of the channels in a binaural recording would serve as a useful cue for a source separator.

2.3 Reliability cues

One final cue is that of reliability. Wilson and Darrell [5] learn filters that predict the reliability of localization cues at particular points in spectrograms using the energy at neighboring points. The optimal such filters act much like the precedence effect in humans, finding that onsets provide the most reliable localization cues. These filters could be used as they are, or similar filters could be built using the same technique to predict local source characteristics which indicate the reliability of source separation. While the training occurs on binaural signals, the resulting filters act on monaural spectrograms and could be used to predict the reliability of each monaural signal.

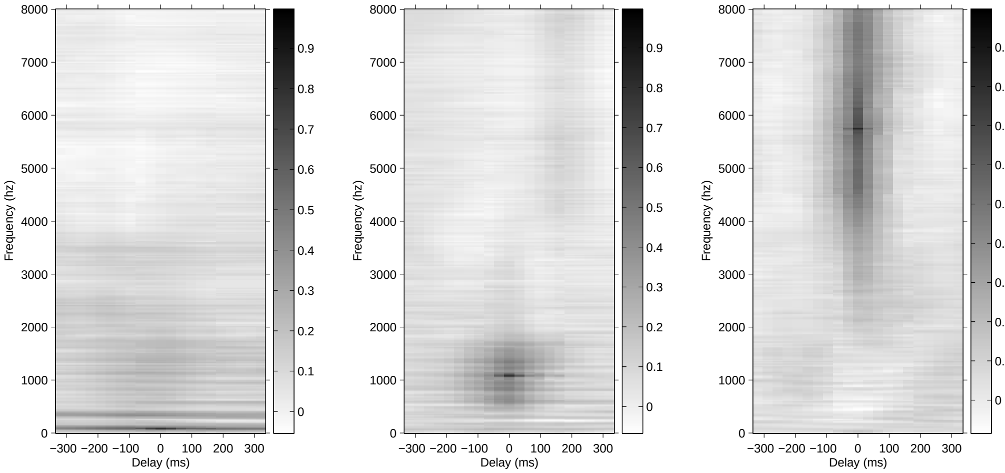
3 Combining cues

Once these cues are extracted, there are a number of ways in which they could be combined into an estimated partitioning of the time-frequency plane. The EM algorithm described in [1] only uses a single cue and assumes that all spectrogram points are statistically independent from one another. A fuller version of such a model, however, could take advantage of the extra information that is ignored by the current model.

Adding another cue, ILD, to the model should not be difficult. The level difference between the two channels is typically log-normally distributed, meaning that when measured in dB it is normally distributed. Although the means of the normal distributions varies with frequency and source location, they can be learned from training data. And since the magnitude and phase differences can be safely assumed to be independent from one another [6], a magnitude term can easily be added into the existing probabilistic model.

The assumption of independence between spectrogram points can be relaxed to account for correlations. As can be seen in Figure 1, there is a not insignificant correlation between the source active at one point in the spectrogram and the source active at its neighbors. This correlation takes different forms in different frequency regions. At low frequencies, the same source tends to maintain dominance in a frequency band for a number of time frames. At high frequencies, sources tend to dominate many frequency bands simultaneously, but only for a single time frame. These correlations can be used to build a Markov random field (MRF) or other grid-like graphical model to allow for variational solutions or loopy belief propagation.

Depending on when the cues are integrated with each other and integrated over the spectrogram, solutions can be more or less involved. In the simplest case, the monaural and binaural cues could be used separately and under the assumption of that all time-frequency points are independent of one another to estimate time-frequency masks of source location. Only after these masks are calculated, a Markov random field could integrate them with one another and across neighboring time-frequency points. A more complicated integration method would be to estimate a single time-frequency mask from all of the cues simultaneously while also taking advantage of correlations between neighboring time-frequency points. Such a method would require a variation solution to the problem instead of the simpler EM solution, but should be feasible.



(a) (b) (c)

Figure 1: From a ground truth mask for two sources in a reverberant room, the correlation of a point in the ground truth binary mask and its neighbors for points at three different frequencies: (a) low frequency, (b) mid frequency, and (c) high frequency

4 Proposal

In our workshop presentation we will further describe our planned framework. We will also present preliminary results on combining IPD and ILD cues in real-world reverberant environments.

References

[1] Michael I. Mandel, Daniel P. W. Ellis, and Tony Jebara. An EM algorithm for localizing multiple sound sources in reverberant environments. In *Proc. Neural Information Processing Systems*, 2006.

[2] Daniel P. W. Ellis and Ron Weiss. Model-based monaural source separation using a vector-quantized phase-vocoder representation. In *ICASSP-06*, pages V–957–960, May 2006.

[3] Barak A. Pearlmutter and Anthony M. Zador. Monaural source separation using spectral cues. In *Proc. Fifth International Conference on Independent Component Analysis ICA-2004*, 2004.

[4] Trausti Kristjansson, John Hershey, Peder Olsen, Steven Rennie, and Ramesh Gopinath. Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system. In *Proc. Int. Conf. Spoken Language Processing*, 2006.

[5] Kevin Wilson and Trevor Darrell. Learning a precedence effect-like weighting function for the generalized cross-correlation framework. *IEEE Transactions on Speech and Audio Processing*, 2006.

[6] Michael I. Mandel and Daniel P. W. Ellis. A probability model for interaural phase difference. In *Workshop on Statistical and Perceptual Audio Processing (SAPA)*, 2006.