Chapter 7

Conclusion

This work has introduced MESSL, a novel, probabilistic system for separating the speech of multiple simultaneous speakers from a stereo mixture. It is built upon a flexible, top-down multi-source localization system that generalizes a number of systems described in the literature. The separation system alternates between computing a probabilistic time-frequency mask for each source based on its model and re-estimating the parameters for each source from observations weighted by these masks.

Of the four goals of the work, it is able to locate the sound sources in space from a realistic binaural recording. It is able to separate the sources enough to determine what is said, in certain conditions. And it collections information about the early echoes and regions of the spectrogram dominated by reveberation that could be used to characterize the room in which the recordings were made.

The localization performance of the systems introduced here is comparable to that of humans in both anechoic and reverberant conditions. MESSL's separation performance in anechoic mixtures is comparable to humans' as well, but in reverberation it only makes up 20–25% of the difference between automatic speech recognition performance on raw mixtures and human performance. While MESSL successfully rejects enough of the direct-path portion of the masking source in reverberant mixtures to improve energy-based signal-to-noise ratio results, it has difficulty rejecting enough reverberation to improve automatic speech recognition results significantly. This problem is shared by other comparable separation systems.

An extensive set of experiments compared the performance of MESSL to a number of other state-of-the-art source separation algorithms. In the first set of experiments, these systems were evaluated on mixtures of two and three equally loud sources in anechoic and reverberant mixtures as the position of the masking sources was varied. MESSL was shown to more accurately localize all of the sources, and more effectively separate the target source than the other algorithms under evaluation. These experiments were also used to compare different versions of MESSL and it was shown that even two and a half seconds of audio provides enough data to fit its most complex parameterization. They also show that taking explicit measures to counteract the effects of reverberation improves

performance. Such countermeasures include treating the reverberation as an additional source and pushing interaural level difference parameters towards their values in anechoic environments.

A second set of experiments kept two sources fixed at 90° apart and varied their relative energies. It used automatic speech recognition to measure the amount of speech information that can be extracted from a mixture by these separation algorithms and compared this quantity to results measuring the intelligibility of the mixtures to humans. It showed that in anechoic mixtures, these algorithms successfully match the performance of humans, but in reverberation, they are still quite far.

These experiments also provide an opportunity to compare a number of objective measures of speech separation and show that they have very different characteristics, especially when the masking speech is much more energetic than the target speech. Notably, at low target-to-masker ratios, PESQ is unreliable and a commonly used signal-to-noise ratio metric finds improved performance in useless separations. Our two proposed metrics, a modified signal-to-noise ratio improvement and measurements of the direct-path, early echoes, and reverberation of the target and masker (DERTM), reveal the shortcomings in these separations and provide insight into the poor automatic speech recognition results, which are caused by a failure of the algorithms to reject reverberation from the masking speaker.

7.1 Future work

There are a number of directions that could be explored in the future based on this work. From the results of the ASR experiments and DERTM analyses, it seems that a dereverberation technique could be very useful in improving ASR performance on these separations. Specifically, it should be possible to construct a similar and compatible probabilistic model that can distinguish diffuse energy from direct-path. This could be as simple as taking account of coherence explicitly in MESSL, although results from (Mandel and Ellis, 2009) imply that even a ground truth measure of reliability is not sufficient for such a task.

Another possibility for overcoming this limitation could be the combination of mask-based separations with direction-based processing. Even though MESSL estimates the interaural delay of each source, this information is not taken into account when reconstructing sources from a mixture. Something as simple as selecting, at each time-frequency point, the ear with the higher estimated signal-to-noise ratio might improve performance. This could also incorporate cancellation-type processing of Sawada et al. (2007) or Buchner et al. (2005), although it is not clear whether such processing can remove diffuse sources.

MESSL could also be extended in various ways to make it more robust or useful in other circumstances. Other changes to the MESSL algorithm itself might improve performance as well. MESSL estimates maximum likelihood parameters for its mixture model. It could instead perform a Bayesian integration over these parameters, including the number of sources. While a closed-form integral is unlikely to be found, samples can be drawn from the resulting posterior distribution using Markov chain Monte Carlo methods. Because

MESSL is very similar to a Gaussian mixture model, it should be easy to adapt a model like the infinite Gaussian mixture model (Rasmussen, 2000) to draw samples from MESSL's posterior.

Another possible addition to MESSL would be in its modeling of early echoes. Currently, there is little structure enforced in them aside from a single dominant interaural time difference. If, however, a large number of impulse responses were generated and the early echoes studied, a prior could be built over them that could add structure to their modeling. Because interaural parameters are much easier to estimate than full binaural impulse responses, it would be interesting to see how much information about the early echoes can be recovered from the interaural parameters. Such a model of early echoes might also be able to infer information about the geometry of a space from its early echo pattern.

MESSL separates sources from a single batch of audio. In order for a realistic system to be fielded, it must be able to function in an online and real-time setting, meaning that the system will need to respond to data as they come in. If the dynamics of the sources were explicitly modeled, then Sequential Monte Carlo methods could perform such tracking. Births and deaths of sources would also need to be modeled, and would estimate the number of sources.

One application of MESSL that has not been explored, but could prove fruitful is its use on musical signals. The main difficulty in adapting MESSL for musical signals is the wide variety of techniques used for mixing and adding reverberation to stereo recordings, most of which do not provide complete interaural cues. For example, simple panning of sources is sufficient to provide human listeners with the impression of a spatial sound field, but does not impart any interaural time differences to the signals. Thus our current initialization, based on cross-correlation, would need to be augmented. Initialization could come from monaural source separators or from estimates of reliable ILD values for each source. With this adaptation, MESSL should be flexible enough to handle such mixtures.

(This page intentionally left blank)