Chapter 1

Introduction

1.1 Motivation

Humans with normal hearing are very good at listening to speech, even in the presence of other speakers, background noise, and reverberation. The canonical example of this ability occurs at a cocktail party, where two guests can carry on a conversation in the din of other such nearby conversations. Listeners can also determine the direction of a sound source by listening, for example turning to face someone who has called out to them.

Machines are less good at these tasks. Current automatic speech recognition systems work quite well for speech recorded in quiet with a microphone close to the speaker's mouth. They work much less well when there is noise present, especially the noise of other people speaking. Performance is also diminished by reverberation, as would be more prominent if the microphone were sitting on a desk instead of in a headset. Humans with hearing loss have difficulty in these situations as well. While current hearing aids improve a wearer's ability to understand a conversational partner in quiet and over the telephone, they do not provide this improvement in noisy or reverberant environments.

The vision of this work is to create a computer system that can "hear" as well as humans in spatially realistic conditions. Such a system should be able to analyze a recording made with two microphones of multiple people speaking in a room and

- infer the speakers' locations in the room
- separate the sounds of their voices "enough"
- transcribe what is said
- characterize the room.

While complete solutions to these goals are still many years away, this work reports progress towards each of them.

Many authors in the field of source separation have focused on the goal of perfect reconstruction of the clean, dry, "original" waveform of a source of interest from a mixture.

We believe that this goal is unnecessarily difficult, and we endeavor instead to separate sources of interest *sufficiently* for successful speech recognition. Existing machine solutions leave much room for improvement, although the fact that such a goal is attainable, at least within certain bounds, is proven by the human auditory system (Bronkhorst, 2000).

1.1.1 Separation and localization are useful

Machines that could separation and localize sounds as well as humans would enable many applications. As mentioned above, automatic speech recognition and hearing aids would be greatly improved in their current roles. Additionally, automatic speech recognition could be fielded in new situations, particularly in "passive" systems that listened for long periods of time. Hearing aids would be able to help their users in noisy social situations, and even improve the ability of normal-hearing individuals to focus on a speaker of interest. Furthermore, musical signals, carefully constructed stereo sound mixtures, might be more easily indexed and organized after being separated into sonic objects. Multimodal interfaces for these systems could provide some indication of the types of sounds present in a space along with information about the space itself, e.g. for the profoundly deaf.

Source separation could benefit automatic speech recognition (ASR) in traditional "active" user interfaces, but could also enable "passive" user interfaces. Active ASR interfaces, in which the user engages in a "conversation" with a computer system, are now available on GPS units, phones, and other mobile devices, and in directory assistance and call centers. Improved robustness to noise would enable these systems to successfully recognize speech from noisy environments. Source separation might be even more useful, however, in facilitating passive ASR interfaces, where the system listens for long periods of time, taking notes, waiting for commands, etc. Such systems might transcribe meetings (Janin et al., 2003), analyze a "life log" recording (Ellis and Lee, 2006), or control various systems in the home (Ducatel et al., 2001). Because these systems would always be on, they would be much more likely to encounter simultaneous speech. If they were installed in rooms, they would be much more likely to encounter reverberation as well.

As will be discussed in section 2.2.3.3, even listeners with normal hearing can become overwhelmed by speech information from non-target sources in a mixture, especially when these sources are too similar to one another. If these masking sources are removed, however, the target source becomes much more intelligible (Brungart et al., 2006). This problem is even more severe for listeners with hearing impairment, who find it much more difficult to understand a target source in reverberant and noisy mixtures (Nábělek and Mason, 1981). One of the most common complaints of hearing aid users is that their hearing aids do not improve the intelligibility of conversations in noisy environments (Kochkin, 2000). Hearing aids that utilize source separation could address these issues by presenting target sources to their wearers while suppressing interfering sources.

While this work does not directly address the problem of musical source separation, our separation system could be applied to musical signals with a few modifications. Musical source separation could be useful in automatically describing and summarizing songs, albums, artists' oeuvres, record labels, genres, etc. If these signals were broken down first into distinct sonic objects, such classifications could be made more easily, without the distraction of competing sounds. Stereo recordings are standard in music, facilitating localization-based separation, although they are not necessarily mixed in a way that presents realistic spatial cues. Humans perform such tasks effortlessly, however, and machines that could match those talents could help humans multiply their productivity in dealing with commercial-scale music collections.

1.1.2 Separation and localization are difficult

While researchers have been able to solve many limited problems in source separation and localization, these solutions rely on simplifying assumptions that are frequently unrealistic. Furthermore, these assumptions are not required by the human auditory system to solve these same problems, leaving room for improvement in difficult conditions. As will be described in more detail in sections 2.4 and 2.5, automatic separation and localization systems typically work well for statistically stationary sources in anechoic conditions with more microphones than sources. They typically work less well on mixtures with more sources, fewer microphones, more complex source statistics, and more reverberation. Realistic recordings, as heard by humans in everyday listening conditions, contain most or all of these complications.

Source separation algorithms can be divided into *overdetermined* and *underdetermined* algorithms. Overdetermined algorithms can separate at most as many sources as they have observations, i.e. microphones. This includes the condition in which the number of sources and observations is equal, known as the *critically determined* condition. The underdetermined case is more difficult than the overdetermined because the system must bring additional prior information to bear on the problem in order to separate the sources. In general, source separation and localization become more difficult as more sources are introduced and as microphones are removed. The human auditory system is able to solve the underdetermined problem: humans have only two ears, but can separate and localize at least three simultaneous sources (Hawley et al., 1999). Additionally, while humans are able to ignore diffuse sources, overdetermined algorithms can only cancel compact sources.

Automatic systems are also more easily analyzed when sources have consistent, idealized behavior including whiteness, Gaussianity, and stationarity. Very few natural signals exhibit these properties; and speech, most notably, exhibits none of them. The lack of these simplifying assumptions makes speech more difficult to separate than noise. Similarly, separation is more difficult when the noise in which speech is embedded does not exhibit these properties. An extension of this limitation is that signals become more difficult to localize and separate as their duration decreases, as there is less of an opportunity to collect statistics about the signals involved.

Human speech separation and localization performance suffers its own limitations due to several factors. These factors also limit automatic solutions and it is not clear whether they can be overcome. First, reverberation makes separation and localization harder, especially as the amount of reverberant energy increases relative to that of non-reverberant energy. Such an increase can be due to either lengthier reverberation, as would be found in a larger or less absorptive room, or louder reverberation, as would be found in a less absorptive room or as a source moves farther away.

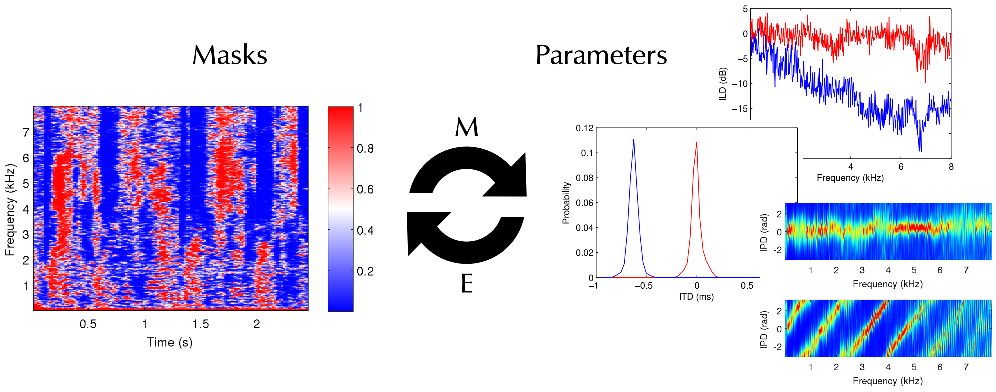

Figure 1.1: Overview of the Model-based EM Source Separation and Localization system (MESSL). In the E-step, models of sources' interaural parameters are used to estimate probabilistic time-frequency masks. In the M-step, these masks are used to re-estimate the parameters of each model from relevant regions of the spectrogram.

One roadblock to human separation performance that might not afflict machines is a limitation on attention. Humans are able to attend to only a small number of sources at a time. If a successful machine source separator were built, increasing the number of sources that it could attend to could be as simple as adding computational resources. Machines also might be able to separate sources based on specific criteria more easily than humans. Humans are adept at exploiting any difference between two speakers' voices when separating them from one another. If such a difference contradicts other information or is distracting, however, it cannot be ignored completely, but its effect can be mitigated (Ihlefeld and Shinn-Cunningham, 2008). The single-minded focus of the computer can enable separation in conditions that humans find confusing or unnatural.

This work focuses on systems that are capable of separating and localizing sources in reverberant, underdetermined mixtures. Frequently, anechoic and critically-determined mixtures are also used as points of comparison, or to facilitate analysis where underdetermined performance is not qualitatively different. It also explores situations in which the target and masker energies are very different, although our system was not designed specifically with these situations in mind.

1.2 Overview of MESSL

The system proposed in this work separates sources by modeling the location of each one probabilistically, with parameters learned using an expectation maximization (EM) algorithm. It is therefore called Model-based EM Source Separation and Localization (MESSL). Each source in a mixture is described by a probabilistic model of the interaural time and level differences, features that are closely tied to location. One of the key ingredients enabling this is a probabilistic localization model that can be evaluated independently at every time-frequency point. Chapter 4 introduces such a model. This model is similar to the computation of the cross-correlation in that it compares a number of interaural delay hypotheses, but unlike the cross-correlation, it is formulated in such a way that it can be applied to arbitrarily-shaped regions of a spectrogram in order to localize multiple simultaneous sources. The separation system described in chapter 5 infers the appropriate regions of the spectrogram in which to apply each localization model and in this way localizes each source using as much relevant data as possible, while avoiding data corrupted by other sources.

As illustrated in figure 1.1, this EM algorithm alternates between refining its estimates of the model parameters (location) for each source and refining its estimates of the regions of the spectrogram dominated by each source. In the expectation step of the algorithm, spectrogram points are assigned to models probabilistically in proportion to the agreement between the observation at that point and each model. In the maximization step, parameters for each source model are re-estimated from points in proportion to their affinity with that model. Basic models include the interaural time difference (ITD) via the interaural phase difference (IPD), and the interaural level difference (ILD), but the flexibility of the framework and its probabilistic nature allow the addition of other information such as models of the sources themselves. This flexibility also allows the system to separate mixtures made with both human-like dummy head recordings and free-field microphones, although the experiments in this work focus on dummy head recordings.

In addition to successfully separating sources, these models are transparent in their representation of the sources, i.e. the model parameters are interesting in and of themselves. It is straightforward to estimate interaural level and time differences for a single source in an anechoic recording. MESSL, however, can estimate these same parameters from a reverberant recording in the presence of other sources, a much more difficult task. This topic is discussed in more detail in section 5.3.

MESSL has a number of advantages over similar systems that will be discussed in sections 2.4 and 2.5. First, in contrast to independent component analysis and other traditional blind source separation techniques, MESSL is able to separate underdetermined mixtures. Unlike generalized cross-correlation or source-modeling techniques, it does not depend on any specifics of the sources involved in the mixture, although explicit models of them can be added if desired. While a number of algorithms generate binary separation masks, MESSL generates real-valued masks with true probabilities that can be easily combined with other probabilistic masks. In contrast to many source separation systems, MESSL has been shown to work in reverberant environments. And unlike beamforming techniques, it requires little prior knowledge of the microphones' geometrical arrangement, the room, or the head transfer function, although again such information can be added if desired.

In order to avoid these limitations, MESSL must make a few assumptions. The first is that sources do not move in space, a result of the batch processing of sounds and the lack of a model for sources' spatial dynamics. It also assumes that one source dominates each time-frequency point, a property known as windowed-disjoint orthogonality that will be explored in section 3.1.4. Finally, it assumes that the number of sources and the possible interaural time delays of the sources are known *a priori*.

Although this work focuses mainly on the first two goals described in section 1.1, it also address the final two goals, determining what is said and characterizing rooms. Experiments carried out in chapter 6 use automatic speech recognition to measure the amount of speech information isolated by various separation algorithms. This experiment was based on the study of intelligibility with human subjects described by Bressler and Shinn-Cunningham (2008), and the results of the two studies are compared. In terms of characterizing rooms, MESSL does capture information that could be relevant to such a task, including the interaural spectrum of early echoes and an estimate of the time-frequency points that are dominated by reverberation. Using this information to infer properties of rooms is left as future work, however.

1.3 Contributions

The primary contribution of this work is the Model-based EM Source Separation and Localization algorithm, a probabilistic separation and localization system. Under this umbrella are a number of additional contributions:

- a tractable probabilistic mixture model of interaural phase differences based on time differences
- a model of interaural level difference (ILD) and a prior on ILD that enables the model to estimate ILD parameters even in reverberation
- an explicit model of reverberation of the same form as other source models

The work also contributes two source separation evaluation metrics:

- a particular formulation of signal-to-noise ratio that penalizes both undesirable energy that is included in a separation and desirable energy that is excluded from the separation.
- an analysis of the effect of mask-based separations on the direct-path, early echo, and late reverberation portions of the target and masker signals.

Other contributions include:

- a bound on the performance of separation algorithms like MESSL, called the ideal interaural parameter mask
- a clarification of the definitions of desirable and undesirable signals used in constructing ground truth masks and evaluating source separation algorithms.

This work was described in five papers, but also includes some previously unpublished material. The localization algorithm was introduced in (Mandel and Ellis, 2006), but has been greatly expanded in chapter 4. The first version of MESSL, which separated sources based only on interaural phase difference cues, was introduced in (Mandel et al., 2007).

This was followed by (Mandel and Ellis, 2007), which added interaural level difference cues, and (Mandel and Ellis, 2009), which added robustness to reverberation in the form of a prior on ILD parameters and an explicit model of reverberation. That paper also introduced the ideal interaural parameter mask, which will be discussed in section 3.2.2. These papers were summarized in (Mandel et al., 2009), which also included a number of additional experiments investigating the behavior of various aspects of the system. This paper serves as the basis for chapter 5. Finally, (Weiss et al., 2008) introduced a model of speech sources into MESSL, which is described in section 5.2.3. The experiments described in chapter 6 have not been published before.

1.4 Outline

This section provides an overview of the structure of the rest of this document. Chapter 2 introduces the binaural room impulse response in detail. It then describes human performance in separating sound sources, particularly in reverberation and in mixtures. This is followed by a discussion of human performance in localizing sources, particularly in reverberation and in mixtures. Machine performance in these two areas is then discussed.

Chapter 3 illustrates, using an extended example, the interaural spectrogram and the observations that are used by all of the algorithms tested in subsequent chapters. These observations motivate the assumptions and models used in subsequent chapters. It also describes various oracle masks that provide upper bounds on mask-based separation performance, including the novel ideal interaural parameter mask.

Chapter 4 describes a probabilistic, top-down localization algorithm. It discusses this algorithm in the context of both generalized cross-correlation (GCC) localization algorithms and localization algorithms based on higher order maximum entropy circular probability distributions. It then presents two localization experiments, which show that the probabilistic algorithm can localize single sources in reverberation more accurately than GCC-based methods and requires shorter observations to do so. Its performance in reverberation is comparable to that of humans in anechoic conditions. Also, a novel region-based localization and separation algorithm based on a smoothed normalized cross-correlation is presented, which is shown to be able to localize sources in reverberant mixtures.

Chapter 5 describes the full MESSL system and evaluates it with a number of experiments. These experiments are based on separating and localizing two and three sources in anechoic and reverberant mixtures, while the angle between the sources is varied. They show that of the five algorithms tested, MESSL separates sources with signal-to-distortion ratios 2.0 dB higher than comparable algorithms, and localizes sources with 40% lower mean absolute error. Experiments on various aspects of MESSL show that its most complex parametrizations perform best without over-fitting the data, that it is more difficult to estimate parameters from mixtures than to estimate masks, but that even from a mixture, MESSL can estimate interaural parameters that are close to those measured in isolation.

Chapter 6 examines the performance of these algorithms along with a number of ground-truth mask-based separations on another experiment, for which human performance data are available. This experiment varies the target-to-masker ratio of mixtures, with the sources at fixed locations, using automatic speech recognition to compare algorithmic performance to human performance. It shows that in anechoic conditions, many of the algorithms perform similarly to humans, but that in reverberation, they perform much worse. Because of a failure of the localization algorithm that initializes it, MESSL cannot separate sources in reverberation when one source is more than 15 dB louder than the other. A number of metrics are also compared for evaluating source separation performance and the limitations of these metrics are discussed.

Finally, chapter 7 summarizes the findings and conclusions of this work and discusses directions for future research.