Underdetermined Source Separation Using Speaker Subspace Models

Ron J. Weiss

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY
2009

(This page intentionally left blank)

(This page intentionally left blank)

Abstract

Underdetermined Source Separation Using Speaker Subspace Models

Ron J. Weiss

Sounds rarely occur in isolation. Despite this, significant effort has been dedicated to the design of computer audition systems, such as speech recognizers, that can only analyze isolated sound sources. In fact, there are a variety of applications in both human and computer audition for which it is desirable to understand more complex auditory scenes. In order to extend such systems to operate on mixtures of many sources, the ability to recover the source signals from the mixture is required. This process is known as source separation.

In this thesis we focus on the problem of *underdetermined* source separation where the number of sources is greater than the number of channels in the observed mixture. In the worst case, when the observations are derived from a single microphone, it is often necessary for a separation algorithm to utilize prior information about the sources present in the mixture to constrain possible source reconstructions. A common approach for separating such signals is based on the use of source-specific statistical models. In most cases this approach requires that significant training data be available to train models for the sources known in advance to be present in the mixed signal. We propose a speaker subspace model for source adaptation that alleviates this requirement.

We report a series of experiments on monaural mixtures of speech signals and demonstrate that the use of the proposed speaker subspace model can separate sources far better than the use of unadapted, source-independent models. The proposed method also outperforms other state of the art approaches when training data is not available for the exact speakers present in the mixed signal.

Finally, we describe an system for binaural speech separation that combines constraints based on interaural localization cues with constraints derived from source models. Although a simpler system based only on localization cues is sometimes able to adequately isolate sources, the incorporation of a source-independent model is shown to significantly improve performance. Further improvements are obtained by using the proposed speaker subspace model to adapt to match the sources present in the signal.

(This page intentionally left blank)

Contents

List of Figures

List of Tables

(This page intentionally left blank)

Acknowledgments

I owe a great deal of gratitude to the people who have helped me on my journey to complete this dissertation. I would like to thank my adviser, Dan Ellis, for his guidance throughout my graduate school career. Dan was a terrific adviser to work with. He allowed me a great deal of freedom in pursuing research topics and was always available to discuss my ideas, no matter how wacky, and set me down the right path. These discussions have proved invaluable to the development of this research.

I would also like to thank Shih-Fu Chang, Rui Castro, Trausti Kristjansson, and Juan Bello for serving on my thesis committee and providing valuable and insightful comments throughout the final stages of preparing this dissertation. I am further indebted to Trausti for serving as my mentor over the course of my visits to Google. I learned a great deal from him and from the rest of the speech recognition group during my tenure there.

Despite being contained in a windowless room with an overactive HVAC system, LabROSA was a great setting for graduate school. This is in no small part due to the great people who occupied the lab. The long discussions about our respective projects, our favorite text editors, and the pros and cons of Matlab as a programming environment were always enjoyable. On the research side of things, I would especially like to thank Michael Mandel for his collaboration on the binaural separation work described in chapter 5.

I am grateful to my parents, Lazslo and Agnes, and my sisters, Michelle, Becky, and Liana, for all of their encouragement and support over the years. Dad, I may not be in medicine, but you can still call me doctor.

And last of all, to Gila: writing this thesis, and surviving graduate school in general, would have been significantly more difficult were it not for your support. Whatever sanity I have left is a testament to your patience. Thank you.

(This page intentionally left blank)

Chapter 1

Introduction

Recognition of audio signals containing contributions from multiple sources continues to pose a significant problem for computer audition. In contrast, human listeners often have little trouble paying attention to a single source in the presence of conflicting sources. Such mixtures are extremely commonplace, and therefore represent an important challenge. For example, this class of scenarios is often referred to as the "cocktail party problem" [13], emphasizing the everyday nature of this challenge. Much research effort has been devoted to building computational systems that can solve this problem.

Audio source separation systems have a variety of potential applications in the context of computer audition, including automatic speech recognition under especially noisy conditions, and multimedia or music analysis where signals are virtually always mixtures of multiple sources. Such systems generally output a high level description of the audio input and do not necessarily require that the underlying source signals be separated and reconstructed with high fidelity. However, it is often possible to configure such a system to operate on single-source inputs and simply offer source separation as pre-processing whenever necessary. This is the approach taken in many of the experiments we will describe in this thesis.

Other important application areas are targeted at human listeners and therefore require that the source signals be reconstructed accurately. Although human listeners are often able to attend to individual sources even in dense mixtures, this ability is significantly reduced in people with impaired hearing. In fact, one of the primary complaints of hearing aid users relates to their poor performance in cocktail party-like situations [51]. Incorporating source separation algorithms into hearing aids could vastly improve their usefulness in situations with significant background noise. Other similar application areas include speech enhancement for telecommunications or other artificial listening situations where many of the perceptual cues humans use to solve the cocktail party problem, e.g. binaural localization cues, are not available, or music remixing when the original multi-track recordings are unavailable.

Separation is especially difficult when the mixed signal is particularly constrained, such as in underdetermined conditions when there are fewer observed channels than there are distinct sources comprising the mixture. In this thesis we focus on the use of statistical source models for separation of underdetermined mixtures in which observations consist of only one or two channels. We primarily focus on speech separation, i.e. the cocktail party problem, where the mixture is composed of speech from multiple talkers. However, the methods we describe could be extended to work on other types of audio signals as well.

Separation algorithms based on such source models have been quite successful under underdetermined conditions, especially in monaural mixtures. However, they suffer from an important disadvantage in that the models must be specific to the sources present in the mixture. This requires that the identities of all sources be known in advance and that sufficient data be available to train models for each of them. In this thesis we focus on the model-based approach to source separation when the precise source characteristics are not known a priori.

1.1 Contributions

The primary contribution of this thesis is the development of a speaker-adaptive speech model for use in speech separation. The use of this model enables a popular approach to model-based source separation (e.g. [91]) to work on mixtures composed from utterances from previously unseen speakers. This removes the unrealistic requirement that training data for all possible speakers be available in advance and therefore makes such separation systems significantly more useful in practice.

We demonstrate the proposed model's effectiveness on monaural mixtures in the context of the 2006 Speech Separation Challenge [16], an organized effort to evaluate monaural speech separation systems for automatic speech recognition. We derive and compare the performance of two different separation algorithms based on this source model, one based on a hierarchical separation and adaptation approach, and a second variational learning algorithm which is significantly faster than the first but not as accurate. Separation based on the proposed speaker-adaptive model significantly outperforms the results obtained using speaker-independent models. We further show that the proposed system outperforms other state of the art approaches when training data is not available for the exact speakers present in the mixed signal.

We also demonstrate the effectiveness of the speaker adaptation model on the easier task of separating binaural speech mixtures. Unlike monaural mixtures, binaural mixtures can often be separated by assuming that the signal from each source arrives from a spatially distinct location without using any prior knowledge of the source statistics. In this thesis we extend the model-based expectation maximization source separation and localization (MESSL) algorithm of Mandel and Ellis [67] to incorporate additional constraints from prior source models. We demonstrate that even the use of relatively simple source models can improve separation performance over the baseline algorithm which utilizes source location alone. Because of the localization information present in binaural mixtures such an improvement is possible even when the models are relatively broad compared to those used for monaural separation, consisting of small speaker-independent models. The additional constraints from the source model essentially serve to guide the estimation of the model's localization parameters. Furthermore, we demonstrate that the addition of source adaptation based on the proposed speaker subspace model improves performance over a similar system that uses speaker-independent models in some conditions, although the improvement is not as large as it is in the monaural case.

Portions of the work in this thesis have been previously published in [29, 116, 117, 118, 120, 119, 68].

1.2 Thesis overview

The remainder of the thesis is structured as follows:

In chapter 2 we give an overview of the commonly used approaches to source separation and introduce the idea of utilizing source-specific constraints that can guide separation under very difficult conditions.

In chapter 3 we describe a speaker subspace speech model appropriate for speech separation. The subspace model can accurately capture the high frequency resolution, speaker-dependent characteristics of speech signals across a broad set of speakers in a parametric manner. This enables a source-adaptive approach to source separation where the subspace model can be used to effectively separate mixtures comprised of utterances from talkers different from those used to train it.

In chapter 4 we describe how the speaker subspace model described in chapter 3 can be applied to monaural speech separation. This requires estimation of the subspace model parameters for each of the sources present in the mixture. We describe two algorithms for learning the adaptation parameters directly from the mixed signal and compare them to an algorithm for model selection from a predefined set of speaker parameters. We then review some experimental results comparing our adaptation based separation system to other state of the art techniques in the context of the 2006 Speech Separation Challenge. We also compare the performance of the proposed model adaptation algorithm and the model selection algorithm on mixtures of utterances from speakers not present in the training set.

In chapter 5 we describe a system for separating multiple sources from an underdetermined, reverberant, two-channel recording using the proposed speaker subspace model. We extend the MESSL model to incorporate the speaker subspace model of source statistics described in chapter 3 and show that this can improve performance over the original algorithm.

Finally, in chapter 6 we present concluding thoughts and sketch out directions for future work.

Chapter 2

Overview of Audio Source Separation

In this chapter we review the state of the art in the field of audio source separation. Many closely related variants of the source separation problem have been studied in the literature, with the earliest approaches dating back over thirty years [75]. We begin by introducing the topic of blind source separation where minimal assumptions are made about the sources present in a mixture. We then describe more challenging separation problems where perfect separation is impossible and strong models of the underlying source signals are required.

2.1 Blind source separation

Many approaches to source separation have been proposed to operate under different situations. There are a number of scenarios under which the problem comes up, each of which requires a different approach. In general, we can express the problem of source separation as the problem of recovering the source signals $\{\mathrm{x}_i(t)\}_{1\leq i\leq I}$ from a set of observations $\{\mathrm{y}_j(t)\}_{1\leq j\leq J}$. In the simplest case, the source signals are assumed to arrive at the microphones at the same time without being filtered, i.e. each source contributes to each observed channel with some multiplicative gain. Such instantaneous mixtures can be written as a linear combination of source signals:

$$\begin{bmatrix} \mathrm{y}_1(t) \\ \mathrm{y}_2(t) \\ \vdots \\ \mathrm{y}_J(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1I} \\ a_{21} & a_{22} & \dots & a_{2I} \\ \vdots & \vdots & \ddots & \vdots \\ a_{J1} & a_{J2} & \dots & a_{JI} \end{bmatrix} \begin{bmatrix} \mathrm{x}_1(t) \\ \mathrm{x}_2(t) \\ \vdots \\ \mathrm{x}_I(t) \end{bmatrix} \tag{2.1}$$

Many approaches are targeted at the more difficult problem of separation of convolutive mixtures due to e.g. reverberation:

$$\mathrm{y}_j(t) = \sum_I \mathrm{h}_{ji}(t) * \mathrm{x}_i(t) \tag{2.2}$$

where $*$ denotes convolution.

The selection of an appropriate separation method varies with the nature of the mixing process. In the case of instantaneous mixing, separation is a matter of estimating and inverting the mixing matrix $A = [a_{ji}]$. The use of longer filters complicate matters in the case of convolutive mixing. One approach to dealing with this is to perform separation in the short-time Fourier transform (STFT) domain where each frequency band can be treated independently. Assuming that the filter impulse responses $\mathrm{h}_{ji}(t)$ are shorter than the STFT window, the convolutive separation problem reduces to instantaneous separation performed separately in each frequency band [102]. This leads to an additional complication related to matching the separated sources in each frequency band with the corresponding source in the other bands.

Note that equations (2.1) and (2.2) both reflect simplified mixing models where the mixing parameters, i.e. the mixing matrix A or filters $\mathrm{h}_{ji}(t)$, are assumed to remain constant across the entire signal, i.e. the sources and microphones are assumed to remain stationary. In fact, we use this simplifying assumption throughout this thesis. However, it is possible to extend many of the algorithms we describe in this chapter to work in an on-line fashion to separate moving sources by allowing the mixing parameters to vary with time, e.g. [52].

2.1.1 Overdetermined mixtures

Another important issue to consider is the relationship between the number of observed channels, J, and the number of underlying sources, I. In the overdetermined or critically determined cases, $J \geq I$ and it is possible in principle to exactly recover the original source signals. It is often possible to separate such overdetermined mixtures without making strong assumptions about the sources or mixing parameters. Such methods are generally referred to as blind source separation (BSS) methods. A common approach to this type of problem is based on independent component analysis (ICA), a higher order generalization of principal component analysis, in which the underlying source signals are constrained to be statistically independent. The sources can be separated by optimizing a measure of their mutual independence or "non-Gaussianity" using higher order statistics such as kurtosis [44], or mutual information [9]. A wide variety of ICA algorithms have been developed since the original work of Bell and Sejnowski [9]. Detailed reviews of ICA and other BSS techniques can be found in Douglas [23] and Choi et al. [14].

While many of the early approaches were designed to work on instantaneous mixtures, there has been significant work on applying to convolutive mixing as well [76]. As described above this is typically done using ICA variants based on frequency-domain analysis [102, 45]. The permutation ambiguity inherent in this approach can be solved by applying additional constraints designed to align sources along frequency. Various approaches have been proposed, based on enforcing that the time domain envelope of the source signals are consistent across all frequency bands [45], source localization using beamforming [94], or correlation across frequency bands [95].

2.1.2 Underdetermined mixtures

The separation problem becomes significantly more difficult if the number of underlying sources is larger than the number of observed channels, i.e. $J \leq I$. Separation of such underdetermined mixtures requires the separation algorithm to make stronger assumptions about the source signals than in the overdetermined case. Even then, because the problem is ill-posed, the sources can only be approximately recovered.

Basis decomposition

Many different methods for underdetermined separation have been proposed. One idea common to many of these methods is that of decomposing the source signals into the weighted sum of a set of basis functions, essentially projecting the signals onto the space spanned by these bases. Assuming that the source signals comprising a mixture can be represented using disjoint basis sets (i.e. they are generated from different subspaces), the mixture can be separated by projecting the mixed signal onto the union of source bases. Each source can then be reconstructed by utilizing only the bases known to be associated with that source.

In general, this process is difficult because the source subspace bases are not known in advance. Lee et al. [61] propose learning source-specific bases directly from underdetermined mixtures using ICA. They extend ICA to learn an overcomplete representation of the mixture containing more basis vectors than there are observed channels. This approach is able to uniquely identify the sources in the overcomplete ICA representation because it incorporates a sparse prior distribution to ensure that each source is represented primarily by a single basis vector. The method is able to separate a two channel mixture of three speech sources. When $J \geq I$, this method is equivalent to the time-domain ICA algorithm described in the previous section.

This is closely related to underdetermined BSS based on sparse coding. Such algorithms use a predefined basis set for all source signals under which the source signals are known to have a sparse distribution, e.g. STFT bases for speech which tends to be well localized in time and frequency [127]. Intuitively, sparsity is a valuable attribute for source separation because enforcing that the distribution over the bases for each source be sparse reduces the likelihood that the distributions overlap. This ensures that the sources are uniquely identified because any two sources are unlikely to be generated by the same bases, i.e. the sources are associated with roughly disjoint subsets of the overall basis set. This has a similar effect to constraining the source distributions to be statistically independent.

The assumption that source signals have a sparse distribution in the time-frequency domain has been shown to hold quite well for speech signals [91, 125]. Given the short-time Fourier transform magnitude of a mixture of two speech signals, it has been empirically observed that over 80% of the time-frequency cells lie within 3 dB of the larger of the corresponding cells of the two constituent clean signals [26]. This idea is demonstrated in figure 2.1. It

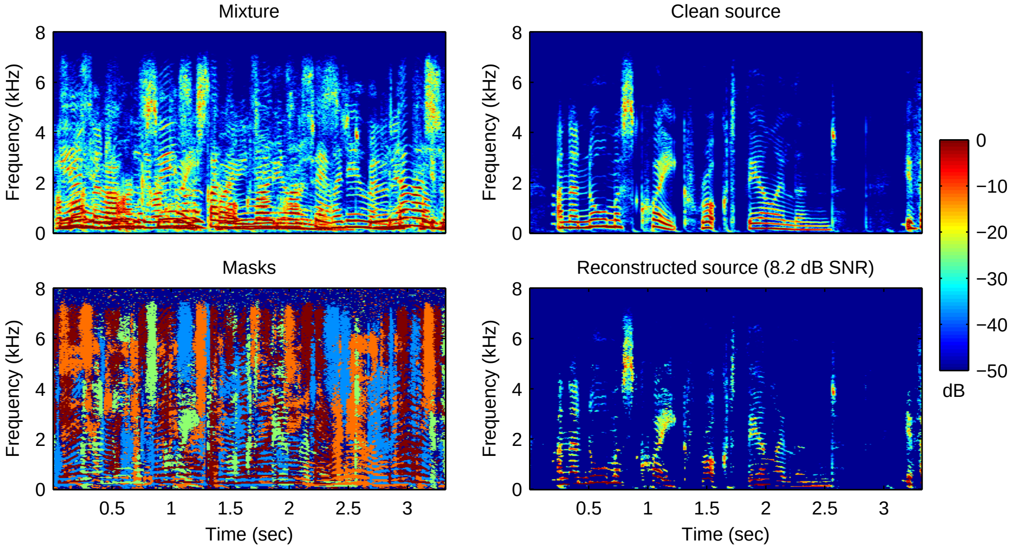


Figure 2.1: Demonstration of the sparse nature of speech signals. The top left panel shows the spectrogram of a monaural mixture of four speech sources. The bottom left panel shows which source dominates each time-frequency cell by segmenting the spectrogram into green, red, orange, and blue regions corresponding to each of the underlying sources. The dark blue regions contain negligible energy from all sources. The bottom right panel shows the reconstruction of the green source found by retaining the time-frequency cells dominated by that source. The corresponding clean source signal is shown in the top right panel.

shows a dense mixture of four speech sources and illustrates which time-frequency regions are dominated by each source. Even in such dense and underdetermined mixtures it is possible to adequately recover the underlying source signals. The reconstruction of the source based on the green mask retains 78% of the energy in the original source signal, despite the fact that 90% of the time-frequency cells are discarded. When listening to these signals the reconstruction remains quite intelligible and sounds quite close to the original.

Spectral masking and source localization

The assumption that source signals are sparsely distributed in the time-frequency domain has inspired many related separation methods based on the STFT decomposition. For example, Abrard et al. [2] describe an early method to leverage the sparse nature of speech signals for underdetermined, multichannel BSS. They show that the mixing coefficients from equation (2.1) can be exactly recovered from portions of the mixed signal that only contain energy from one source. Such regions are identified by searching for peaks in the distribution of the ratio of the STFT of the two observed channels, similar to the interaural level difference cues used in human auditory perception [69].

Yilmaz and Rickard [125] describe a similar method for underdetermined two channel separation by clustering the same localization cues used by the human auditory system to identify time frequency regions dominated by a single source. Unlike [2], the spectral masks derived from this process are used directly to separate a particular source by simply discarding the time-frequency cells of the mixed signal that are dominated by interfering sources, as in the bottom right pane of figure 2.1. It does not attempt to estimate and invert the mixing matrix. Because this spectral masking approach does not rely on knowledge of the mixing process, it can be used to separate convolutive mixtures as well. The separation algorithm of [125] and other similar approaches for underdetermined binaural separation is discussed in detail in chapter 5.

2.1.3 Monaural mixtures

Perhaps the most difficult source separation problem occurs when only a single channel observation is available. None of the underdetermined separation methods described in the previous section have been shown to work on such mixtures. However, it is possible to extend some of these approaches to work in such conditions. In this section we first review monaural extensions to the underdetermined basis decomposition methods described in the previous section. Although it is sometimes possible for these algorithms to operate in blind manner without making any strong assumptions about the underlying source signals other than their sparsity or mutual statistical independence, obtaining the best performance typically requires the separation algorithm to incorporate additional prior knowledge about the nature of the source signals. This is especially important on mixtures composed from statistically similar source signals, e.g. mixtures of speech signals.

Independent subspace analysis

Casey and Westner [12] describe Independent Subspace Analysis (ISA), an approach similar to the multidimensional ICA approach for separation of overcomplete convolutive mixtures. ISA overcomes the usual ICA requirement that there be at least as many sensors as sources by operating in the STFT domain. This simulates an observation with dimensionality increased from 1 to N. Performing ICA on the transformed signal yields a set of N independent bases which must then be grouped together into I subsets corresponding to the sources present in the mixture. Davies and James [19] show that this approach is only appropriate when the underlying sources have disjoint spectral support which guarantees that the ICA bases will be linearly independent. If the sources have overlapping support in frequency, as is generally the case for mixtures of speech signals, then the separation algorithm must utilize strong prior information to obtain high quality separations.

Non-negative matrix factorization

A similar method for unsupervised monaural separation that has shown promise is non-negative matrix factorization (NMF). In its simplest implementation, NMF decomposes a non-negative matrix representation of a mixed signal, i.e. the magnitude or power STFT, into the product of two low rank, non-negative matrices: $Y \approx BS$. The columns of B contain a set of basis vectors that define the spectral structure of the spectrogram, corresponding to spectral templates that appear repeatedly throughout Y. Similarly, the rows of S describe the temporal structure of the mixture, defining the activations and scaling of the bases in B. A variety of algorithms have been proposed for learning B and S from Y in an unsupervised manner [60], potentially incorporating sparsity constraints similar to the sparse coding method described above [97, 113] and temporal continuity constraints [113] to improve separation quality. The only parameter that needs to be specified is the rank of B and S, i.e. the number of basis vectors to use. Assuming that the underlying sources are composed from disjoint subsets of the basis vectors, it is straightforward to recover the source signals by partitioning B and S into parts that are unique to each source: $X_i \approx B_i S_i$.

In the simplest case, each source can be represented by a single basis vector. However, this makes it difficult to describe complex source signals such as speech that contain many different sounds with significantly varying spectral support. Virtanen [111], Smaragdis [100] describe an extension to the basic NMF approach to incorporate convolutive bases that extend over multiple frames. This enables the basis set to accurately represent more complex source signals. An alternative approach to modeling complex signals, similar to that taken by ISA and sparse coding, is to use an overcomplete basis set, i.e. set the rank of B to be larger than the number of sources and allow each source signal to be composed from multiple basis vectors. This requires the use of an additional stage prior to separation to group together the bases corresponding to each source. It is possible to use clustering approach similar to that used in ISA to do such grouping [12]. However, accurately performing this grouping is difficult to do in a completely unsupervised manner, especially if the source signals are very similar. As a result, many NMF-based speech separation systems operate in a supervised manner by learning the source specific bases B_i using clean training data specific to the type of source (e.g. specific talker, musical instrument, etc.) known to be in the mixture [97, 101].

2.2 Model-based source separation

All of the source separation methods approaches described in this chapter have been based in some form on the idea of utilizing constraints on the source signals comprising a mixture to disambiguate and separate them. In cases where there is redundancy in the observed mixture, as in overdetermined mixtures, signal-blind constraints based on mutual statistical independence or distinct spatial location can be used to effectively isolate individual sources. On their own, such constraints are much less effective when the problem is underdetermined. In the most challenging situation, where only a monaural observation of the mixture is available, high quality separation generally requires the use of constraints that are specific to particular sources. Such source-dependent constraints are often combined with the more general assumptions about signal independence or sparsity, as in the supervised NMF approach described in the previous section.

In this section we expand on the idea of model-based source separation, where constraints based on a predefined model of the underlying source signals is used for separation. This is in contrast to many of the other separation methods we have already described, where the relationship between source signals in a mixture is exploited, often without making any strong prior assumptions about the sources themselves.

The approaches described in this section can be broadly characterized as being based on explicit or implicit models of source characteristics [26]. Methods that fall into the former category can be described as supervised algorithms which rely on the use of training data to learn a prior model or distribution over the underlying sources signals. This approach is attractive in that the source constraints are learned directly from the data without making any prior assumptions. In contrast, the second category consists of unsupervised approaches that are instead based on more loosely defined constraints that are not specific to a particular source, e.g. based on consistent periodicity, or common frequency modulation of harmonic series. This has the advantage of not being dependent on the collection of training data, but comes at the cost of being more complex and difficult to implement. Hybrid approaches that combine both of these ideas have also been explored. A detailed review of model-based source separation can be found in Ellis [26].

Supervised ICA

Blind source separation under the most constrained conditions, i.e. when the observation consists of a single-channel, remains quite difficult. Relaxing the blindness requirement opens up wider possibilities for the design of separation algorithms. By incorporating information about the source signals learned from a set of training data it is possible to improve the performance of many of the unsupervised algorithms described above. Taking a supervised approach allows algorithms such as time-domain ICA, which are unable to separate monaural mixtures at all, to be extended to work in such situations. Jang and Lee [46] propose such a method which separates sources using predefined, source-specific ICA bases learned from training data. This is quite similar to the basis decomposition methods for undetermined source separation described in section 2.1.2 except the bases are known in advance. Given such predefined bases, separation can be thought of as simply finding the projection of the mixed signal onto each basis set. If the subspaces spanned by the different basis sets do not overlap, then the sources can be accurately recovered. In the results reported in [46], this approach was shown to work well on mixtures of speech with a music source, but performed significantly worse on mixtures of male and female speech due to fact that the speech signal bases had significantly more overlap, despite the inherent gender differences.

Computational auditory scene analysis

Another class of models for source separation that have been widely studied are based on a computational model of the auditory scene analysis processes used in human audition [11]. Research in this area dates back many decades [75, 115] and comprises some of the oldest work on source separation. Computational auditory scene analysis (CASA) systems work by using low level perceptual cues to segment a time-frequency representation of a mixture into regions consistent with being generated by a single source. Such *grouping* cues include: consistent pitch and harmonic structure as used in very early CASA systems [75], synchronous changes of spectral energy as in common onset, offset, amplitude and frequency modulation, and spatial location cues when at least two channels are observed [73]. Cooke and Ellis give a detailed review of CASA grouping cues in [15]. Time-frequency cells with consistent cues are grouped together using a heuristic, rule-based approach (e.g. [42]) to form spectral masks for each source which can then be used for separation. More advanced machine learning methods have also been proposed for segmentation, e.g. [6] which segments sources using spectral clustering of perceptual grouping cues.

Most of these perceptual cues are only applicable to harmonic signals such as voiced speech and thus are limited in terms of the types of signals they can separate. Ellis [27] describes a CASA separation system based on a broader set of parametric signal models suitable to describe segments of more complex signals, e.g. noise bursts, transients, or harmonic series. Similarly, for separation of mixtures of speech signals which contain both voiced and unvoiced sounds, Wang and Hu [114] extend the traditional CASA approach to use a classifier to identify contiguous unvoiced time-frequency regions to improve speech separation performance. Unlike most CASA separation systems, which often operate in an unsupervised manner, this approach requires the use of training data.

Separation as a classification problem

Given training data specific to a particular target source, one of the simplest approaches to monaural source separation is to train a set of classifiers to predict whether or not a particular time-frequency cell of a mixed signal spectrogram is dominated by the target source. Seltzer et al. [98] describe such an approach to identifying speech-dominated time-frequency regions in mixtures of speech and non-stationary noise. They use features that are independent of the interfering signal including periodicity-based features to detect voiced speech regions and more coarse energy-based features to detect unvoiced regions. Although the work is targeted at noise robust automatic speech recognition, the ideas are also applicable to simple separation tasks. Weiss and Ellis [116] describe a similar classifier approach to mask estimation based directly on the observed mixture spectrogram without any more involved feature extraction.

The primary disadvantage of this approach is that the classifier is only able to recognize a specific type of target signal in mixtures similar to those on which it was trained. Separation quality will be poor if the target or interfering sources are sufficiently different from the training data, or the target signal is not sufficiently distinct from the interference. For example, this approach would be unable to separate mixtures of two speech signals unless it was trained on mixtures composed of utterances from the same set of speakers. While such an approach could work if the identities of the sources present in a mixture are known in advance and sufficient training data is available, these assumptions are quite unrealistic. It might be possible to train classifiers specific to many possible source combinations, but such an approach quickly becomes intractable as the number of possible source identities increases. Because of these problems, this approach is generally only suitable for simpler denoising tasks.

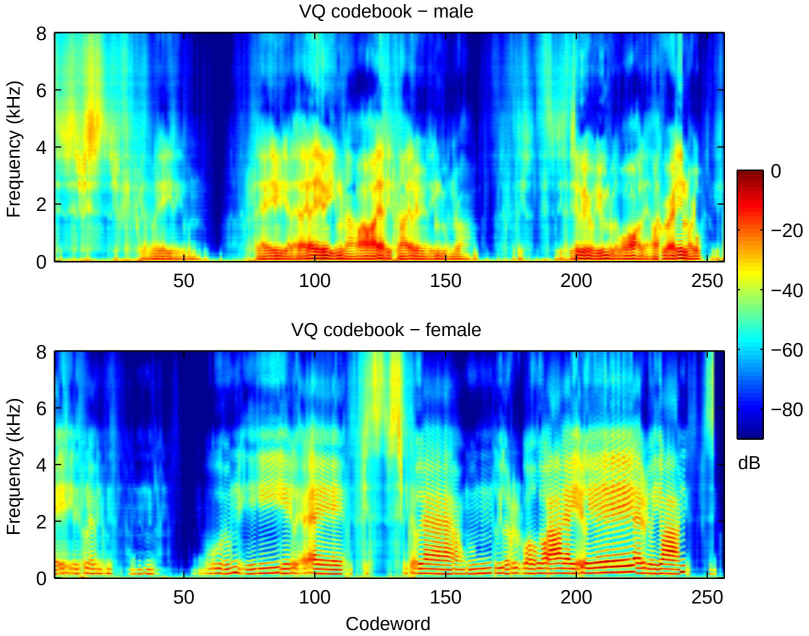


Figure 2.2: Example VQ codebooks of the power spectra in decibels learned from both male (top panel) and female (bottom panel) speakers.

Statistical model based separation

The idea of utilizing source-dependent information described above has merit, provided that the constraints learned from the training data do not simultaneously depend on the specific identity of all of the sources in the mixture. Roweis [92] describes the "refiltering" approach to spectral mask estimation from a monaural mixture of two speakers based on simple vector quantizer (VQ) [37] models of each of the underlying sources. The source models are trained independently on clean training data from each speaker. They consist of a codebook of exemplars that characterize the magnitude STFT of different sounds produced by a specific speaker. The training process essentially involves "memorizing" example STFT frames. An example of such VQ speech models for two different speakers is shown in figure 2.2. Given a set of speaker specific VQ codebooks, separation can be achieved by finding the best-matching codewords consistent with an observed mixture. However, the resulting quantized representation, which only coarsely represents the true source and lacks phase, is not directly invertible. Instead, estimated magnitudes from the inference are used to derive a spectral mask which can then be used to recover an estimate of the original target source.

More recently, this approach has been refined to use more sophisticated source models and allow for more accurate signal reconstruction [54, 88, 81, 41]. These extensions will be described in detail in chapters 3 and 4. The improved source reconstruction method of Reddy and Raj [81] highlights one of the advantages of statistical model-based separation when compared to other spectral masking based methods. Source reconstruction based on spectral masking simply discards time-frequency regions dominated by interfering sources. However, given the model of the underlying source spectrum used to estimate the mask, it is also possible to use the constraints from the model to estimate the source signal even in regions where the mask is zero. In the experiments reported in [81], this gives an improvement of 3 to 5 dB signal-to-noise ratio (SNR) over reconstruction based on simple spectral masking.

Source model considerations

The VQ approach to model-based separation is quite similar to the supervised NMF separation system described earlier. Both methods utilize source-specific constraints in the time-frequency domain in the form of a set of spectral templates learned from clean training data similar to those shown in figure 2.2 (although the model representation should be in units that combine linearly for NMF as opposed to the log-power spectra depicted in the figure). In fact, recent work by Rennie et al. [85] describes a separation system that combines ideas from NMF and sparse coding with spectral model based separation that has both approaches as special cases. The primary difference between the two approaches is in the way the models are used to reconstruct the sources. In the NMF case, linear combinations of source templates in the B_i matrix are used to reconstruct source spectrograms in a process similar to time-domain basis decomposition whereas in the VQ case the templates are used to define a probability distribution over the sources where each time frame of the source signal can only be derived from only one of the codewords. This distinction is further blurred if the NMF signal model incorporates sparsity constraints which enforces that relatively few NMF bases are used to represent each frame.

Obtaining high quality separation using either method requires the use of high quality source models that are specific to the different sources comprising the mixture and sufficiently different from one another to disambiguate the sources. If identical source models are used to represent each source, the resulting separations will be ambiguous, with some parts corresponding to one source while other parts corresponding to others, essentially *source permutations*. This makes e.g. separation of a mixture of multiple speech signals using a speaker-independent source model a very difficult problem. Although this can be overcome to some extent by enforcing additional temporal constraints through the use of convolutive NMF [101] or Markovian dynamic constraints in the VQ case [91], separation quality will almost certainly be improved if source-dependent models are used instead.

It is therefore important to use a representation in which the sources will be easy to differentiate when constructing the source models. Because natural audio sources tend to be sparsely distributed in time and frequency, it is therefore necessary to use an STFT representation with a high frequency resolution. Selecting a representation that exploits this property also allows for efficient inference because the full range of combinations of source model states need not be considered when their overlap is minimal [92].

In the particular case of speech sources, which are the focus of this thesis, the use of such a representation allows the source models to capture the wideband features such as the high-frequency formant resonances characteristic of vowel sounds and noise bursts characteristics of fricatives, and to fully resolve the narrowband features, highly source-

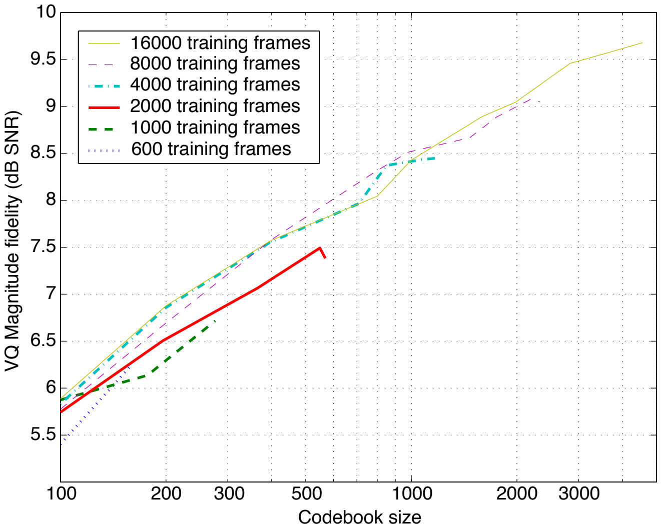


Figure 2.3: VQ fidelity (as SNR of STFT magnitudes) on a held-out test signal as a function of codebook size (horizontal axis) and training set size (different traces). Source: Ellis and Weiss [29]

dependent details such as the harmonic series characteristic of pitched speech sounds. Such details that serve to discriminate between sources are clearly visible in the different source models shown in figure 2.2. The significantly higher fundamental and formant frequencies in the female codebook which range around 200 Hz and up to 5 kHz respectively, are easily distinguishable from those in the male codebook whose fundamentals are not as well resolved as in the female model and are closer to 100 Hz, and whose formants only extend up to about 4 kHz.

It is worth noting that the wideband and narrowband spectral features of speech are often treated independently, e.g. in the source-filter model commonly used in applications such as speech coding [20, Chapter 5]. If these features are truly independent then it should be possible to model each of them independently using simpler models whose combination should be equivalent to the corresponding monolithic VQ model. This idea was proposed in [53, 33, 80]. Radfar et al. [80] describe a model-based separation system based on this idea where simpler VQ codebooks are only used to model the wideband spectra. This system works well when compared to VQ separation using identical speaker-independent source models, but does not perform as well as VQ separation using source-dependent models. This implies that it is important for the model to capture the relationship between narrow- and wide-band features which is implicitly learned by the monolithic VQ model.

The final consideration to be taken into account when designing template-based source models is the number of NMF bases or VQ codewords to use. Most interesting source types are quite complex and generally require large models to adequately cover the space of possible signals that can be generated by a particular source. Ellis and Weiss [29] investigated the relationship between VQ codebook size and quantization error on speech data from a single male speaker. The results are reproduced in figure 2.3. Based on this plot it is clear that even with very large models it is difficult to exactly model the source signal with very high fidelity. Even with a 2000 entry codebook, the magnitude of the source signal can only be represented with an SNR of 9 dB. However, the idea of representing the source signal exactly using a VQ model is somewhat naive. In separation algorithms like those described in the previous section, the source models are used in conjunction with each other to estimate a binary mask which describes which source dominates each time-frequency cell. The source signals are not reconstructed using the quantized representation based on the VQ codebook, so it is not important that they be extremely detailed. The codebooks serve to more loosely constrain the signal space. Separation is based on the combined constraints obtained from the full set of source models used to describe the mixture.

VQ and NMF source models containing a few hundred entries have been shown to work well in practice. Radfar and Dansereau [79] report similar performance to those shown in figure 2.3 using VQ speech models on a speech separation task where increasing the size of the codebook from 16 to 512 entries improved the average source reconstruction SNR by about 3 dB. The overall SNR using the largest models was about 7 dB. Schmidt and Olsson [97] describe roughly comparable results on a similar task and data set using a sparse NMF signal model.

Most statistical model or supervised NMF based systems in the literature use models based on representations similar to the one depicted in figure 2.2. Although quite powerful, these approaches have a significant disadvantage in their dependence on predefined, source-dependent models. This requires that the identities of all sources be known in advance and that sufficient data be available to train large models for each of them. One of the main contributions of this thesis is an adaptable source model that relaxes these assumptions. This will be described in detail in chapter 3. Although we will describe the adaptation model in the context of statistical model based separation similar to the VQ method described in this section, the underlying ideas are more universal and could be derived to work in the context of NMF based separation systems as well.

2.3 Discussion

In this chapter we have surveyed a wide variety of methods for audio source separation. In a broad sense, all approaches can be thought of as forms of constrained optimization where source signals are estimated to be consistent with the observed mixture under constraints such as mutual statistical independence. The underlying theme is that as the number of observations decreases and the similarity of the underlying sources increases, the more constrained the separation algorithm must be. In the extreme case of separation of monaural mixtures of multiple speech signals (the cocktail party problem), the most successful separation methods rely on access to source-specific training data to learn tight constraints in the form of source models.

Recently there have been a number of efforts to systematically evaluate many of the different separation methods described in this chapter. Evaluations of multichannel separation include the 2007 Stereo Audio Source Separation Evaluation Campaign [109] and the 2008 Signal Separation Evaluation Campaign [110] which evaluated a wider variety of separation tasks. The results of the most recent evaluation led the organizers to conclude that many of the separation problems outlined in this chapter, namely separation of overdetermined mixtures and two-channel, underdetermined instantaneous mixtures, are essentially solved with many algorithms reconstructing source signals with source-to-interference ratios (SIR) of about 15–20 dB. The separation of reverberant mixtures remains a challenge, with the best performing systems obtaining an average SIR of around 6 dB.

The 2006 Speech Separation Challenge [16] similarly evaluated separation of instantaneous, monaural mixtures. Source reconstruction results on this data set are similar to those obtained on reverberant mixtures in [110], with reconstruction SNR of about 6 dB reported [97, 79]. The best performing systems on this task utilized a statistical model based approach. In general such systems tended to outperform those based on CASA. We review the results of this challenge in more detail in section 4.3.

Based on these and other results discussed in this chapter, we can conclude that statistical source models are an important tool in the arsenal of source separation algorithms, especially in extremely underdetermined situations. This approach is generally complementary to other methods even in cases where unsupervised methods are known to work well. For example, the addition of source models has been shown to improve performance of separation systems based on spatial location [122, 120]. We will describe such a system in detail in chapter 5. Similarly, source-dependent models have also been usefully used as additional grouping cues in CASA systems to help group source segments over time [103].

Although the use of constrained source models has the seemingly onerous requirement of access to source-specific training data, in many applications this is not debilitating. In many computer audition tasks where model-based separation can often be of significant benefit, such as automatic speech separation, input signals are often known in advance to come from a constrained set of signals (e.g. speech). Furthermore, the availability of training data is often a prerequisite for such a system, justifying the use of pre-trained source models. Of course, it is generally not possible to have access to training data specific to *all* possible sources which makes it difficult to leverage source-dependent constraints for separation. In this thesis we propose the idea of using source adaptive models which can be used to learn speaker specific constraints from a limited set of training data that can be generalized to previously unseen test data. The design of such a model will be described in detail in the next chapter.

Chapter 3

Speaker Subspace Model

In this chapter we describe an extension to traditional hidden Markov speech models that incorporates ideas from the basis decomposition methods described in the previous chapter and can accurately capture the speaker-dependent characteristics of speech signals across a broad set of speakers. It is based on subspace methods for dimensionality reduction commonly used across a variety of pattern recognition applications, e.g. for face recognition [107], gene expression analysis [5], and drum pattern visualization and classification [28]. Initially proposed as a method for rapid speaker adaptation in automatic speech recognition given a small amount of adaptation data [32, 57], this type of model has also had much success in the area of speaker verification [105, 65, 49, 50].

We begin by discussing the construction of a speech model suitable for model-based source separation including considerations for feature representation and model topology. We then discuss methods for modeling a set of speaker constraints. Finally, we discuss the eigenvoice method for speaker adaptation and introduces some novel extensions for adapting additional model parameters and compensating for channel mismatch between train and test data.

3.1 Signal modeling for source separation

As described in the previous chapter, the success of statistical model based source separation generally depends on the strength of the signal constraints enforced by the model. Perhaps the best studied application of statistical speech modeling is automatic speech recognition (ASR) in which a speech recording is automatically transcribed into a string of words. The standard statistical approach combines top down constraints on allowed word sequences using a language model, with bottom up signal constraints using an acoustic model. A model similar to this should be appropriate for model-based source separation, however there are significant differences in the type of acoustic features that are appropriate for these different applications.

3.1.1 Spectral constraints

Typically, ASR acoustic models are deliberately designed to ignore acoustic features that are irrelevant to word identity. Removing uninformative and potentially confounding features can improve modeling performance for a particular task. However, many of these features are exactly those that are important for high quality separation. ASR acoustic models use a feature representation such as Mel-frequency cepstral coefficients (MFCC) which, while appropriate for capturing the linguistically relevant speech features (i.e. formant frequencies), discard high resolution spectral details that distinguish different speakers (e.g. pitch which only carries prosodic information in most Western languages). In fact, significant lengths are usually taken to normalize features across different classes of speakers using e.g. vocal tract length normalization [25]. Furthermore, ASR models are often trained in a speaker-independent manner by averaging across a wide variety of speakers, which further decreases the amount of spectral detail that is captured.

In contrast to ASR, spectral characteristics particular to individual speakers are key to separating different voices in a mixture. These include details such as the fundamental frequency range and characteristic formant resonances. As described in the previous chapter, in order to effectively leverage the sparse nature of natural audio, and speech in particular, this type of model generally requires the use of a representation with high frequency resolution. This ensures that the harmonics of pitched speech sounds are individually resolved, decreasing the likelihood that distinct sources will overlap when mixed together. Again, this is very different to MFCC-like features which are by design based on very broad frequency bands.

These considerations motivate us to model source statistics using a standard time-frequency representation based on a linear frequency short-time Fourier transform with a long analysis window to capture frequency detail with high resolution. Throughout this thesis our models use a log power spectral representation based on a STFT with a 40 ms Hamming analysis window and 10 ms hop between adjacent frames. Each frequency band of this representation has a bandwidth of 25 Hz.

Another important observation is that the space of possible speech signals, even if limited to those generated by a single speaker, is quite complex. Thus the distribution over speech spectra is inherently multimodal. For example, the spectra of vowel sounds are quite different from those of unvoiced consonants. As discussed in the previous chapter, this necessitates the use of a complex distribution. The standard approach to solving this problem is to model the full space using a mixture of simpler probability distributions, typically multivariate Gaussian distributions. The likelihood of a speech frame in the representation described above, $\boldsymbol{x}(t)$, under the GMM distribution can be written as follows:

$$P_{GMM}\big(\boldsymbol{x}(t)\,|\,\theta\big) = \sum_{c=1}^{C} \pi_c\, \mathcal{N}\big(\boldsymbol{x}(t);\, \boldsymbol{\mu}_c,\, \Sigma_c\big) \tag{3.1}$$

The distribution is broken into C distinct components, corresponding to a distribution with up to C modes. Each component consists of a single Gaussian distribution parametrized by mean $\boldsymbol{\mu}_c$, covariance matrix Σ_c, and mixing weight π_c that ensures that the distribution integrates to one. We use $\theta = \{\boldsymbol{\mu}_c, \Sigma_c, \pi_c\}_{c=1..C}$ as shorthand for the set of model parameters. Given a sufficient number of mixture components, such a Gaussian mixture model (GMM) is capable of approximating any continuous distribution.

This is a natural extension of the VQ dictionary model described in chapter 2. Each component of the GMM essentially behaves as an exemplar with some variance around it. Finally, we note that we use diagonal covariance matrices for all speech models described in this thesis. This implies that given a particular mixture component, each frequency subband of the observation is treated independently. Although this reduces each component's ability to tightly capture constraints on the signal, this can be mitigated by using a larger number of Gaussian components. As we will see in chapter 4, this limitation allows for significant speedups in the separation algorithm.

Because each time frame is treated independently under the GMM model, the likelihood of the overall observed signal, $\boldsymbol{x}(1..T)$, can be written as:

$$P_{GMM}\big(\boldsymbol{x}(1..T)\,|\,\theta\big) = \prod_t P_{GMM}\big(\boldsymbol{x}(t)\,|\,\theta\big) \tag{3.2}$$

It is clear however that this independence assumption is a significant weakness. Speech signals tend to evolve on time scales longer than a single frame. Therefore, adjacent frames are typically highly correlated, so an accurate speech model would have to take the relationship between adjacent frames into account. This is the subject of the following section.

3.1.2 Temporal constraints

In some situations, e.g. in mixtures of multiple talkers of the same gender, the spectral differences between sources are not always sufficient for effective separation. In such cases, the addition of temporal constraints on the signal has proved to be quite effective. Such constraints can operate over different time scales, from simply enforcing that adjacent STFT frames do not differ too much (e.g within an individual phoneme), to enforcing that certain phones do not follow others, or at the extreme, to enforcing that the signal adhere to a particular grammar or predefined phone sequence as in ASR systems.

As shown in Kristjansson et al. [55], incorporating temporal dynamics into source models can significantly improve separation performance. This is especially true when all sources in a speech mixture are being fit using the same model. In this case, the only hope for obtaining high quality separation is to enforce sequential constraints on the temporal evolution of the signals to ensure that the recovered source signals are unique [26]. Given such strong top down constraints, it is possible to so tightly constrain the space of possible source signals that mixtures of different utterances by the same speaker can be well separated. By leveraging a priori knowledge of the constrained grammar from which the utterances where generated, Kristjansson et al. [55] were able to obtain performance on par with human listeners on such mixtures mixed at 0 dB signal-to-noise ratio.

This motivates the use of a framework similar to that commonly used in automatic speech recognition (ASR), which combines temporal constraints on a short time scale within individual phones, and constraints on a longer time scale using a language model to define allowable phone and word sequences [20, Chapter 13]. Unfortunately, the incorporation of

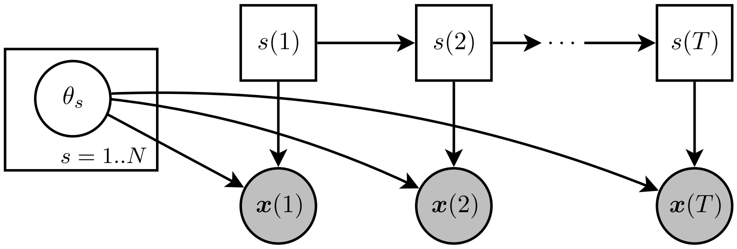


Figure 3.1: Graphical model representation of a hidden Markov model. Square nodes represent discrete variables and round nodes represent continuous variables. Shaded nodes correspond to observed variables. The observation at time t, $\boldsymbol{x}(t)$, is generated by a particular model state $s(t)$ which corresponds to a particular emission distribution parametrized by θ_s. N is the total number of states. Each state depends on the previous state, enforcing constraints on temporal dynamics.

a typical large vocabulary ASR language model is infeasible to use in practice due to its large cost in terms of computational complexity. While it is straightforward to incorporate simpler language constraints from a limited grammar as in [55], the resulting speech model is unable to capture utterances generated outside of that grammar.

3.2 Single speaker speech model

Based on the motivations outlined in the previous section, we model the speech signal using a hidden Markov model (HMM). The HMM is a natural extension of the GMM to incorporate temporal constraints by explaining different observations using different GMMs which are selected using a time-varying prior distribution. The overall likelihood of the signal under this model can be written as follows:

$$P_{HMM}\big(\boldsymbol{x}(1..T) \mid \theta\big) = \sum_{s(1..T)} \prod_t P\big(s(t) \mid s(t-1)\big)\, P_{GMM}\big(\boldsymbol{x}(t) \mid \theta_{s(t)}\big) \tag{3.3}$$

This resembles the GMM likelihood in equation (3.2), except the emission distribution P_{GMM} depends on a state variable $s(t)$. This variable selects a particular GMM for each observation which depends on the state used at the previous observation through the transition distribution $P\big(s(t) \mid s(t-1)\big)$. Because of this dependence, a speech signal under this model can be represented simply as a sequence of states. A graphical model representation of this structure is shown in figure 3.1.

To simplify the notation in the remainder of this thesis we will assume that all HMM emission distributions are single Gaussians (i.e. GMMs with a single mixture component). Such a model can exactly represent a more complex GMM emission distribution with multiple mixture components by separating each component into its own single Gaussian

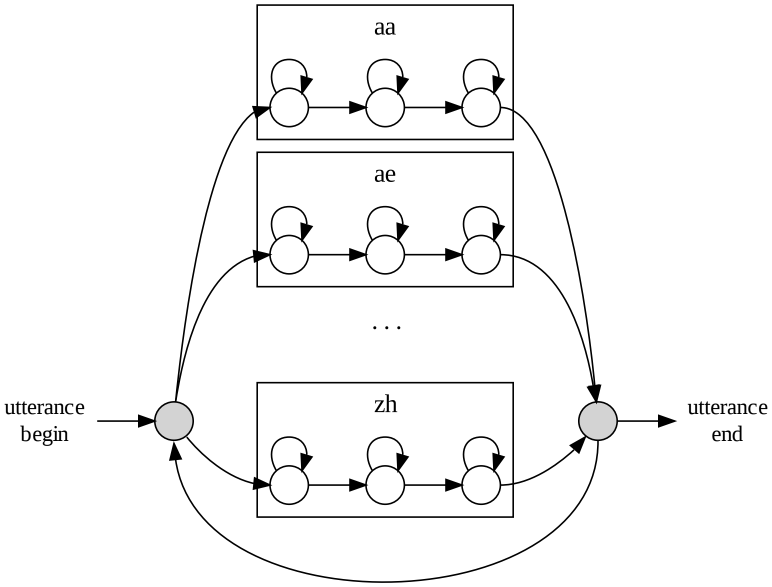


Figure 3.2: Finite state machine representation of the speech HMM topology. Non-emitting states are shaded. Individual phonemes use the standard 3 state forward structure. Transitions between phones all have the same probability.

state and utilizing identical transition probabilities for all Gaussians derived from a given GMM.

In contrast to Kristjansson et al. [55], we are interested in creating a more generic speech model that is not specific to a given grammar, so we only incorporate limited temporal dynamic constraints that do not depend on any prior high level knowledge about signal content. The resulting acoustic model therefore follows the ″phonetic vocoder″ approach for low bitrate speech coding [77], which models temporal dynamics only within each phone. The transitions from each phone to all others have equal probability, which we have found to work as well as more phonotactically-informed values. The resulting structure is illustrated in figure 3.2. It allows us to incorporate some knowledge of speech dynamics without being specific to any grammar. This is similar to the approach taken by Schmidt and Olsson [97] who utilize phone models trained using non-negative matrix factorization.

3.3 Generalizing across speakers

As described earlier, the key to good performance on source separation tasks is the use of source models that are specific to a particular source. If a source model is too general it is impossible to distinguish between the different sources for which it is a good fit. In the simplest case, the sources present in the mixture are known in advance, and models specific to each of those sources are available. This makes separation a matter of finding the best combination of signals under each model that, if combined appropriately, would match the observed mixture. In general, this is not a very realistic scenario because the space of possible source identities in an arbitrary mixture is quite large, and it is unlikely that sufficient training data will be available for all possible sources.

3.3.1 Model selection

A slightly more difficult variation on the separation problem is one in which the constituent sources are known to come from a closed set for which models have already been trained. This is the situation in the 2006 Speech Separation Challenge [16]. Model-based separation in this case can be done in a number of ways. The brute force approach is to try all possible model combinations and choose the best performing separation as measured using e.g. model likelihood [7, 97]. This obviously scales quite poorly as the total number of combinations to be evaluated grows exponentially with the number of source models in the training database. A variation on this approach is to evaluate the entire set of models in parallel (e.g. by building an HMM with the topology in figure 3.2 but with the phonemes models replaced by full speaker models), and allow the separation algorithm to find the model path with the highest likelihood [112].

The most scalable approach, however, is to break the problem down into two stages. First the source identities must be determined by choosing the set of models from the pre-trained database that best fit the mixture, then the signals can be separated using only the selected models. This is the approach taken by Kristjansson et al. [55] and Barker et al. [8]. It can be significantly more efficient than the previous approaches because the first stage can utilize a simple algorithm that is insufficient for high quality separation but good enough for model selection. Such an algorithm is described in detail in [83] and in section 4.2.1.

The most general setting however is when the precise source characteristics are not known a priori, and in fact, there is no training data available for the sources in question. This is the situation with the most general applicability, and the one we are interested in addressing. Source-dependent speaker modeling then becomes a matter of utilizing whatever training data is available to help construct a model that is a good fit to previously unseen sources.

One approach to this problem would be to simply ignore the potential mismatch between the speakers in the training set and those in the mixture, and proceed with the model selection approach outlined above. This will select the models from the training set that best fit the speakers in the mixture. If the training set adequately covers the full space of speakers, then this approach can be sufficient. It is analogous to quantizing the space of speaker models based on available training data and then selecting the closest “quantization level” at evaluation time. The downside to this approach is that the selected models could be a poor fit to the source in question, resulting in poor separation performance. This might occur if the training speakers represent a sparse sample of the full space of possible speaker models and the mixture speakers were far away from those used in training (e.g. if only male speakers models are available to model female speakers).

3.3.2 Model adaptation

The model selection approach is not the only way mismatched training and test data can be utilized. Ozerov et al. [74] address a similar situation in the setting of monaural singing voice separation. They propose the idea of beginning with a source-independent (SI) model and adapting it to the target source using maximum likelihood linear regression (MLLR) [62]. This is analogous to acoustic model adaptation used to improve the performance of automatic speech recognition in situations where a limited amount of adaptation data is available for a particular speaker [123]. This approach can separate previously unseen sources far better than one that uses unadapted models, but it requires a substantial amount of adaptation data. In this work we consider adaptation when there is much less data available, only a single utterance, requiring a more constrained model space.

The MLLR adaptation approach used in Ozerov et al. [74] defines a speaker-adapted (SA) model as an affine transform of the parameters of the SI model, $\bar{\theta}$, which was trained over a large corpus of speech from multiple talkers. This involves using a small set of adaptation data from a single speaker to learn a set of warping and offset coefficients, W and b, by which the SI model means, $\bar{\boldsymbol{\mu}}$, will be rotated and offset, respectively, to more closely match the adaptation data. In the simplest case, a global warping is applied across all states. The MLLR-adapted likelihood of a frame of data under HMM state s can be written as follows:

$$P_{MLLR}\big(\boldsymbol{x}(t)\,|\,s,\,W,\,b,\,\bar{\theta}\big) = \mathcal{N}\big(\boldsymbol{x}(t);\,W\bar{\boldsymbol{\mu}}_s + b,\,\bar{\Sigma}_s\big) \tag{3.4}$$

where s indexes the particular state against which $\boldsymbol{x}(t)$ is being matched.

While this approach has been extremely successful in improving ASR performance given fairly limited amounts of training data, the amount of adaptation data needed to properly learn the warping parameters is on the order of a few tens of utterances, too many for our intended application. The amount of adaptation data needed to adequately learn a set of parameters without over-fitting generally scales with the dimensionality of the parameters being learned [10, Chapter 1]. The high dimensionality of our signal representation relative to that typical of ASR features further exacerbates this problem.

One way to solve this problem would be to enforce tighter constraints on how the model can be adapted. For example, we could do this by lowering the dimensionality of the adaptation parameters by constraining the form of W to have a particular structure, e.g. block diagonal. This has the detrimental effect of reducing the extent to which the model can adapt to the speaker characteristics.

An alternative method for reducing the amount of adaptation data needed is to directly embed additional knowledge into the adaptation framework. In MLLR adaptation the only prior knowledge needed are the parameters for a pre-trained SI model. In many of the scenarios outlined in the previous section a set of pre-trained speaker-dependent (SD)

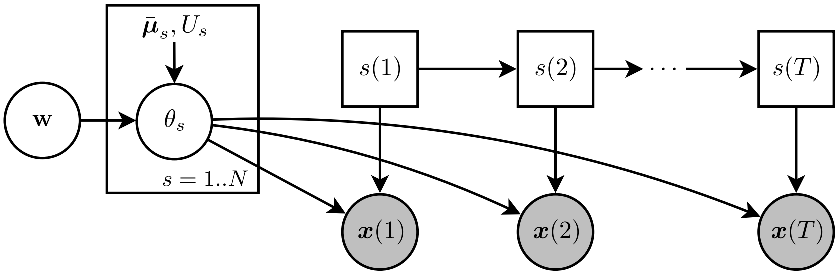


Figure 3.3: Graphical model representation of the eigenvoice adapted speech model. The observed speech signal $x(t)$ is generated by a factor analyzed hidden Markov model with adaptation parameters determined by the value of $\mathbf{w}$. Note that in contrast to figure 3.1, θ_s, the Gaussian parameters for state s, are now random variables dependent on $\mathbf{w}$.

models is available as well. Kuhn et al. [57] propose the idea of "eigenvoice" adaptation for exactly this situation. The idea is to break down each SD model as a linear combination of a "mean voice", essentially corresponding to the SI model, and a set of basis vectors U. The likelihood under such a model can be written as follows:

$$P_{EV}\big(\boldsymbol{x}(t)\,|\,s,\mathbf{w},\bar{\theta}\big) \;=\; \mathcal{N}\big(\boldsymbol{x}(t);\,\boldsymbol{\mu}_s(\mathbf{w}),\,\bar{\Sigma}_s\big) \tag{3.5}$$

$$\boldsymbol{\mu}_s(\mathbf{w}) \;=\; \bar{\boldsymbol{\mu}}_s + \sum_k w_k\,\hat{\boldsymbol{\mu}}_{s,k} \;=\; \bar{\boldsymbol{\mu}}_s + U_s\,\mathbf{w} \tag{3.6}$$

where $\hat{\boldsymbol{\mu}}_{s,k}$ is the kth basis vector for state s, and $U_s = [\hat{\boldsymbol{\mu}}_{s,1}, \hat{\boldsymbol{\mu}}_{s,2}, \ldots, \hat{\boldsymbol{\mu}}_{s,K}]$. The graphical model representation of the resulting adapted speech model is shown in figure 3.3. The observations are generated by an HMM whose parameters are defined by a factor analysis model [124].

Essentially, the very high dimensional model parameters for a particular speaker are represented as a function of $\mathbf{w}$. In contrast to the MLLR adaptation approach in equation (3.4), the only parameters to be learned here are the weights, $\mathbf{w}$, which are generally of relatively low dimension. Because the number of parameters needed to describe a particular speaker is so small, this technique has the advantage of requiring very little adaptation data, making it suitable for adapting models to a single utterance. The bulk of the knowledge of speakers characteristics is embedded in the predefined speaker basis vectors U. Adaptation is just a matter of learning the ideal combination of bases, essentially projecting the observed signal onto the space spanned by U. As with MLLR, this can be done using an expectation maximization algorithm such as the maximum likelihood eigen-decomposition (MLED) algorithm [58] which will be described in detail in chapter 4.

It is worth noting that there are no restrictions placed on U. This adaptation framework was independently developed with slight differences in Gales [32] and Kuhn et al. [57]. Gales [32] interprets the bases as a set of speaker clusters (e.g. two bases devoted to male and female speakers) and describes a number of approaches to learning them directly from the training corpus. Taken to the extreme, U could be composed of the pretrained

SD model parameters in which case an adapted model is found by linear interpolation across the SD models. Because the adaptation parameters are continuous, this approach allows for smooth interpolation across the entire space, enabling it to capture a wider variety of SD models than were used in training. Alternately, if constraints are applied to $\mathbf{w}$ to make it behave more like an indicator variable (e.g. by enforcing sparsity constraints through a sparse prior distribution over $\mathbf{w}$), then the adaptation model essentially reduces to speaker dependent modeling and is analogous to speaker selection methods described in the previous section.

3.4 Eigenvoices

Kuhn et al. [57] take a different view of the basis set U. Their idea is to utilize a previously trained set of speaker-dependent models to generalize to other speakers. Each of these SD models can be thought of as a sample from a very high dimensional "model space" of possible models of a certain parametrization. The entire space of such models is extremely broad and covers far more than only valid speech models. If the space of valid speech models is assumed to form a "speaker subspace" embedded in this high dimensional model space, as illustrated in figure 3.4, then the properties of this subspace can be learned from the samples.

From this perspective, equation (3.6) should not be thought of as interpolation between arbitrary sets of parameters. Instead, an adapted speaker model is represented as a "mean voice" $\bar{\mu}$, essentially corresponding to the SI model, plus a linear combination of the basis vectors for the speaker subspace. The speaker subspace can be thought of as a space of inter-speaker variation, i.e. the model characteristics that distinguish different speakers.

Furthermore, it is reasonable to assume that the latent dimensionality of the speaker subspace will be relatively low, certainly considerably lower than that of the overall model space. This is guaranteed to be true primarily because there are at most as many degrees of freedom as there are SD samples. In any conceivable scenario there will be at most a few hundred speaker models, far fewer than the thousands of dimensions needed to fully specify the natural parametrization of a speech model, which must parameterize each of the Gaussians used to describe the dozens of different phonemes. Still, it is usually desirable to further reduce the dimensionality of $\mathbf{w}$ to minimize the amount of adaptation data needed. Furthermore, it is intuitively attractive to assume that there is a relatively small number (perhaps tens) of degrees of freedom in the space of possible speaker variation, related to different physical proportions of the vocal tracts.

Kuhn et al. [57] propose learning the subspace bases U using principal component analysis (PCA) [24, Chapter 3], a technique that naturally leads itself to dimensionality reduction. The eigenvalue corresponding to each "eigenvoice" basis vector found using PCA corresponds to the variance of the speaker models across that basis. So, retaining the K eigenvoice bases with the highest eigenvalues will result in a basis set that captures the maximum variation possible with K basis vectors. As a secondary benefit to this construction, the resulting bases are by definition orthogonal which has a side effect of making the adaptation parameters $\mathbf{w}$ uncorrelated, a fact that will be taken advantage of in section 4.2.2.1.

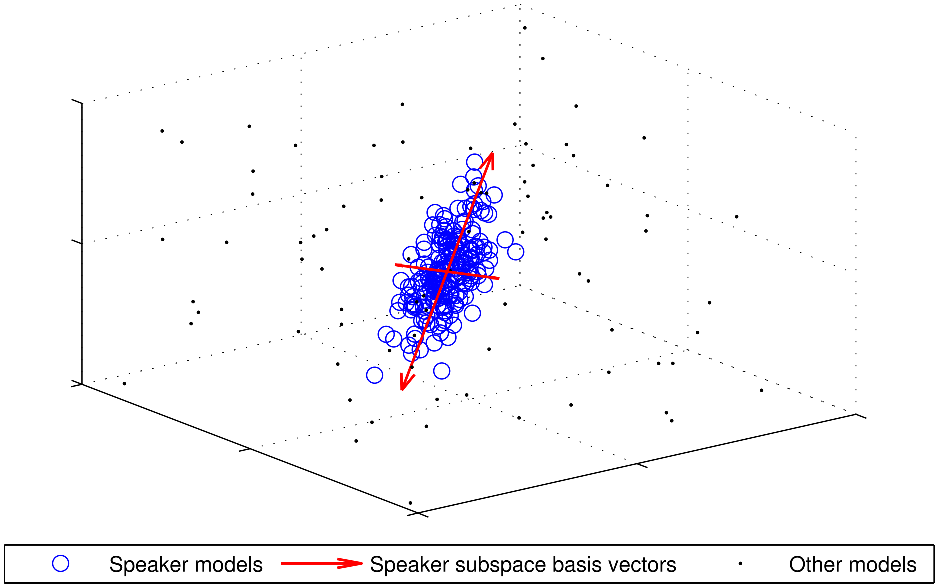


Figure 3.4: Illustration of a speaker subspace with latent dimensionality of two embedded in a three-dimensional model space. Valid speaker models are denoted by blue circles. Black dots denote random points in the space that do not correspond to valid speaker models. The principal axes of the subspace are drawn in red. These correspond to a rotation of the natural x, y, z bases centered on the mean of the SD model population. Any point in the subspace can be represented as linear combination of these bases.

3.4.1 Training

As described in the previous section, each pre-trained SD model can be represented as a single point in a very high dimensional model space. Each point is defined by concatenating the set of Gaussian means for all states in the model for speaker i into parameter supervector $\boldsymbol{\mu}_i$. The construction of the eigenvoice basis vectors requires the use of a speaker model parametrization that is consistent across all speakers. Although the ordering of states in the supervectors is arbitrary, care must be taken to ensure that the ordering is consistent for all speakers. This ensures that there is a one-to-one correspondence of HMM states across all speaker models. If GMM emissions are used, further complications are possible if there is no correspondence between mixture components across the speaker models. This will occur if the speaker models are trained independently using a process such as mixture splitting [126].

A simple way to guarantee a consistent mapping is to use an identical initialization for all speaker models. We begin by pooling all of the training data across all speakers and using it to train a speaker independent model. This model is then used to bootstrap each SD model to ensure that each state of the SD models corresponds directly to the same state in the SI model. We do this using MAP adaptation of the SI model [123] whereby the SI model is treated as a prior and the SD parameters are re-estimated in proportion to the amount of training data available. This ensures that the resulting SD models do not contain any unused states with degenerate parameters (e.g. zero covariances) in instances when there is insufficient training data to fully estimate the parameters. Such outliers would skew the resulting eigenvoice bases and lead to sub-par adaptation.

Parameter supervectors are constructed for all M speaker models and used to construct a parameter matrix $P = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_M]$ that spans the space of speaker variation. The mean voice $\bar{\boldsymbol{\mu}}$ is found by taking the mean across columns of P. Performing the singular value decomposition on $P - \bar{\boldsymbol{\mu}}$ then yields orthonormal basis vectors for the eigenvoice space, U.

Example eigenvoice parameters learned from the 34 speakers of the GRID training set Cooke and Lee [16] are depicted in figure 3.5. The mean voice and the three eigenvoices with the highest variance are depicted. The mean voice is very similar to the speaker-independent model and very coarsely models the overall spectral shape characteristic of different phones. Successive eigenvoices are used to add additional high resolution detail to this model. They each reflect easily interpretable and linguistically relevant features of speech spectra.

The eigenvoices with the highest associated variance reflect very broad differences in vocal tract shape. Successive bases incorporate additional fine spectral detail corresponding to characteristic pitch peaks. Eigenvoice 1, $\hat{\boldsymbol{\mu}}_1$, emphasizes formant resonances that are characteristic of female speakers. In fact, as shown in figure 4.5, the corresponding eigenvoice weight is perfectly correlated with gender; female speakers have positive w_1 and male speakers have negative w_1. This implies that the dimension of the training set with the most variation relates to gender. This is not surprising because the physical dimensions of the vocal tracts of male and female speakers tend to be quite different. Eigenvoice 2 emphasizes different formants in consonant states and introduces some fundamental frequency information and high frequency resonances into the vowel states. Finally, $\hat{\boldsymbol{\mu}}_3$ incorporates additional pitch trajectory detail into voiced phones.

3.4.2 Incorporating covariances

An important disadvantage of the speaker adaptation approaches described above is that only the mean parameters are adapted to match a specific speaker. As shown in [118], this significantly reduces the model's ability to unambiguously capture the statistics of a specific source. This is especially important when attempting to separate a mixture of two talkers of the same gender because the model means of the two sources can be quite similar, leaving the variance parameters to distinguish between the speakers.

Extending the eigenvoice model to incorporate covariance parameters introduces a few complications. First, the basic idea behind the adaptation model is to represent the adapted model parameters as a simple linear combination of some set of basis vectors. The covariance analog to mean-only adaptation in equation (3.6) would be to represent the adapted covariance as $\Sigma_s(\mathbf{w}) = \bar{\Sigma}_s + S_s\mathbf{w}$. But, unlike Gaussian means, covariances must remain non-negative (since our covariance matrices are all diagonal, covariance terms

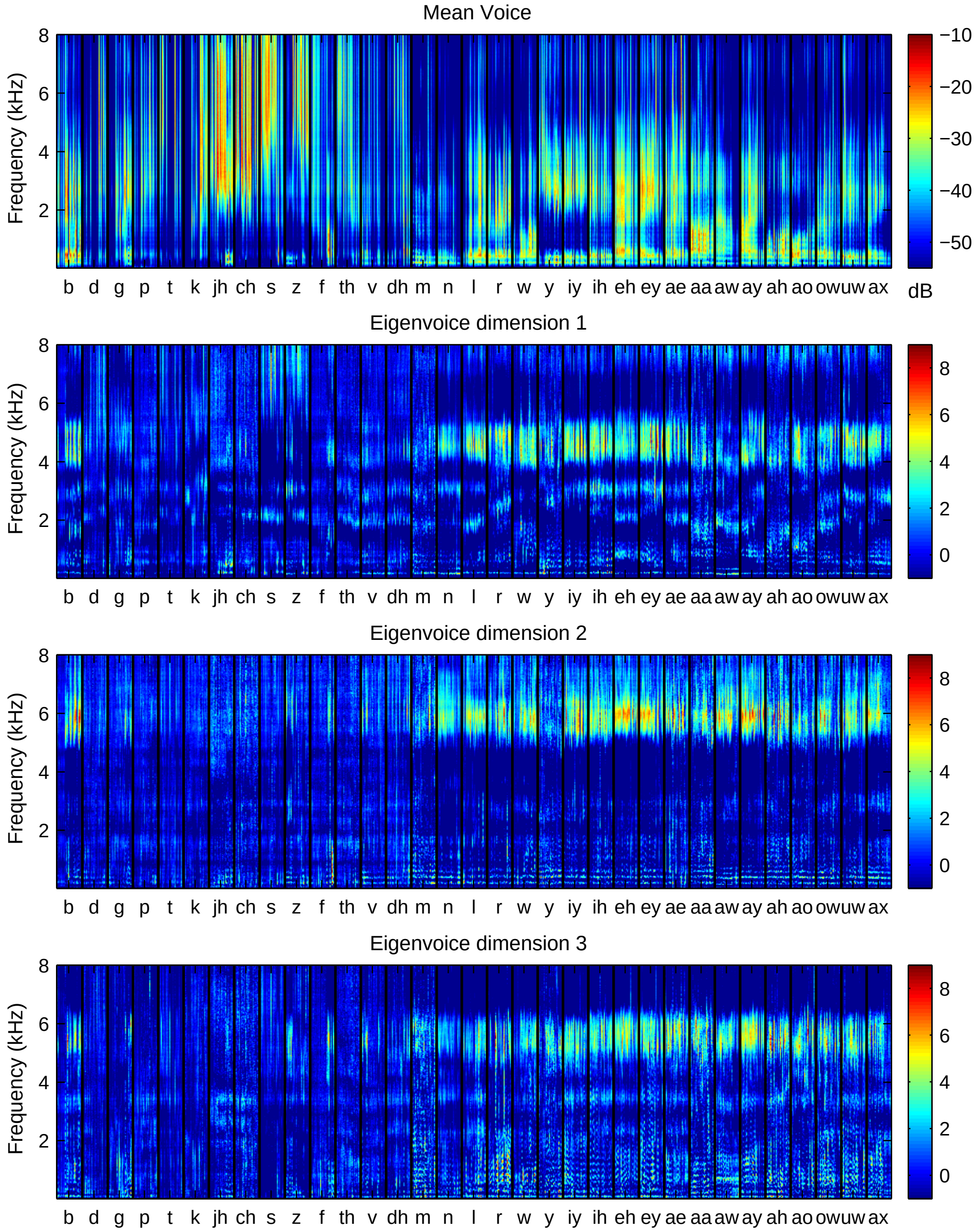


Figure 3.5: Eigenvoice basis vectors learned from the GRID data set [16]. The top panel shows the mean voice $\bar{\mu}$ which closely resembles the speaker-independent model. The remaining panel show the three eigenvoices with the largest variance, $\hat{\mu}_1, \hat{\mu}_2$, and $\hat{\mu}_3$ respectively. The black lines in all panels indicate the borders between different phonemes.

between different dimensions, which in general might be negative, are always zero and thus not modeled). In the formulation outlined above, there is no way to guarantee that an arbitrary weighted sum of basis vectors will satisfy these constraints.

One potential solution would be to represent U and $\mathbf{w}$ in such a way that they are both non-negative using non-negative matrix factorization [60]. This would require extending the MLED algorithm to ensure that $\mathbf{w}$ remain non-negative. An alternative solution, which we adopt, is to represent the covariance parameters in the log domain. This ensures that the adapted covariance matrices are always non-negative, regardless of the sign of $\mathbf{w}$. The adapted covariance matrix for state s can then be written as follows:

$$\Sigma_s(\mathbf{w}) = \exp\left(\log(S_s)\,\mathbf{w} + \log\bar{\Sigma}_s\right) \tag{3.7}$$

Training such a model now utilizes both the Gaussian means and the log-covariances. The parameters for all states for speaker i are concatenated into a parameter supervector $\boldsymbol{p}_i = [\boldsymbol{\mu}_i;\, \log\Sigma_i]$, representing the speaker model in the very high dimensional model space. The space spanned by all M training speakers can then be described by the matrix $P = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_M]$. As in section 3.4.1, a set of orthonormal basis vectors for the speaker subspace are constructed by applying PCA to P.

The final complication has to do with the learning algorithm for $\mathbf{w}$. As described in [43, 119], the straightforward extension to the MLED algorithm where both the Gaussian mean and covariance parameters depend on $\mathbf{w}$ requires performing a non-convex optimization. A potential solution is suggested in [58]. Assuming that the mean and covariance parameters have some correlation (e.g. the correlation coefficient of the mean and covariance parameters of the speaker models described in section 4.4 is 0.3), a reasonable approximation is to use the MLED algorithm that optimizes for the mean statistics alone and rely on the correlation between the different parameters implicit in the learned subspace to simultaneously adapt the covariances. This turns out to work quite well as we will see in the next chapter.

3.4.3 Channel compensation

In many instances, there will be additional mismatch between an observed signal and the corresponding SD or adapted model if the recording conditions do not match those of the training data. This could be caused by the use of a different microphone or communication channel used in recording the data, or in the mixture case, by different gains applied to each source relative to the others.

Since our signal model is based on log spectral features, such filtering appears as an additive offset to the observations. This is a natural fit to the subspace adaptation model. Because the adaptation model described in equation (3.6) makes no assumptions about the basis set combined to form the adapted model, it is straightforward to extend it to compensate for such channel effects. We do this by extending the eigenvoice basis set to include a rudimentary set of channel bases, B, tied across all model states. The resulting speaker and channel adapted model is parametrized by the speaker-dependent eigenvoice weights $\mathbf{w}$ and the channel weights $\mathbf{h}$:

$$\boldsymbol{\mu}_s(\mathbf{w}, \mathbf{h}) = \bar{\boldsymbol{\mu}}_s + U_s\,\mathbf{w} + B\,\mathbf{h} \tag{3.8}$$

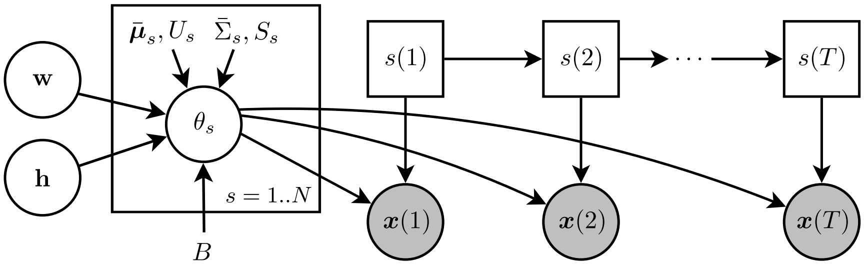


Figure 3.6: Graphical model representation of the full speaker subspace model including the extensions described in sections 3.4.3 and 3.4.2. The adapted model parameters for state s are a function of random variables $\mathbf{w}$ and $\mathbf{h}$ corresponding to the eigenvoice and channel weights respectively, and the predefined eigenvoice and channel basis parameters $\bar{\boldsymbol{\mu}}_s, U_s, \bar{\Sigma}_s, S_s$ and B.

Because B is independent of the signal state, it can capture filtering that is constant across the entire signal, provided that it has an effect that is consistent within each short-time analysis window, i.e. that its impulse response is compact relative to that window. The time varying portions of the observations, corresponding to the underlying speech source, are still well modeled by the eigenvoice speech model.

The number and type of channel bases to use depends on the particular data set being modeled. In general it is reasonable to assume that the channel filter has a short impulse response and thus will be relatively smooth across frequency. Given this assumption, a good choice of channel basis is a set of low-order discrete cosine transform (DCT) bases:

$$B_{ij} = \frac{2}{\sqrt{\Omega}} \cos\left(\frac{\pi}{\Omega}\left(i + \frac{1}{2}\right) j\right) \quad i = 0..\Omega - 1,\ j = 0..L - 1 \tag{3.9}$$

where Ω is the dimensionality of the feature representation, which, depending on the signal sampling rate, is generally on the order of a few hundred. L is the maximum DCT frequency, generally on the order of 10.

This channel compensation with a single, flat (i.e. DC) basis is used to compensate for unknown relative signal gains in section 4.3. A larger channel basis set is especially important for the binaural separation system described in chapter 5.

The graphical model representation of the full adaptation model, including adapted covariance parameters and channel compensation, is shown in figure 3.6. We note that the channel basis B and parameters $\mathbf{h}$ can simply be subsumed into the overall speaker subspace parameters:

$$U_s' = [U_s\ B], \quad \mathbf{w}' = \begin{bmatrix} \mathbf{w} \\ \mathbf{h} \end{bmatrix} \tag{3.10}$$

This convention is used throughout the remainder of this thesis to keep the notation simple. In situations where **w** and **h** need to be addressed independently, we will explicitly decouple them.

3.5 Discussion

The proposed subspace model shares some similarities with the basis decomposition models described in the previous chapter. One disadvantage of unadapted exemplar-based source models such as HMMs is that each frame of the observation is explained by a single fixed state. In contrast, basis decomposition models like NMF represent each frame with a combination of rescaled basis vectors. The NMF model also decouples the model for spectral shape from the scale of the observation. It therefore implicitly handles simple gain compensation through the per-basis scaling applied in the factorization. This allows an NMF model significantly more flexibility than simple HMM or VQ model of the same size. This is because a similar HMM model would need more states to explicitly model all valid combinations of NMF bases at different scales. The proposed subspace approach extends some of these advantages to the HMM signal model.

Figure 3.7 contrasts the eigenvoice adaptation approach with the model selection method described in section 3.3.1. The parameters of any model in the subspace can be represented exactly by equation (3.6), while the selection approach is limited by the coarseness of the training set. For example, points far from the mean voice are very poorly represented by the nearest model in the dictionary, but given an appropriate basis the interpolation method can still represent them accurately.

Of course the flexibility of the adaptation approach could potentially be a disadvantage in some situations. For example, if the training data is not very well matched to the test data, e.g. if presented with a mixture of speech and music the adaptation model should not be not so general that it would be able to model portions of the music signal as speech. This would lead to very poor separation performance. This could be solved to some extent through the use of a prior distribution on **w** to limit the extent to which an adapted model is allowed to stray from the training set. However, it is important to note that this is a problem with model-based approaches to source separation in general. The goal of the adaptation approach described in this chapter is to strike a balance between the flexibility of the model and the potential for such over-production.

A further concern with the eigenvoice construction is that it is based on the assumption that the space of SD models exists as a subspace of the full model space. There is no a priori reason to think that this will be the case. In fact, it is quite possible that the SD models actually live on a nonlinear manifold of even lower latent dimensionality. While this could be reasonably approximated with the linear decomposition of PCA, it will generally require the use of more basis vectors than are strictly necessary. Additionally, it runs a greater risk of overproducing speech models, i.e. the basis vectors could be combined to create models which do not represent any conceivable speaker. Tipping and Bishop [106] propose an extension to PCA to mixtures of locally linear PCA models for capturing this sort of non-linearity. If the linearity assumption is indeed a problem, the adaptation framework described here could be easily extended to use this mixture PCA model.

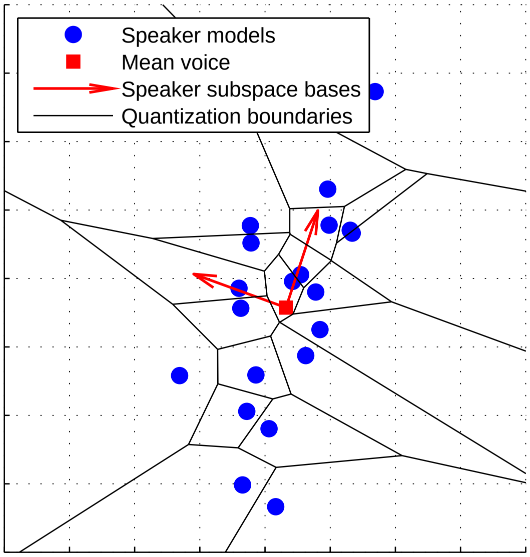


Figure 3.7: Illustration of the model selection and subspace adaptation approaches for representing arbitrary speaker models. The model selection approach (section 3.3.1) represents a point in the speaker model space by the closest of the speaker models in the training database (blue dots). The boundaries between these quantization levels are drawn in black. Alternatively, the subspace adaptation approach can represent any point in the space exactly as the sum of the mean voice $\bar{\mu}$ (red square) and a linear combination of the basis vectors U (red arrows).

Finally, if dimensionality reduction is not a priority, we note that it is not necessary to use PCA to find a good basis for the subspace. Other methods might provide useful advantages. For example, if the training set can be clustered into different classes as in [31] (e.g. by gender), learning the basis set using linear discriminant analysis (LDA) [24, Chapter 3] would result in bases that make the classes linearly separable (as was already the case for different genders when using PCA). This could make for more robust initialization of the EM algorithm based on such clustering. This is especially important for separation applications because initial separation of the sources has a significant effect on the system's performance (see section 4.2.2.1 for a detailed discussion). Bases based on independent component analysis (ICA) [44], which attempts to find bases that are statistically independent, could provide similar advantages. PCA retains the advantage of enabling optimal dimensionality reduction of the speaker subspace over these other methods. Furthermore, we note that PCA has worked quite well thus far, so an exploration of alternate constructions of the basis set remains as future work.

3.6 Summary

We have reviewed the necessary considerations for constructing high quality speech statistical models appropriate for source separation. We introduced the notion of the space of speaker models and describe various methods of accurately representing individual models in this space. The simplest and most often used approach is to quantize the space to a dictionary of pretrained speaker models. Instead we propose learning a linear basis set for the model space and represent a specific speaker model as an interpolation between these bases. The prior knowledge built into the subspace model allows it to rapidly adapt to match the statistics of a particular speaker given a small amounts of adaptation data. This approach is more flexible than simply quantizing the space and is better able to represent previous unseen speakers. Finally, we presented an extension to the adaptation model to compensate for any channel variation that was not present in the training data.

Chapter 4

Monaural Speech Separation

In this chapter we describe how the speaker subspace model described in chapter 3 can be applied to monaural speech separation. We model a monaural mixture of utterances from two speakers using a factorial hidden Markov model composed from the speaker HMMs. If, as is typically the case, the speaker-specific model parameters are not known a priori, they must be estimated from the mixture. We derive two algorithms for learning the adaptation parameters directly from the mixed signal and compare them to an algorithm for model selection from a predefined set of speaker parameters. We then review some experimental results comparing our adaptation based separation system to other state of the art techniques in the context of the 2006 Speech Separation Challenge. Finally we compare the performance of the proposed model adaptation algorithm and the model selection algorithm on mixtures of utterances from speakers not present in the training set.

4.1 Mixed signal model

In the time-domain, a monaural mixture of I sources, $y(t)$, is simply the sum of the contributions of each source, $x_i(t)$. As described in section 3.4.3, we assume that each source has been passed through an arbitrary channel with impulse response $h_i(t)$. We can therefore express the mixed signal in the time-domain as follows:

$$y(t) = \sum_{i=1}^{I} (x_i * h_i)(t) \tag{4.1}$$

Our source models operate in the log power spectral domain, which complicates the linear mixing in equation (4.1). In fact, the non-linear nature of the log operation makes the model of a particular frame of the log power spectrum quite complex:

$$\boldsymbol{y}(t) = \log\Big(\sum_i \exp\left(\boldsymbol{x}_i(t) + \mathbf{h}_i\right)\Big) \tag{4.2}$$

where $\boldsymbol{y}(t)$ denotes a frame of the short-time log power spectrum of the waveform $y(t)$ in decibels. Because our chosen representation of $\boldsymbol{y}$ uses high frequency resolution, the interaction between the sources is quite often negligible. This motivates the common "max" approximation [71] to describe the way two natural speech signals mix in this domain:

$$\boldsymbol{y}(t) \approx \max_i \left(\boldsymbol{x}_i(t) + \mathbf{h}_i\right) \tag{4.3}$$

where max denotes the element-wise maximum. In the remainder of this section we describe how to combine the source models described in the previous chapter using this mixing model to form a probabilistic model of the monaural mixture.

Throughout the remainder of this chapter we assume that mixtures are composed of only two sources. While the extension to more sources is straightforward to implement, it is in general quite slow as many of the algorithms that we will describe have a run time that grows exponentially with the number of sources.

4.1.1 Factorial hidden Markov model

As described in chapter 3, each clean source signal is modeled using a hidden Markov model. We can therefore use the combination of these source HMMs to model the mixed signal. This can be accomplished in a straightforward manner using a factorial HMM (FHMM) [35] constructed from the individual source models. The graphical model representation of our mixed signal model is shown in figure 4.1. Each source signal $\boldsymbol{x}_i(t)$ is generated by the factor-analyzed HMM described in the previous chapter. The speaker-dependent characteristics of source i are compactly described by the parameters $\mathbf{w}_i$ which are used to generate the Gaussian means and covariance parameters of the HMM emission distributions. Finally, the observed mixture $\boldsymbol{y}(t)$ is explained by the combination of the two hidden source signals.

The overall likelihood of all of the random variables in the model given the adaptation parameters $\mathbf{w}_1$ and $\mathbf{w}_2$ can be written as follows:

$$\begin{aligned} &P\big(\boldsymbol{y}(1..T), \boldsymbol{x}_1(1..T), s_1(1..T), \boldsymbol{x}_2(1..T), s_2(1..T) \,|\, \mathbf{w}_1, \mathbf{w}_2\big) \\ &\qquad = \prod_t P\big(\boldsymbol{y}(t) \,|\, \boldsymbol{x}_1(t), \boldsymbol{x}_2(t)\big) \prod_i P\big(s_i(t) \,|\, s_i(t-1)\big)\, P\big(\boldsymbol{x}_i(t) \,|\, s_i(t), \mathbf{w}_i\big) \end{aligned} \tag{4.4}$$

where

$$P(x_i \,|\, s_i, \mathbf{w}_i) = \mathcal{N}\big(\boldsymbol{x}_i; \boldsymbol{\mu}_{s_i}(\mathbf{w}_i), \Sigma_{s_i}(\mathbf{w}_i)\big) \tag{4.5}$$

If covariance adaptation described in section 3.4.2 is not used then the covariance in equation (4.5) is independent of $\mathbf{w}$, i.e. $\Sigma_s(\mathbf{w}) = \bar{\Sigma}_s$. Where convenient we will drop the dependence on $\mathbf{w}$ in the notation for the model parameters and use $\boldsymbol{\mu}_{i,s} = \boldsymbol{\mu}_s(\mathbf{w}_i)$ and $\Sigma_{i,s} = \Sigma_s(\mathbf{w}_i)$ to refer to the speaker model adapted to match speaker i. Finally, a prior on $\mathbf{w}_i$ can optionally be included as well:

$$P(\mathbf{w}_i) = \mathcal{N}(\mathbf{w}_i; \boldsymbol{\nu}, \Xi) \tag{4.6}$$

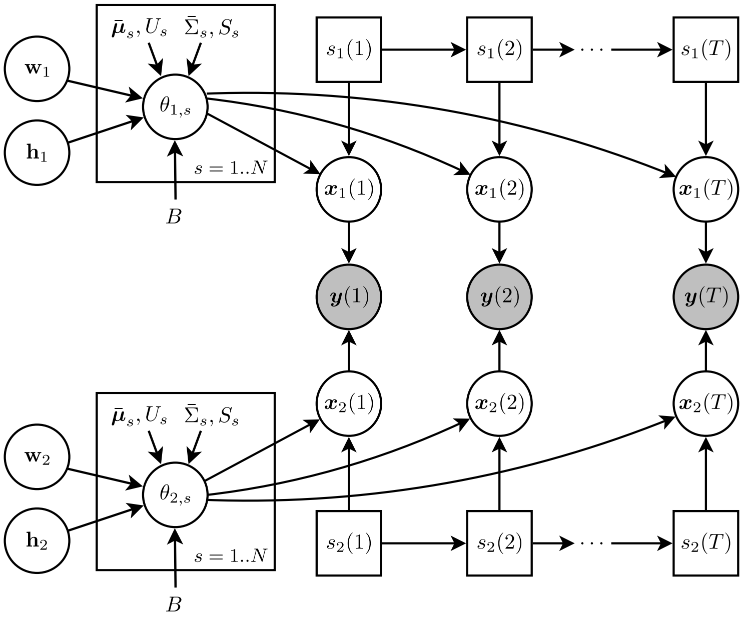


Figure 4.1: Graphical model representation of the proposed mixed signal model. The mixture observations $y(t)$ are explained as the combination of two hidden source signals $x_1(t)$ and $x_2(t)$. Each source signal is modeled by a separate speaker-adapted hidden Markov model, that is derived from the speaker subspace model described in section 3.4. $\theta_{i,s}$ denotes the adapted model parameters for state s in source i derived from the speaker parameters $\mathbf{w}_i$ and channel parameters $\mathbf{h}_i$.

The right hand terms of equation (4.4) very closely resemble the HMM likelihood given in equation (3.3). So the overall likelihood can be thought of as the product of two HMMs coupled by the observed mixed signal $y(t)$. In fact, the model is equivalent to an HMM whose state space is comprised of all possible combinations of s_1 and s_2. This is made clear by marginalizing the latent source signals x_1 and x_2 out of the distribution:

$$\begin{aligned} &P\big(y(1..T),\, s_1(1..T),\, s_2(1..T) \,|\, \mathbf{w}_1,\, \mathbf{w}_2\big) \\ &\qquad = \prod_t P\big(s_1(t) \,|\, s_1(t{-}1)\big) P\big(s_2(t) \,|\, s_2(t{-}1)\big) P\big(y(t) \,|\, s_1(t),\, s_2(t),\, \mathbf{w}_1,\, \mathbf{w}_2\big) \end{aligned} \tag{4.7}$$

where

$$P\big(y \,|\, s_1,\, s_2,\, \mathbf{w}_1,\, \mathbf{w}_2\big) = \int_{x_1}\int_{x_2} P\big(x_1 \,|\, s_1,\, \mathbf{w}_1\big)\; P\big(x_2 \,|\, s_2,\, \mathbf{w}_2\big)\; P\big(y \,|\, x_1,\, x_2\big) \tag{4.8}$$

This is analogous to the HMM likelihood in equation (3.3) except the observations depend on *two* state variables whose temporal dynamics are independent. Therefore, given predefined source models, the only additional detail needed to model a monaural mixture is the interaction distribution of equation (4.8), $P(\boldsymbol{y} \mid \boldsymbol{x}_1, \boldsymbol{x}_2)$.

The construction of the emission distribution depends on the exact relationship between the mixture $\boldsymbol{y}(t)$ and the underlying source signals $\boldsymbol{x}_i(t)$. As shown in equation (4.2) the exact relationship of signals in the log-spectral domain is highly nonlinear and complex to model. Kristjansson [56] describes the Algonquin method for modeling this interaction accurately, but it comes at significant computational cost. Instead we utilize the "max" model [71, 108, 92] which relies on the sparsity of audio sources in the short-time Fourier transform representation. The key assumption here that is not made by Algonquin is that the sources will not overlap. So, in most cases, wherever there is significant energy in one signal, the other will be close to zero. Since only one source makes a significant contribution to any one point in time-frequency, the mixed signal will essentially be equal to the dominant source at each point. This motivates the approximation given in equation (4.3). The corresponding interaction likelihood between the two source signals has all of its mass at a single point and can be written as follows:

$$P(\boldsymbol{y} \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = \delta\big(\boldsymbol{y} - \max_i\{\boldsymbol{x}_i\}\big) \tag{4.9}$$

where δ is a Dirac delta function. Using this approximation, the emission distribution can be written as follows [108]:

$$\begin{aligned} P(\boldsymbol{y}(t) \mid s_1, s_2) &= \prod_d P(y^d(t) \mid s_1, s_2) && (4.10) \\ P(y^d(t) \mid s_1, s_2) &= \mathcal{N}\big(y^d(t);\, \mu^d_{1,s_1}, \Sigma^{dd}_{1,s_1}\big)\, \mathcal{C}\big(y^d(t);\, \mu^d_{2,s_2}, \Sigma^{dd}_{2,s_2}\big) \\ &\quad + \mathcal{N}\big(y^d(t);\, \mu^d_{2,s_2}, \Sigma^{dd}_{2,s_2}\big)\, \mathcal{C}\big(y^d(t);\, \mu^d_{1,s_1}, \Sigma^{dd}_{1,s_1}\big) && (4.11) \end{aligned}$$

where $\mathcal{C}$ is the Gaussian cumulative distribution function (CDF), x^d denotes the dth dimension of $\boldsymbol{x}$, and $\mu^d_{i,s}$ and $\Sigma^{dd}_{i,s}$ denote the adapted mean and diagonal covariance, respectively, for state s in model i.

Since one source is almost always significantly louder than the other, a further approximation is possible. If $(y^d - \mu^d_1)/\sigma^d_1 \gg (y^d - \mu^d_2)/\sigma^d_2$, where $\sigma^d_i = \sqrt{\Sigma^{dd}_i}$ is the standard deviation in subband d, then $\mathcal{C}(\mu^d_2) \gg \mathcal{C}(\mu^d_1)$ and the first term of equation (4.11) will dominate and vice versa. The CDF terms effectively behave like soft masks selecting which model dominates. This is illustrated in figure 4.2. By assuming that these masks are binary, the CDFs do not have to be evaluated, saving computation. We can then rewrite equation (4.11) as a single Gaussian likelihood:

$$P(\boldsymbol{y}(t) \mid s_1, s_2) \approx \mathcal{N}\big(\boldsymbol{y}(t);\, M_1 \boldsymbol{\mu}_{1,s_1} + M_2 \boldsymbol{\mu}_{2,s_2},\, M_1 \Sigma_{1,s_1} + M_2 \Sigma_{2,s_2}\big) \tag{4.12}$$

where M_i behaves as a binary mask that isolates frequency bands dominated by source i. M_1 is a diagonal matrix containing ones for dimensions where model 1 is larger than the other model as described above and zeros elsewhere:

$$M_1^{ij} = \begin{cases} 1, & \text{if } i = j \text{ and } \frac{y^j - \mu^j_{1,s_1}}{\sigma^j_{1,s_1}} > \frac{y^j - \mu^j_{2,s_2}}{\sigma^j_{2,s_2}} \\ 0, & \text{otherwise} \end{cases} \tag{4.13}$$

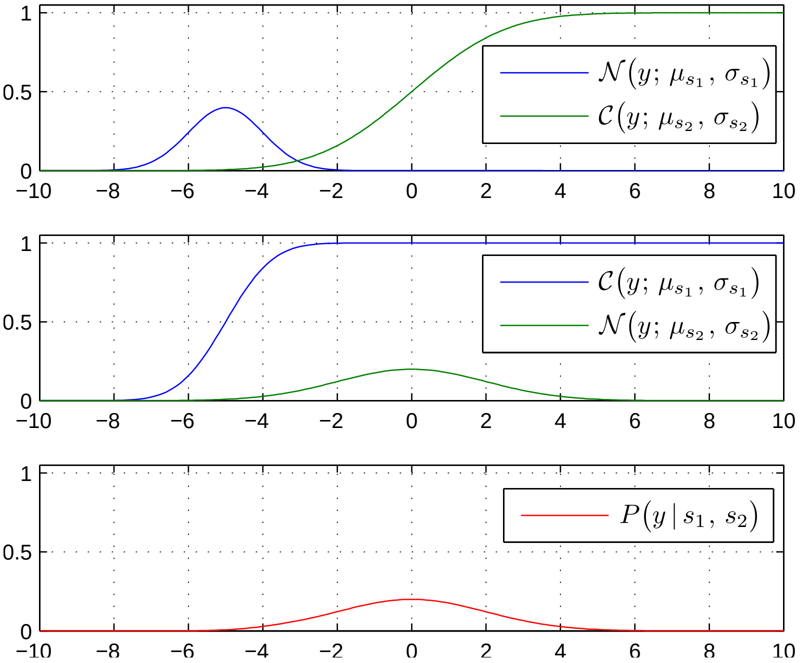


Figure 4.2: Illustration of the max model in equation (4.11). The CDF of the Gaussian with the larger mean (green) effectively cancels the likelihood under the other model (top pane). Similarly the CDF of the smaller Gaussian (blue) passes the likelihood through (middle pane). The overall joint likelihood is therefore determined by the Gaussian with the larger mean (bottom pane).

Similarly, $M_2 = I - M_1$.

Because of the partitioning of the observation by the binary masks in equation (4.12), the two sources decouple given the binary mask. Equation (4.12) can therefore be rewritten as follows:

$$P(\boldsymbol{y}(t)\,|\,s_1,\, s_2) \approx P_{M_1}(\boldsymbol{y}(t)\,|\,s_1)\; P_{M_2}(\boldsymbol{y}(t)\,|\,s_2) \tag{4.14}$$

where

$$P_{M_i}(\boldsymbol{y}(t)\,|\,s_i) = \prod_d M_i^{dd} \mathcal{N}(y^d(t);\, \mu^d_{i,s_i},\, \sigma^d_{i,s_i}) \tag{4.15}$$

4.1.2 Signal separation

Given the mixed signal and the speaker specific parameters $\mathbf{w}_1$ and $\mathbf{w}_2$, the model described above can be used to separate the signals corresponding to each model. We will skip over the task of learning these parameters for now and return to it in section 4.2, where we will describe a variety of such learning algorithms. In this section it is assumed that $\mathbf{w}_i$ are already known.

Reconstructing the signal for source i essentially amounts to finding the expected value of $\boldsymbol{x}_i(1..T)$ given the mixed signal and adaptation parameters: $E\big(\boldsymbol{x}_i(1..T)\,|\,\boldsymbol{y}(1..T),\,\mathbf{w}_1,\,\mathbf{w}_2\big)$. Computing this expectation is not trivial because of the other hidden variables in the model: the HMM state sequences $s_1(1..T)$ and $s_2(1..T)$ and the other underlying source signal $\boldsymbol{x}_j(1..T)$. Marginalizing over these hidden variables before taking the expectation results in the minimum mean square error (MMSE) estimate for each frame of source i:

$$\hat{x}_{i_{MMSE}}(t) = E\big(\boldsymbol{x}_i(t)\,|\,\boldsymbol{y}(1..T)\big) = \sum_{s_1}\sum_{s_2} \gamma_{s_1,s_2}(t)\, E\big(\boldsymbol{x}_i(t)\,|\,\boldsymbol{y}(t),\,s_1,\,s_2\big) \tag{4.16}$$

where $\gamma_{s_1,s_2}(t) = P\big(s_1(t),\,s_2(t)\,|\,\boldsymbol{y}(1..T)\big)$ is the posterior probability of source 1 being in state s_1 and source 2 being in state s_2 at time t given the entire observed sequence. This can be computed using the forward-backward algorithm for factorial HMMs [78, 35].

Given a particular state combination, the expectation under the max model from equation (4.11) at time t can be written as follows [71, 41]:

$$E\big(x_1^d\,|\,y^d,\,s_1,\,s_2\big) = p\,y^d + (1-p)\left(\mu_{1,s_1}^d - \sigma_{1,s_1}^d\,\frac{\mathcal{N}\big(y^d;\,\mu_{1,s_1}^d,\,\sigma_{1,s_1}^d\big)}{\mathcal{C}\big(y^d;\,\mu_{1,s_1}^d,\,\sigma_{1,s_1}^d\big)}\right) \tag{4.17}$$

$$p = \frac{\mathcal{N}\big(y^d;\,\mu_{1,s_1}^d,\,\sigma_{1,s_1}^d\big)\;\mathcal{C}\big(y^d;\,\mu_{2,s_2}^d,\,\sigma_{2,s_2}^d\big)}{P\big(y^d\,|\,s_1,\,s_2\big)} \tag{4.18}$$

and similarly for x_2^d. As noted earlier, the CDF computations needed above can be quite expensive. Under the max model from equation (4.12) this can be simplified as follows:

$$E\big(x_1^d\,|\,y^d,\,s_1,\,s_2\big) \approx \begin{cases} y^d & \text{if } \frac{y^d-\mu_{1,s_1}^d}{\sigma_{1,s_1}^d} > \frac{y^d-\mu_{2,s_2}^d}{\sigma_{2,s_2}^d} \\ \mu_{1,s_1}^d & \text{otherwise} \end{cases} \tag{4.19}$$

Because the mixing model assumes that there is little overlap between the source signals, equation (4.19) simply assigns the observed frequency bin to the dominant source and uses the model mean wherever the source is masked.

As an alternative, the maximum a posteriori (MAP) reconstruction can be found using only the most likely state path through the factorial HMM. This can be computed efficiently using the FHMM Viterbi algorithm which is more efficient to compute than the full posterior distribution needed by equation (4.16). This has the additional advantage of identifying the path that best explains the observations and is not distracted by competing hypotheses with lower likelihood, and is the reconstruction most commonly used in the literature [91, 88, 83].

The MAP reconstruction can be written as follows:

$$\hat{x}_{i_{MAP}}(t) = E\big(\boldsymbol{x}_i(t)\,|\,\boldsymbol{y}(t),\,\hat{s}_1(t),\,\hat{s}_2(t)\big) \tag{4.20}$$

where $\hat{s}_1(t)$ and $\hat{s}_2(t)$ denote the state indices at time t on the Viterbi path for sources 1 and 2, respectively:

$$\hat{s}_1(1..T),\ \hat{s}_2(1..T) = \operatorname*{argmax}_{s_1(1..T),s_2(1..T)} P\big(s_1(1..T),\ s_2(1..T)\,|\,\boldsymbol{y}(1..T)\big) \tag{4.21}$$

Computing the full distribution over all possible state sequences is typically intractable. However, $\hat{s}_i(1..T)$ can be computed using a simple dynamic programming recursion similar to that used in the forward-backward algorithm:

$$\begin{aligned}&P_{MAP}\big(s_1(t),\ s_2(t)\,|\,\boldsymbol{y}(1..t)\big)\\ &\quad= \max_{s_1(t\text{-}1)} P\big(s_1(t)\,|\,s_1(t-1)\big) \max_{s_2(t\text{-}1)} P\big(s_2(t)\,|\,s_2(t-1)\big) P\big(s_1(t-1),\ s_2(t-1)\,|\,\boldsymbol{y}(1..t-1)\big)\end{aligned} \tag{4.22}$$

This recursion defines the maximum likelihood path through the first t observations that end in $s_1(t)$ and $s_2(t)$. By running this recurrence through to the final frame T and by keeping track of the states that maximize equation (4.22) at each time step ("backtracking") it is possible to find the state sequence efficiently.

Despite this speed-up, computing $P_{MAP}\big(s_1,\ s_2\,|\,\boldsymbol{y}(1..T)\big)$ exactly remains intractable if the number of possible state combinations, $N_1 \times N_2$, is too large. To speed it up we prune the number of active state combinations at each frame to the 200 most likely. Because many state combinations are discarded, the resulting state sequences are no longer guaranteed to correspond to the exact MAP solution. Despite this, however, the separation quality remains acceptable.

The speech models we use operate only on features derived from the STFT magnitudes of the source signals, which do not characterize the full time-domain signals exactly. In some applications, the magnitude estimates can be used directly without reverting to the time-domain, e.g. when the target application is speech recognition, not signal recovery, as in the 2006 Speech Separation Challenge. However, sometimes it is desirable to revert to the time-domain. In order to revert the signal estimates $\hat{\boldsymbol{x}}_i(t)$ to the time-domain, we generally combine the magnitude estimates with the phase of the mixed signal and use the inverse STFT to estimate $\hat{x}_i(t)$. This generally leads to errors due to the mismatch between the mixture phase and the magnitude estimates which use the source model parameters to "fill in the blanks" that were dominated by competing sources. These phase errors can be minimized using the iterative phase estimation algorithm described in [38] (and extended in [90]) which estimates phases based on consistency with the given magnitudes.

Figure 4.3 demonstrates the importance of using speaker specific models for the separation algorithm described in this section. Separation performance is quite poor when using the same speaker independent model for both sources in the mixture because, as described in section 3.2, the speech model does not enforce strong temporal constraints. This is due to ambiguity in the Viterbi path through a factorial HMM composed of identical models [26]. The state sequences can permute between sources whenever the Viterbi path passes through the same state in both models at the same time. Since our models only include basic phonetic constraints, the resulting separated signals can permute between sources whenever the two sources have (nearly) synchronous phone transitions. Such permutations are evident in the middle row of figure 4.3. This permutation problem can be solved using models matched to each source as shown in the bottom row of the figure. In the following

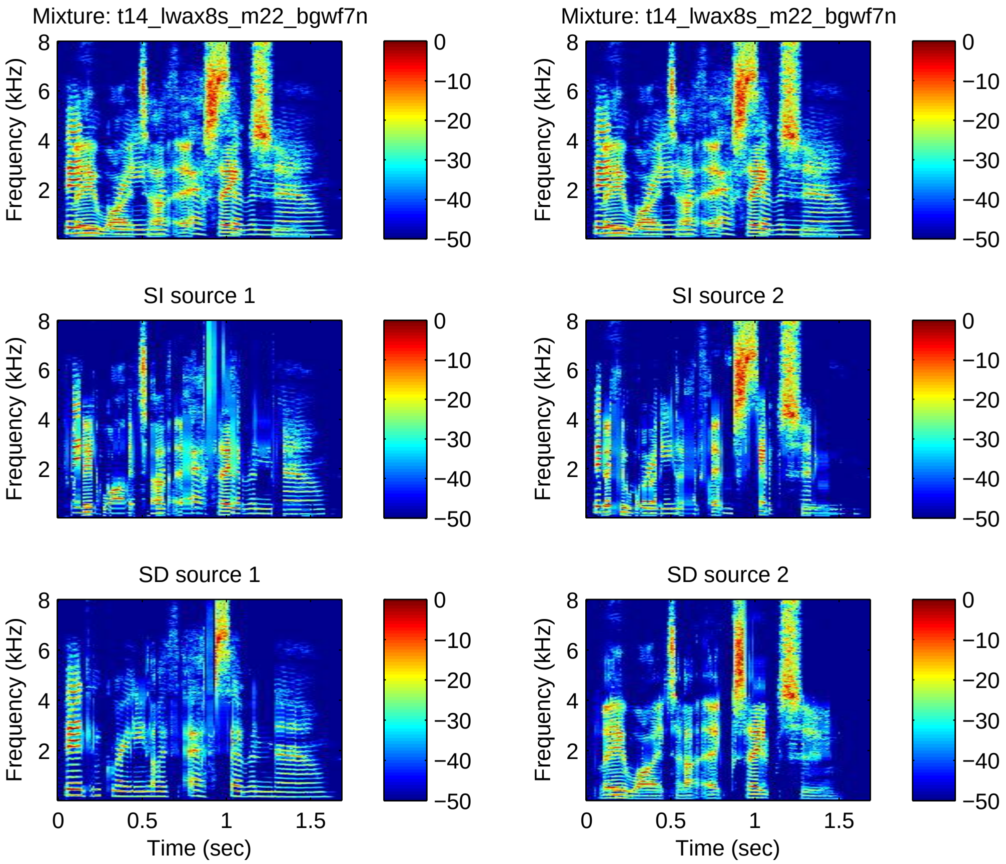


Figure 4.3: Phone permutations found in the Viterbi path through a factorial HMM that models the mixture of two sources. Each source is modeled with the same speaker-independent (SI) models. Neither of the resulting source reconstructions (middle row) is a good match for either of the sources found by separation using speaker-dependent (SD) models (bottom row). Each SI reconstruction contains short regions from both of the SD sources. For example, the final phone in SI source 1 is quite close to that of SD source 1, but the first half second of the signal is closer to SD source 2.

section we discuss methods for learning adaptation parameters for the sources comprising a mixture.

4.2 Learning algorithms

We separate the sources composing a speech mixture in two stages using the model described in the previous section. First, the subspace parameters are estimated for each source in the mixture, producing a set of speaker-adapted models capturing the speaker-dependent statistics of the constituent talkers. Then, given the adapted models, the clean source signals are reconstructed by finding the MAP reconstruction of the signals given the model, as described in section 4.1.2.

The remaining detail to be described is the method for learning the speaker adaptation parameters $\mathbf{w}_i$ and channel responses $\mathbf{h}_i$ from the mixed signal. In this section we describe three different approaches to this problem. First, we describe the speaker identification approach described in Rennie et al. [83] which utilizes a pre-defined set of speaker parameters and selects the subset that best match the observations. If the sources comprising the mixture are known to correspond to entries in this database, this approach can approximate the upper bound performance of using known models. However, in mismatched conditions this does not work as well as finding the optimal point in the continuous eigenvoice space that correspond exactly to the given speakers. We describe two alternate algorithms to learn the adaptation parameters directly from the mixed signal in sections 4.2.2 and 4.2.3.

4.2.1 Model selection

As described in chapter 3, the simplest approach to "adapting" to a set of sources is to simply select the model parameters from a closed set of predefined settings that best fit the mixture, essentially finding the nearest neighbor in model space. Rennie et al. [83] describe an efficient algorithm for doing exactly this as a part of their separation system. In this section we extend their method to work with our channel model described in section 3.4.3.

The approach attempts to identify frames that are dominated by a single source and use these frames to evaluate the set of speaker parameters and identify those that fit best. Given a set of C speaker parameters $\{\mathbf{w}_c\}_{c=1..C}$, we define the corresponding set of model parameters for state s as in the previous section: $\boldsymbol{\mu}_{c,s} = \boldsymbol{\mu}_s(\mathbf{w}_c)$ and optionally $\Sigma_{c,s} = \Sigma_s(\mathbf{w}_c)$. Note that in this section, $\mathbf{w}_c$ only refers to the eigenvoice coefficients and does not include the channel response $\mathbf{h}$. This is because the channel is assumed to be independent of speaker identity and is likely to vary between training and testing data. We will describe estimation of $\mathbf{h}$ given $\mathbf{w}$ in section 4.2.1.1.

Because we are interested in selecting model parameters independent of the channel response, it needs to be integrated out of the model. Rennie et al. [83] used a very constrained channel model consisting of a simple fixed gain which was assumed to come from a quantized set. However, as described in chapter 3, we use a more general model of channel variation. Here, we utilize a simple Gaussian prior on the channel response $\mathbf{h}$: $\mathcal{N}(\mathbf{h};\, \boldsymbol{\nu}_{\mathbf{h}},\, \Xi_{\mathbf{h}})$. The likelihood of observation y being in state s of source c can then be written as follows:

$$P(\boldsymbol{y} \mid c,\, s) = \int_{\mathbf{h}} \mathcal{N}(\boldsymbol{y};\, \boldsymbol{\mu}_{c,s} + B\mathbf{h},\, \Sigma_{c,s})\, \mathcal{N}(\mathbf{h};\, \boldsymbol{\nu}_{\mathbf{h}},\, \Xi_{\mathbf{h}}) \tag{4.23}$$

$$\propto \mathcal{N}(\boldsymbol{y};\, \boldsymbol{\mu}_{c,s},\, \Sigma_{c,s}) \frac{\mathcal{N}(\mathbf{0};\, \boldsymbol{\nu}_{\mathbf{h}},\, \Xi_{\mathbf{h}})}{\mathcal{N}(\mathbf{0};\, \boldsymbol{d},\, D)} \tag{4.24}$$

where

$$D = \left(\Xi_{\mathbf{h}}^{-1} + B^T \Sigma_{c,s}^{-1} B\right)^{-1} \tag{4.25}$$

$$\boldsymbol{d} = D\left(\Xi_{\mathbf{h}}^{-1} \boldsymbol{\nu}_{\mathbf{h}} + B^T \Sigma_{c,s}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_{c,s})\right) \tag{4.26}$$

Alternatively, if the channel is known to be very simple, e.g. the test data is simply offset from the training data by a constant gain as is the case in the Speech Separation Challenge, then a simpler multinomial prior on h can be used. This is described in detail in Rennie et al. [83]. Similarly, if the channel distribution is known to be more complex than a single Gaussian, it is straightforward to extend equation (4.24) to a mixture of Gaussians.

Since the only speaker-dependent parameters in our source models are the Gaussian means and covariances, we ignore the temporal dynamics throughout the model selection process. Instead we reduce the HMM transition matrix to the corresponding steady-state distribution parametrized by π_s and represent the model as a GMM where mixture component s corresponds to the Gaussian in state s of the HMM with prior probability π_s. Once the model parameters have been estimated, we revert to the full HMM distribution to separate the sources. The posterior probability of model c given a particular frame can then be written as follows:

$$P\big(c \,|\, \boldsymbol{y}(t)\big) = \frac{\sum_s \pi_s \, P\big(\boldsymbol{y}(t) \,|\, c,\, s\big)}{\sum_c \sum_s \pi_s \, P\big(\boldsymbol{y}(t) \,|\, c,\, s\big)} \tag{4.27}$$

The final decision is made by pooling all frames dominated by a single source:

$$P\big(c \,|\, \boldsymbol{y}(1..T)\big) = \frac{\sum_{t \in F} P\big(c \,|\, \boldsymbol{y}(t)\big)}{\sum_c \sum_{t \in F} P\big(c \,|\, \boldsymbol{y}(t)\big)} \tag{4.28}$$

where F is the set of all frames for which a single source is more likely than the rest, i.e. $F = \{t \mid \max_c P\big(c \,|\, \boldsymbol{y}(t)\big) > \alpha\}$ where α is some predefined threshold. Only utilizing frames dominated by a single source minimizes the influence of frames that are good matches to multiple source models (i.e. silence or frames containing significant energy from multiple sources). Because of the natural short pauses often present in speech, this approach gives reasonable glimpses of individual speakers in a mixture.

The I models with highest posterior probability as given by equation (4.28) are assumed to correspond to the sources in the mixture. Because no state combinations need to be evaluated, this algorithm is quite fast, essentially requiring the computation of posterior probabilities for each of N states in each of C models for each of T observations. The algorithm has been shown to perform extremely well, accurately identifying 98% of the speakers in the SSC data set [83].

4.2.1.1 Channel adaptation

Given the speaker specific model parameters $\mu_{i,s}$ and $\Sigma_{i,s}$, the corresponding channel response h_i still needs to be determined. In [83], Rennie et al. follow the speaker identification stage described above by a process to estimate the unknown gain applied to the observed signal. They do this using a simple quantized representation of possible gains consistent with the SNRs found in the 2006 Speech Separation Challenge dataset. Once again, we generalize this approach to arbitrary channel filtering, characterized by the channel parameter vector $\mathbf{h}_i$.

$\mathbf{h}_i$ is estimated iteratively, first inferring the state sequences to separate the signals, then updating $\mathbf{h}_i$ to match the source reconstructions $\hat{x}_i(t)$. To minimize necessary computation we utilize the same GMM representation of the speaker model as in the model selection stage. Because the models are generally sufficiently distinct from one another, the lack of temporal constraints does not significantly decrease performance.

In order to efficiently separate the sources, we use the MAXVQ branch and bound inference algorithm described in Roweis [92] to estimate the MAP state combination under the max mixing model. The source reconstructions $\hat{x}_i(t)$ are estimated as described in section 4.1.2.

Given each of the separated signals, the channel parameters are estimated using the MLED expectation maximization (EM) algorithm [58] (which will be described in detail in section 4.2.2.2), except only the channel parameters are updated. The E-step computes $\gamma_s(t)$, the posterior distribution of $\hat{x}_i(t)$ under each model state using the current estimate of the channel parameters. This posterior distribution is then used in the M-step to update the channel parameters.

The expected complete log likelihood of the data under the model can be written as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{h}_i) = -\frac{1}{2} \sum_t \sum_s \gamma_s(t) \Big(\big(\boldsymbol{\mu}_{i,s_i} + B\mathbf{h}_i - \hat{x}_i(t)\big)^T \Sigma_{i,s}^{-1} \big(\boldsymbol{\mu}_{i,s_i} + B\mathbf{h}_i - \hat{x}_i(t)\big) \\ + \big(\mathbf{h}_i - \boldsymbol{\nu}_{\mathbf{h}}\big)^T \Xi_{\mathbf{h}}^{-1} \big(\mathbf{h}_i - \boldsymbol{\nu}_{\mathbf{h}}\big) \Big)\end{aligned} \tag{4.29}$$

Maximizing this objective for source i yields the following update for $\mathbf{h}_i$:

$$\mathbf{h}_i = \left(\sum_t \sum_s \gamma_s(t)\, B^T \Sigma_{i,s}^{-1} B + \Xi_{\mathbf{h}}^{-1} \right)^{-1} \left(\sum_t \sum_s \gamma_s(t)\, B^T \Sigma_{i,s}^{-1} \big(\hat{x}_i(t) - \boldsymbol{\mu}_{i,s_1}\big) + \Xi_{\mathbf{h}}^{-1} \boldsymbol{\nu}_{\mathbf{h}} \right) \tag{4.30}$$

The process is iterated until it converges. Figure 4.4 shows an example of the convergence behavior of this algorithm on a mixture of utterances from a male and female speaker from the Switchboard data set described in section 4.4 using a channel basis set composed of 10 DCT bases. The channel responses for both sources converge after about 10 iterations.

4.2.2 Hierarchical learning

In the previous section we assumed that a set of pre-trained speaker parameters was available, in which case adaptation is simply a matter of selecting the parameters that best fit the observations. Such a database is not generally available, but even if one is, there is no guarantee that the optimal parameters for a particular source signal are contained in

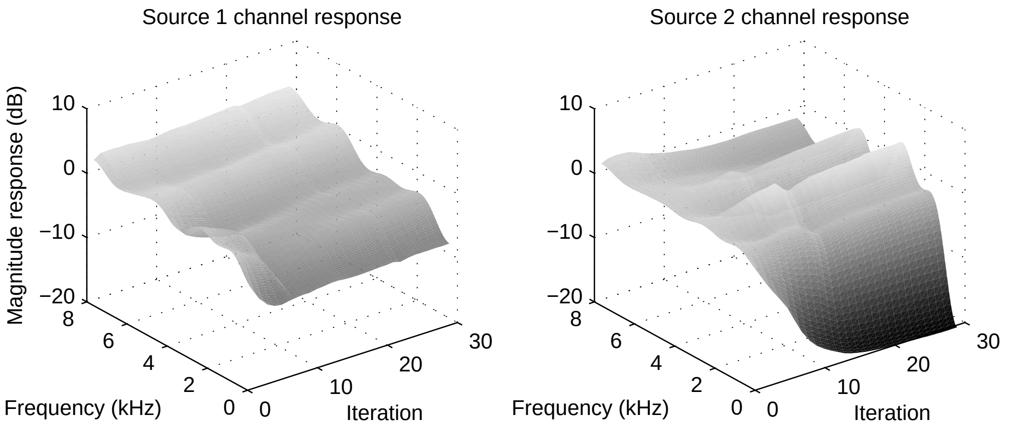


Figure 4.4: Example of the convergence behavior of the channel adaptation algorithm.

that set. In this section we describe a method for learning adaptation parameters directly from the mixture.

Kuhn et al. [58] describe an EM algorithm for learning eigenvoice parameters **w** given a clean source signal. This algorithm cannot be directly applied to a mixture of multiple sources. Instead, we use a hierarchical approach which iteratively separates the sources and then uses the aforementioned EM algorithm to adapt each source to its reconstruction. An outline of this approach is given below:

1. Obtain initial model estimates for each source
2. Separate signals as described in section 4.1.2
3. Update model parameters
4. Repeat 2 – 3 until convergence

The intuition behind the iterative approach is that each of the reconstructed source estimates will resemble one source more closely than the other (i.e. more than half of it will match one of the sources), even if the initial separation is quite poor. As a result, the model parameters inferred from these estimates will also be a better match to one source than to the other. This in turn should improve the separation in the next iteration.

Initially the dynamic constraints in the model partially make up for the lack of source-dependent feature constraints. But the reconstructions are still quite prone to permutations between sources. The permutations tend to get corrected as the algorithm iterates because the adaptation allows the models to better approximate the speaker characteristics. Unfortunately the algorithm takes many iterations to converge.

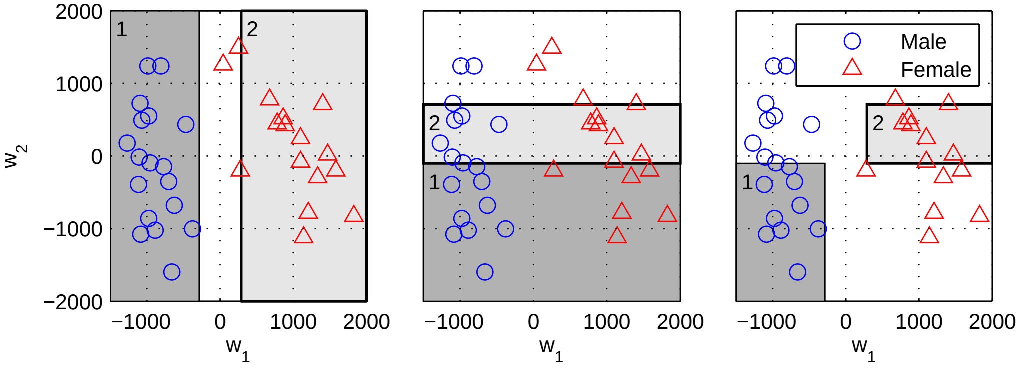


Figure 4.5: Quantization of eigenvoice coefficients w_1 and w_2 across all GRID training speakers. In the first two panes, the regions labeled 1 and 2 denote the settings for the corresponding eigenvoice coefficient that best fit the observations. The rightmost pane shows the best joint setting of w_1 and w_2.

4.2.2.1 Initialization

As with many iterative algorithms, this method is vulnerable to becoming stuck in local optima. Good initialization is crucial to finding good solutions quickly. One approach would be to use the speaker identification algorithm described in section 4.2.1 to search through a set of adaptation parameter vectors corresponding to the speakers in the training set, and use them to initialize the learning algorithm with the best matching prototypes.

However this may not work well on sources that are not in the training set. We are interested in producing a more general algorithm that does not require as much prior knowledge. Toward that end we propose an alternative scheme for initialization. We begin by projecting the mixed signal onto the eigenvoice bases to set the parameters for both sources (see section 4.2.2.2). Obviously these parameters will not likely be a good match to either isolated source and, as described earlier, using the same model for both sources will lead to poor performance.

Further steps are taken to differentiate the two speakers using the speaker identification algorithm. We note that by design the eigenvoice dimensions are decorrelated, which allows us to treat each of them independently. The idea is to build an approximation of **w** for each source from the bottom up, beginning from w_1 and adding consecutive weights.

During training we learn prototype settings for each weight w_j by coarsely quantizing the corresponding weights of the training speakers to three quantization levels using the Lloyd-Max algorithm [63]. This allows w_j for any given speaker to be approximated by one of the quantized values $\{\hat{w}_j^1, \hat{w}_j^2, \hat{w}_j^3\}$. The first two panels of figure 4.5 show example quantization levels for w_1 and w_2.

Given the mixed signal, we can approximate the eigenvoice adaptation parameters for each speaker, $\mathbf{w}_1$ and $\mathbf{w}_2$, using the following bottom up construction:

- Initialize $\mathbf{w}_1$ and $\mathbf{w}_2$ to zero, i.e. set $\boldsymbol{\mu}(\mathbf{w}_1)$ and $\boldsymbol{\mu}(\mathbf{w}_2)$ to $\bar{\boldsymbol{\mu}}$.
- For each speaker i and eigenvoice dimension j:
  - Use the quantized weights to construct prototype models $\{\boldsymbol{\mu}^k(\mathbf{w}_i)\}_{1 \leq k \leq 3}$ where $\boldsymbol{\mu}^k(\mathbf{w}_i) = \boldsymbol{\mu}(\mathbf{w}_i) + \hat{w}_j^k \hat{\boldsymbol{\mu}}_j$.
  - Use the Iroquois speaker identification algorithm (see section 4.2.1) to select the most likely prototype model given the mixed signal and update $\mathbf{w}_i$ and $\boldsymbol{\mu}(\mathbf{w}_i)$ accordingly.

An example of this process is shown in figure 4.5. The first and second panels show the quantization levels of eigenvoice dimensions 1 and 2 respectively. The shaded regions show the prototypes chosen for speaker 1 (dark gray) and speaker 2 (light gray). Finally, the rightmost panel shows the joint selection of w_1 and w_2 for both speakers.

This is only done for the 3 eigenvoice dimensions with the highest variance. The remaining parameters are the same for both sources, set to match the mixture. This technique is not very accurate, but in most cases it suffices to differentiate the two sources. It works best at differentiating between male and female speakers because the eigenvoice dimension with the most variance tends to be highly correlated with speaker gender. When the sources are very similar this procedure does not guarantee that they will not have identical initialization. In this case, the algorithm relies on the initial separation to break the symmetry.

4.2.2.2 Eigenvoice parameter estimation

Given source specific weights $\mathbf{w}_i$, the sources are separated using the factorial HMM Viterbi algorithm as described in section 4.1.2. Then the speaker models are updated to better match the source estimates $\hat{x}_1(t)$ and $\hat{x}_2(t)$. This amounts to projecting the reconstructed source signal onto the eigenvoice (and channel) bases U. The model parameters $\mathbf{w}_i$ are estimated iteratively using an extension of the maximum likelihood eigen-decomposition (MLED) EM algorithm described in Kuhn et al. [58] to incorporate a prior distribution over the adaptation parameters. This was first described by Huang et al. [43]. As described in section 3.4.3, for the purposes of deriving the learning algorithm, the channel parameters for source i, $\mathbf{h}_i$, can be subsumed into $\mathbf{w}_i$.

The E-step of the algorithm involves computing $\gamma_s(t) = P\big(s(t) \mid \hat{x}(t)\big)$, the posterior probability of the source occupying state s at time t given the observations $\hat{x}_i(1..t)$ using the HMM forward-backward algorithm [78]. It should be emphasized that in contrast to the algorithm described in section 4.2.3, the E-step here uses the source reconstructions, and not the mixed signal, thus eliminating the need for taking all possible state combinations into account. This task was offloaded to the separation algorithm.

Given these posteriors, the M-step maximizes the likelihood of the observed sequence $\hat{x}_i$ under the model. Once again, this is done by maximizing the expected complete log likelihood $\mathcal{L}(\mathbf{w}_i) = E\big(\log P\big(\boldsymbol{y}(t),\, \mathbf{w}_i \,|\, s\big)\big)$:

$$\mathcal{L}(\mathbf{w}_i) = -\frac{1}{2}\sum_t\sum_s \gamma_s(t)\,\left(\hat{\boldsymbol{x}}_i(t) - \boldsymbol{\mu}_s(\mathbf{w}_i)\right)^T \bar{\Sigma}_s^{-1}\,\left(\hat{\boldsymbol{x}}_i(t) - \boldsymbol{\mu}_s(\mathbf{w}_i)\right) \\ - \frac{1}{2}\big(\mathbf{w}_i - \boldsymbol{\nu}\big)^T \Xi^{-1}\big(\mathbf{w}_i - \boldsymbol{\nu}\big) \quad (4.31)$$

Maximizing this objective function yields the following update for $\mathbf{w}_i$:

$$\mathbf{w}_i = \left(\sum_{t,s} \gamma_s(t)\, U_s^T\, \bar{\Sigma}_s^{-1}\, U_s + \Xi^{-1}\right)^{-1} \left(\sum_{t,s} \gamma_s(t)\, U_s^T\, \bar{\Sigma}_s^{-1}\, (\hat{\boldsymbol{x}}_i(t) - \bar{\boldsymbol{\mu}}_s) + \Xi^{-1}\boldsymbol{\nu}\right) \quad (4.32)$$

This EM algorithm is applied in turn to each source estimate $\hat{\boldsymbol{x}}_i$ until convergence to infer $\mathbf{w}_i$ and $\mathbf{h}_i$. The updated source models are then used to re-estimate the source signals $\hat{\boldsymbol{x}}_i$, and the algorithm iterates.

Figure 4.6 shows an example of the convergence behavior of this algorithm on a mixture of utterances from a male and female speaker. The algorithm has mostly converged after about 15 iterations.

Figure 4.7 gives an example of the algorithm outlined in this section. The initial separation does a reasonable job at isolating the target, but it make some errors. For example, the phone at $t = 1\ s$ is initially mostly attributed to the masking source. The reconstruction improves with subsequent iterations, getting quite close to the reconstruction based on SD models (bottom panel) by the fifth iteration.

4.2.3 Variational learning

In the previous section, we described an ad-hoc learning algorithm for $\mathbf{w}_i$ based on well known algorithms for model-based separation and eigenvoice parameter estimation from clean signals. In this section we derive a similar learning algorithm that is better motivated theoretically. The standard approach for estimating the value of latent parameters in a model such as that depicted in figure 4.1 is to use the expectation maximization (EM) algorithm [21]. The parameters are estimated iteratively by alternately using the current estimate of the parameters to infer the posterior distribution over the remaining latent variables (E-step) and then using the new distribution to re-estimate the parameters in question (M-step). In general, this approach can find the maximum likelihood estimates of the latent parameters.

In the context of the model in figure 4.1, it is possible to derive an EM algorithm to estimate $\mathbf{w}_i$ and $\mathbf{h}_i$ similar to the factorial HMM EM training algorithm described in [35], but the exact computation of the posterior probabilities in the E-step is intractable due to the combinatorial nature of the state space. I.e. if each of the I speaker models contains N states, the statistics needed by the full EM algorithm must take into account all possible state combinations from all speaker models, leading to an equivalent state space containing N^I states. Instead, we derive an approximate E-step with a significantly reduced complexity

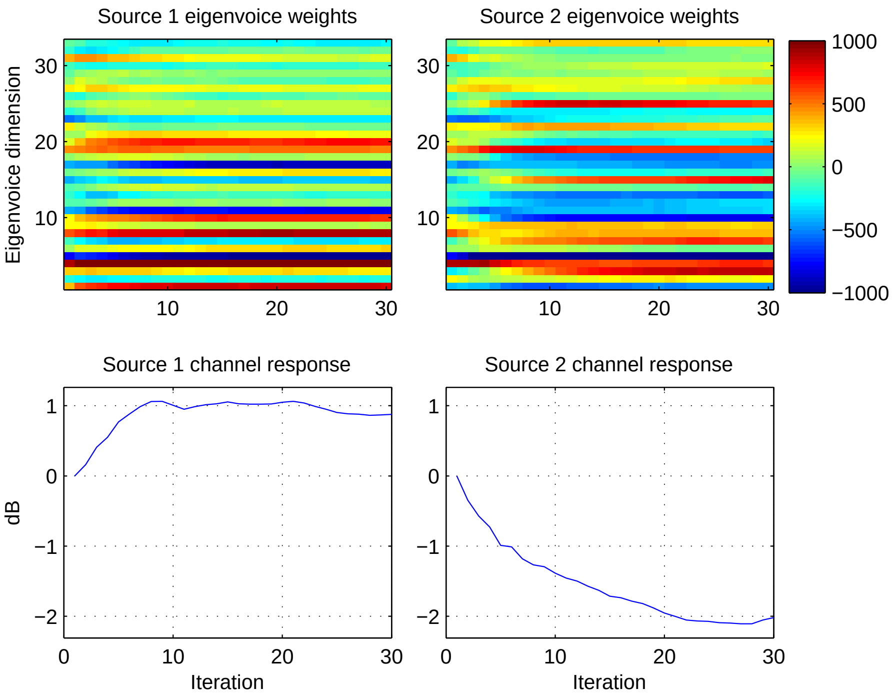


Figure 4.6: Example of the convergence behavior of the hierarchical adaptation algorithm on a mixture from the Speech Separation Challenge data set (see section 4.3). The top panes plot the adaptation parameters $\mathbf{w}_1$ and $\mathbf{w}_2$ at each iteration, and the bottom panes show the channel response $B\,\mathbf{h}_1$ and $B\,\mathbf{h}_2$ at each iteration. Note that this example used a single channel basis vector, corresponding to a constant gain across all frequency bands.

based on an approximation to the posterior distribution. This approach was described in [119].

The full posterior distribution over all possible state combinations given the observations can be written as follows:

$$P\big(s_1(1..T), s_2(1..T) \mid \boldsymbol{y}(1..T)\big) \propto \prod_t P\big(\boldsymbol{y}(t) \mid s_1(t), s_2(t)\big) \prod_i P\big(s_i(t) \mid s_i(t-1)\big) \quad (4.33)$$

This is based on equation (4.4) where the latent source signals $\boldsymbol{x}_i$ have been marginalized out.

Following the structured variational approximation described in [35], we lower bound the joint distribution in equation (4.33) with an approximate distribution in which the HMM

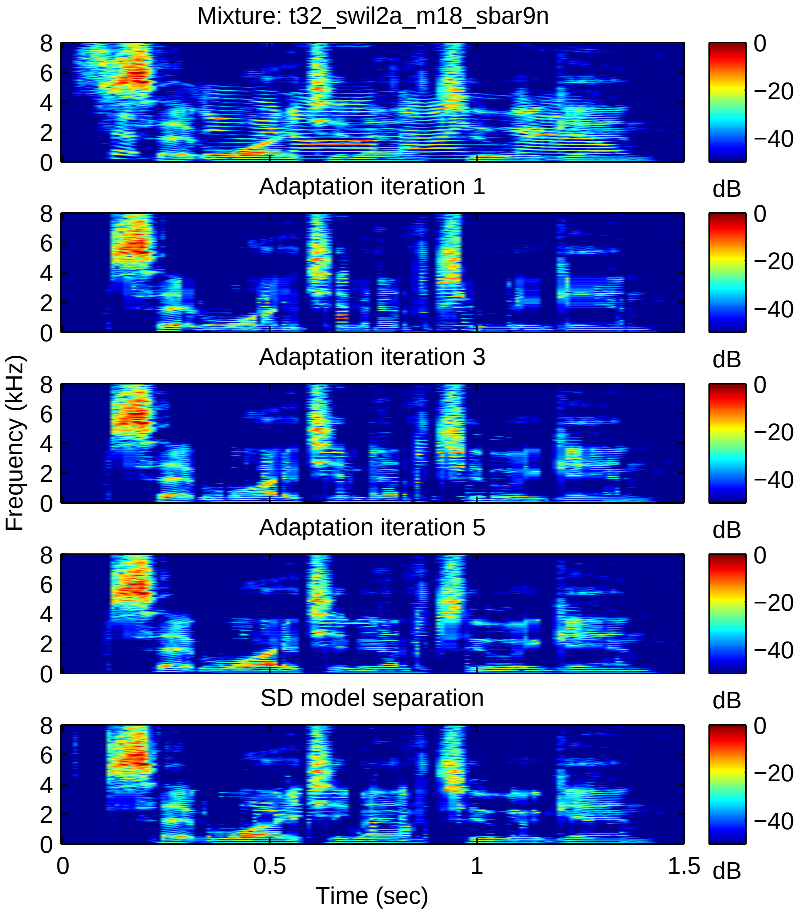
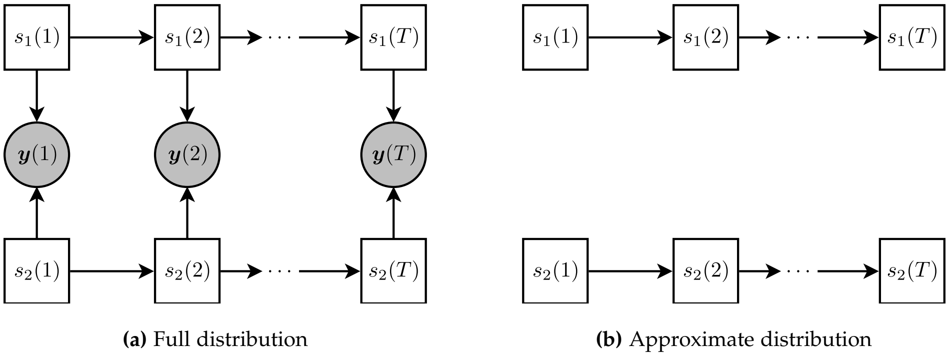


Figure 4.7: Example separation using speaker-adapted models. The top plot shows the spectrogram of a mixture of female and male speakers. The middle three show the reconstructed target signal ("set white in l 2 again") from the adapted models after iterations 1, 3, and 5. The bottom plot shows the result of separation using the speaker-dependent model for target speaker.

chains for each speaker are assumed to be independent:

$$Q\big(s_1(1..T),\, s_2(1..T) \,|\, \boldsymbol{y}(1..T)\big) \propto \prod_i Q_i\big(\boldsymbol{y}(1..T),\, s_i(1..T)\big) \tag{4.34}$$

This is in contrast to the true distribution in which the two models are explicitly coupled by the likelihood term $P\big(\boldsymbol{y}(t) \,|\, s_1, s_2\big)$. Instead, a looser coupling is incorporated into Q_i in



(a) Full distribution

(b) Approximate distribution

Figure 4.8: Graphical model representation of the full distribution from equation (4.33) (a) and variational approximation from equation (4.35) (b) where the underlying source signals have been marginalized out. In (b) the coupling between the two HMM chains by the observations $\boldsymbol{y}$ has been removed.

the form of variational parameters $\rho_{i,s_i}(t)$:

$$Q_i\big(\boldsymbol{y}(1..T),\, s_i(1..T)\big) = \prod_t \rho_{i,s_i}(t)\, P\big(s_i(t) \mid s_i(t-1)\big) \tag{4.35}$$

This closely resembles the HMM likelihood in equation (3.3) with the variational parameters replacing the observation likelihood. A graphical model depiction comparing the exact distribution to this approximation is shown in figure 4.8.

An outline of the overall variational learning algorithm is described below. Details are given in the following sections.

- Initialize w_1 and w_2 as in the hierarchical algorithm.
- E-step: Iteratively infer the posterior distribution over state combinations of both speaker models.
  1. For each model: compute the state occupancy probabilities, $\gamma_{i,s}(t)$, using the HMM forward-backward algorithm with the observation likelihoods replaced by the corresponding variational parameters.
  2. Compute the variational parameters $\rho_{i,s}(t)$ based on $\gamma_{i,s}(t)$.
  3. Iterate 1 – 2 until convergence.
- M-step: Update the model parameters using the posteriors computed in the E-step.

4.2.3.1 E-step

For each iteration of the inner loop in the E-step, new state occupancies $\gamma_{i,s}(t)$ are calculated by evaluating the HMM forward-backward algorithm [78], taking the current variational parameters as the observation likelihoods as described in [35]. Then, the variational parameters are updated by minimizing the Kullback-Leibler divergence between the approximation Q and the full distribution P. I.e. we seek to minimize:

$$\begin{aligned} &KL(Q||P) \\ &\quad = \sum_{t,s_1,s_2} \gamma_{1,s_1}(t)\, \gamma_{2,s_2}(t) \Big(\log P\big(\boldsymbol{y}(t) \,|\, s_1,\, s_2\big) - \log \rho_{1,s_1}(t) - \log \rho_{2,s_2}(t) \Big) + c \end{aligned} \tag{4.36}$$

where c is a constant that is independent of the variational parameters. This implies the following updates for the variational parameters:

$$\log \rho_{1,s_1}(t) = \sum_{s_2} \gamma_{2,s_2}(t)\, \log P\big(\boldsymbol{y}(t) \,|\, s_1,\, s_2\big) \tag{4.37}$$

$$\log \rho_{2,s_2}(t) = \sum_{s_1} \gamma_{1,s_1}(t)\, \log P\big(\boldsymbol{y}(t) \,|\, s_1,\, s_2\big) \tag{4.38}$$

The variational parameter for model i is found simply by marginalizing the joint likelihood over the other models, essentially holding one chain constant while updating the other. Using the likelihood from equation (4.12), equation (4.37) can be rewritten as follows:

$$\begin{aligned} \log \rho_{1,s_1}(t) = \sum_d \Big(&\log \mathcal{N}\big(y^d(t);\, \mu^d_{1,s_1},\, \Sigma^{dd}_{1,s_1}\big) \sum_{s_2} \gamma_{2,s_2}(t) M_1^{dd} \\ &+ \sum_{s_2} \gamma_{2,s_2}(t) M_2^{dd} \log \mathcal{N}\big(y^d(t);\, \mu^d_{2,s_2},\, \Sigma^{dd}_{2,s_2}\big)\Big) \end{aligned} \tag{4.39}$$

where M_i depends on both s_1 and s_2 as defined in equation (4.13). The equation for $\rho_{2,s_2}(t)$ can be derived analogously. The new values for $\rho_{i,s}(t)$ are then used in the forward-backward algorithm to calculate new $\gamma_{i,s}(t)$ values, and this loop continues until convergence.

Unfortunately, s_1 and s_2 do not fully decouple in equation (4.39), so the run time is not guaranteed to be linear in the number of sources. However, because $\gamma_{i,s}(t)$ is generally quite sparse (i.e. very few states per frame have significant probability mass), it is still fast to compute the variational parameters. In order to retain this performance advantage in the first iteration, the initial $\gamma_{i,s}(t)$ must be sparse. Initializing the posteriors to be uniformly random results in the initial computation of equation (4.39) being extremely slow as all possible state combinations need to be evaluated. In our experience, initializing $\gamma_{i,s}(t)$ in the first EM iteration based on a computation of equation (4.33) using aggressive pruning of the forward-backward lattice has worked reasonably well. For all remaining iterations the approximate E-step described above is used.

4.2.3.2 M-step

Given the posterior distribution over the hidden state sequences, the speaker subspace parameters $\mathbf{w}_i$ can be updated by maximizing the expected log likelihood of the model:

$$\mathcal{L}(\mathbf{w}_1, \mathbf{w}_2) = \sum_{t,s_1,s_2} \gamma_{1,s_1}(t)\, \gamma_{2,s_2}(t)\, \log P\big(\boldsymbol{y}(t),\, \mathbf{w}_1,\, \mathbf{w}_2 \,|\, s_1,\, s_2\big) + k \tag{4.40}$$

where k is a constant that is independent of the parameters w_1 and w_2. As shown in [43], this objective function is not convex when both the Gaussian means and covariances depend on the subspace parameters being optimized. Instead, as described in section 3.4.2, we derive an update based only on the mean statistics and rely on the correlation between the mean and covariance parameters implicit in the subspace bases to adapt the model covariances. The simplified objective can be written as follows:

$$\mathcal{L}(\mathbf{w}_1, \mathbf{w}_2) = -\frac{1}{2} \sum_{t,s_1,s_2} \gamma_{1,s_1}(t)\, \gamma_{2,s_2}(t)\, \big(\boldsymbol{y}(t) - \boldsymbol{\mu}_{s_1 s_2}\big)^T \Sigma_{s_1 s_2}^{-1} \big(\boldsymbol{y}(t) - \boldsymbol{\mu}_{s_1 s_2}\big) \\ - \frac{1}{2}\big(\mathbf{w}_1 - \boldsymbol{\nu}\big)^T \Xi^{-1} \big(\mathbf{w}_1 - \boldsymbol{\nu}\big) - \frac{1}{2}\big(\mathbf{w}_2 - \boldsymbol{\nu}\big)^T \Xi^{-1} \big(\mathbf{w}_2 - \boldsymbol{\nu}\big) \quad (4.41)$$

where

$$\boldsymbol{\mu}_{s_1 s_2} = M_1\, \boldsymbol{\mu}_{s_1}(\mathbf{w}_1) + M_2\, \boldsymbol{\mu}_{s_2}(\mathbf{w}_2) \quad (4.42)$$

$$\Sigma_{s_1 s_2} = M_1\, \bar{\Sigma}_{s_1} + M_2\, \bar{\Sigma}_{s_2} \quad (4.43)$$

and $\boldsymbol{\nu}$ and Ξ denote the parameters of the prior distribution from equation (4.6).

Unfortunately, this objective function is non-differentiable due to the step function (i.e. the binary masks M_i) inherent in the max approximation in equation (4.12). This makes it difficult to maximize exactly. To find an approximate solution we hold M_i constant in the optimization, i.e. we assume that the resulting update will not change which source dominates the mixture at any time-frequency point. Because of this, the log likelihood is not guaranteed to increase. However, in practice we have found that it works quite well.

The binary mask construction has the advantage of decoupling the two sources, allowing them to be updated independently. The decoupled objective function for source i can be written as follows:

$$\mathcal{L}(\mathbf{w}_i) = -\frac{1}{2} \sum_{t,s} \gamma_{i,s}(t)\, \big(\boldsymbol{y}(t) - \boldsymbol{\mu}_s(\mathbf{w}_i)\big)^T M_i \bar{\Sigma}_s^{-1} \big(\boldsymbol{y}(t) - \boldsymbol{\mu}_s(\mathbf{w}_i)\big) \\ - \frac{1}{2}\big(\mathbf{w}_i - \boldsymbol{\nu}\big)^T \Xi^{-1} \big(\mathbf{w}_i - \boldsymbol{\nu}\big) \quad (4.44)$$

This is maximized by solving the following equation for $\mathbf{w}_1$ and $\mathbf{w}_2$:

$$\left(\sum_{t,s_1,s_2} \gamma_{1,s_1}(t)\, \gamma_{2,s_2}(t)\, U_s^T M_i\, \bar{\Sigma}_{s_i}^{-1} U_s + \Xi^{-1} \right) \mathbf{w}_i \\ = \sum_{t,s_1,s_2} \gamma_{1,s_1}(t)\, \gamma_{2,s_2}(t)\, U_s^T M_i\, \bar{\Sigma}_{s_i}^{-1} \big(\boldsymbol{y}(t) - \bar{\boldsymbol{\mu}}_{s_i}\big) + \Xi^{-1} \boldsymbol{\nu} \quad (4.45)$$

The updates are quite similar to the clean signal eigenvoice updates derived in section 4.2.2.2, except for the binary masks M_i which partition the observations into regions dominated by a single talker, causing the algorithm to ignore interference-dominated time-frequency regions when updating the parameters for a particular talker.

Figures 4.9 and 4.10 show an example of the convergence behavior of this algorithm on the same mixture as figure 4.6. Figure 4.9 shows the log likelihood. after each iteration,

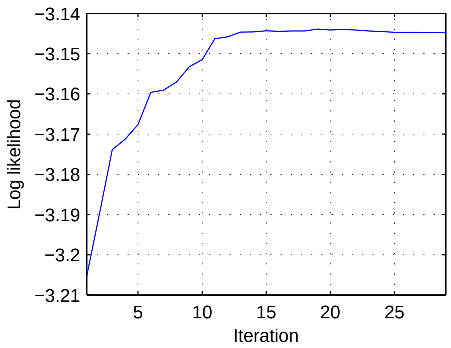


Figure 4.9: Convergence of the log likelihood of the variational EM adaptation algorithm for the same mixture as figure 4.6.

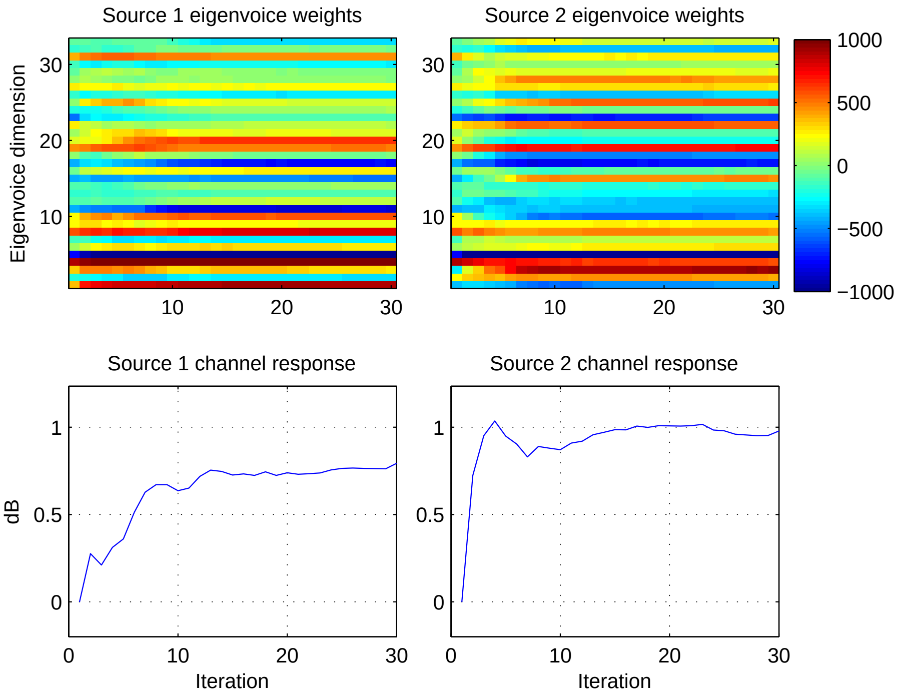


Figure 4.10: Example of the convergence behavior of the hierarchical adaptation algorithm on the same mixture as figure 4.6. The top panes plot the adaptation parameters $\mathbf{w}_1$ and $\mathbf{w}_2$ at each iteration, and the bottom panes show the channel response $B\mathbf{h}_1$ and $B\mathbf{h}_2$ at each iteration.

command	color	preposition	letter	digit	adverb
bin	blue	at		0-9	again
lay	green	by	a-v		now
place	red	in	x-z		please
set	white	with			soon

Table 4.1: Grammar used to generate sentences in the GRID data set [18]. One element is chosen from each column to form a sentence, e.g. "place red by c 6 now".

and figure 4.10 shows the adaptation parameters. As with the hierarchical algorithm, the variational EM algorithm takes about 15 iterations to converge on this mixture.

4.3 Experiments: 2006 Speech Separation Challenge

The learning algorithms described in the previous section were evaluated on the test data from the 2006 Speech Separation Challenge (SSC) [16]. This data set is composed of 600 artificial speech mixtures composed of utterances from 34 different speakers, each mixed at signal to interference ratios varying from -9 dB to 6 dB. Each utterance follows the pattern *command color preposition letter digit adverb* as shown in table 4.1. The task is to determine the letter and digit spoken by the source whose color is "white".

About 15 minutes of training data was available for each speaker in the test set. We used the HTK toolkit [126] to train models for each speaker. Prior to training, the data was downsampled to 16 kHz and pre-emphasized using a simple first order high pass filter with transfer function $1 - 0.97z^{-1}$ to whiten the spectrum. Each of the 35 phones used in the task grammar are modeled using a standard 3-state forward HMM topology. Each state's emissions are modeled by a GMM with 8 mixture components.

The training data for all 34 speakers was used to train a speaker-independent (SI) model. We also constructed speaker-dependent models for each speaker and learned eigenvoice bases as described in section 3.4.1. We learned two such sets of eigenvoice bases: one based only on the mean parameters and one that included both the mean and covariance parameters. Because the training and test data were matched in terms of channel response, we use a single flat channel basis to estimate the frequency independent gain for each source signal, corresponding to the scaling used to generate the mixtures at different SNRs.

Note that in the experiments described in this section we do not discard any low variance eigenvoice dimensions (i.e. U_s has 33 columns) so it is possible to exactly reconstruct the original model for speaker i given the corresponding weights w_i. However, in practice the adapted models do not match the original speaker models perfectly because the adaptation is based only on the relatively small set of model states used to generate a single utterance.

We evaluate the speaker parameter estimation algorithms described in the previous section using these models. We compare the eigenvoice speaker adaptation algorithm (SA) to separation based on model selection of speaker-dependent parameters (SD) as described in section 4.2.1. Because of the simple nature of the channel variation we use the discrete channel prior for model selection. No prior distributions are used in the eigenvoice and channel updates. In all cases, once the speaker parameters are known, we use them to separate the signals as described in section 4.1.2. Then, we reconstruct the time-domain sources $x_i(t)$ from the STFT magnitude estimates $\hat{x}_i$ using the phase of the mixed signal. The final stage is to select the target source from the reconstructed signal. The two reconstructed signals are passed to a speech recognizer; assuming one transcription contains "white", it is taken as the target source.

We use the default HTK speech recognizer provided by the challenge organizers [16], retrained on 16 kHz data. The acoustic model consists of whole-word HMMs based on MFCC, delta, and acceleration features. We select the target by recognizing both signals using two different grammars: one containing only "white" and one without it. The combination of white and non-white with the highest combined likelihood are labeled target and masker, respectively. Finally, given the target signal, performance is measured using word accuracy of the letter and digit spoken by the target speaker. The learning algorithms described in section 4.2 are compared using this common framework.

4.3.1 Comparison of subspace parameter estimation algorithms

We begin by comparing the performance of the three algorithms described in the previous section on the 0 dB SNR subset of the SSC data set. This consists of 200 single-channel mixtures of two talkers of different gender, and 179 mixtures of two talkers of the same gender. All systems were evaluated using models where only the means were speaker-dependent or adapted (Mean Only) as in [117, 118], as well as using models where both the means and covariances were speaker-dependent or adapted (Full).

In order to explore the upper bound performance of source separation based on the proposed model, we evaluate separation using oracle knowledge of the speaker identities and relative gains to fix the settings of $\mathbf{w}_i$ and $\mathbf{h}_i$ (Oracle). We also attempt to isolate the effects of subspace adaptation from the task of inferring the adaptation parameters from a mixture by using oracle knowledge of the clean source signals for adaptation and then performing separation using the pre-adapted models (Estimated Oracle). The adaptation parameters for each source are estimated from the clean signals using the MLED adaptation algorithm (initialized to zero) described in section 4.2.2.2.

The oracle systems are compared to the three adaptation algorithms: selection of SD models using the algorithm described in section 4.2.1 (SD model selection), hierarchical separation and adaptation (Hierarchical), and variational EM adaptation (VEM). Finally, a lower bound on the performance of the separation algorithms is obtained by running the speech recognizer over the mixture (Baseline).

Three variations of the hierarchical adaptation algorithm are compared. The first, labeled Hierarchical (SD init), is initialized using the SD model selection algorithm. The second, labeled Hierarchical, uses the initialization scheme described in section 4.2.2.1. Both Hierarchical (SD init) and Hierarchical are based on the separation algorithm and state pruning described in section 4.1.2 which was found to give a reasonable trade-off between computation time and performance. The third, labeled Hierarchical (fast), uses

	Mean Only		Full	
Algorithm	Same Gender	Diff Gender	Same Gender	Diff Gender
Oracle	75.42 (+0.0)	80.00 (+0.0)	81.56 (+0.0)	81.75 (+0.0)
Estimated Oracle	73.46 (+1.4)	77.00 (+0.5)	74.02 (+0.6)	80.75 (+0.0)
SD model selection	72.07 (+2.1)	76.00 (+2.0)	83.52 (+0.5)	80.00 (+1.2)
Hierarchical (SD init)	75.42 (+0.8)	77.50 (+0.3)	80.17 (+0.8)	82.25 (+1.0)
Hierarchical	56.15 (+0.6)	66.75 (+3.5)	62.29 (+0.2)	79.25 (+0.2)
Hierarchical (fast)	46.63 (+7.3)	57.25 (+4.8)	50.28 (+4.2)	66.25 (+1.5)
Variational EM	47.49 (+0.8)	61.75 (+2.5)	58.10 (+0.2)	69.75 (+5.0)
Baseline	36.03	34.75	36.03	34.75

Table 4.2: Digit-letter recognition accuracy (in percent) on the 0 dB SNR two-talker subset of the 2006 Speech Separation Challenge data set. Numbers in parenthesis correspond to the absolute improvement in accuracy using corrected target speaker selection based on distance from ground-truth target signal instead of output of the speech recognizer.

more aggressive pruning in the separation stage where only the 40 most likely state combinations are retained at each frame (versus 200 in the previous case). This speeds up the hierarchical algorithm to be about as fast as the variational algorithm, allowing for a more fair comparison between them.

The recognition results are summarized in table 4.2. All of the evaluated separation systems show very large improvements over the baseline. The addition of speaker specific covariance parameters gives a significant performance improvement of between 5% and 10% absolute to all systems under all conditions, with the exception of the oracle systems whose improvements are sometimes smaller. Separation using models adapted to the clean signals (Estimated Oracle) performs almost as well as Oracle, indicating that the subspace adaptation approach is able to accurately adapt to a single clean utterance. Any reduction in performance in the Hierarchical and VEM algorithms is therefore the result of poor adaptation when only the mixture is observed.

After the oracle systems, the system based on SD model selection and Hierarchical (SD init) perform best overall, with accuracy on par with Oracle. Hierarchical (SD init), which combines quantization of the speaker model space with continuous adaptation of the model parameters, tends to slightly outperform SD model selection alone. Comparing the remaining adaptation algorithms, the full hierarchical algorithm outperforms variational EM which further outperforms Hierarchical (fast).

The large performance gap between Hierarchical (SD init) and the remaining adaptation algorithms demonstrates their sensitivity to initialization. The likelihood surface of the mixed signal model is inherently multimodal because it contains a large number of unknown parameters, including the HMM state sequences for each source signal and the corresponding adaptation parameters. As a result, if the initialization does not adequately differentiate the sources, the adaptation algorithms are prone to getting stuck in local maxima that correspond to suboptimal separation. When an appropriate set of SD models is available, using SD model selection to initialize the adaptation parameters eliminates

	Mean Only		Full	
Algorithm	Same Gender	Diff Gender	Same Gender	Diff Gender
Oracle	6.83 (+0.00)	7.68 (+0.00)	7.76 (+0.00)	8.51 (+0.00)
Estimated Oracle	6.19 (+0.18)	7.40 (+0.02)	6.63 (+0.01)	7.38 (+0.07)
SD model selection	6.83 (+0.04)	7.26 (+0.07)	7.64 (+0.03)	8.08 (+0.12)
Hierarchical (SD init)	6.77 (+0.16)	7.83 (+0.06)	7.30 (+0.05)	8.13 (+0.03)
Hierarchical	4.16 (+0.34)	6.25 (+0.26)	5.14 (+0.37)	7.47 (+0.12)
Hierarchical (fast)	3.67 (+0.39)	4.64 (+0.62)	4.11 (+0.32)	5.85 (+0.15)
Variational EM	3.29 (+0.50)	5.34 (+0.33)	4.40 (+0.28)	6.30 (+0.40)

Table 4.3: Magnitude SNR (in dB) of target reconstruction on the 0 dB SNR two-talker subset of the 2006 Speech Separation Challenge data set. Numbers in parenthesis correspond to the absolute improvement in accuracy using corrected target speaker selection based on distance from ground-truth target signal instead of output of the speech recognizer.

this problem.

Comparing SD model selection to Hierarchical (SD init), it is clear that adapting the models to better match the mixture improves performance over the SD initialization. The exception is on same gender mixtures using full models. This is a result of the fact that the covariance parameters are not optimized directly, so the optimization of equations (4.31) and (4.41) can sometimes lead to estimates of the model covariances that are worse than the initialization. The effect of this problem is more apparent in same gender mixtures because they are more sensitive to the settings of the model covariances than different gender mixtures. When compared to the Oracle system, the Estimated Oracle system exhibits similar behavior under these conditions. However, despite this problem, the performance of a given separation system using full models in these experiments always improves over the corresponding mean only variant.

The SD system significantly outperforms the best performing adaptation systems that use the default initialization on same gender mixtures. The advantage on different gender mixtures is not as pronounced. This is because same gender sources have more overlap, which makes it more difficult to segregate them, which in turn makes it difficult for the adaptation algorithm to isolate regions unique to a single source. Therefore, the adaptation algorithms sometimes converge on solutions which are partial matches for both speakers, leading to source reconstructions which contain phone permutations across sources as described earlier. This hypothesis is confirmed by the target reconstruction SNR performance shown in table 4.3. The metric is based on the SNR of the magnitude of the target reconstruction, ignoring phase:

$$\mathrm{SNR}_{\mathrm{mag}} = 10 \log_{10} \frac{\| \boldsymbol{x}_{\mathrm{mag}}(1..T) \|^2}{\| \boldsymbol{x}_{\mathrm{mag}}(1..T) - \hat{\boldsymbol{x}}_{\mathrm{mag}}(1..T) \|^2} \tag{4.46}$$

where $\boldsymbol{x}_{\mathrm{mag}}(t) = 10^{\frac{\boldsymbol{x}(t)}{20}}$, $\boldsymbol{x}(t)$ is the ground truth target signal, $\hat{\boldsymbol{x}}(t)$ is the source reconstruction, and $\| \cdot \|$ is the Frobenius norm operator. As with the recognition results, the performance on same gender mixtures is significantly reduced (by about 3 dB) on the

	Mean Only		Full	
Algorithm	Same Gender	Diff Gender	Same Gender	Diff Gender
Oracle	100.00	100.00	100.00	100.00
Estimated Oracle	94.97	99.50	97.21	98.00
SD model selection	97.77	97.00	98.88	98.00
Hierarchical (SD init)	96.09	99.00	98.88	99.00
Hierarchical	83.24	92.00	87.71	98.00
Hierarchical (fast)	80.34	82.50	85.47	93.50
Variational EM	74.86	88.00	83.24	90.50

Table 4.4: Target speaker selection accuracy (in percent) on the 0 dB SNR two-talker subset of the 2006 Speech Separation Challenge data set.

adaptation based systems. This indicates that the performance reduction is due to source reconstruction errors (e.g. source permutations), not only recognition errors.

These source permutation problems could be improved through the use of a different initialization scheme which is better able to segregate same gender mixtures. This is demonstrated by the small difference in performance between different gender and same gender mixtures in the Hierarchical (SD init) system. The initialization scheme described in section 4.2.2.1 does not always suffice to differentiate between sources in same gender mixtures, leading to the same (or very similar) initializations for both sources. This occurs because the eigenvoice dimensions with high variance, which are the ones that are used to differentiate the sources, are well correlated with gender. When the initialization is symmetric across all sources, they can only be segregated using the model's temporal constraints. By design, the temporal constraints in our model are somewhat weak, so nearly symmetric initialization on same gender mixtures leads to source permutations.

The variational EM algorithm performs almost as well as the full hierarchical algorithm, particularly on same gender mixtures when both means and covariances are adapted. Unfortunately, its recognition accuracy is reduced by as much as 9% absolute under the other conditions. This is at least partially a result of recognition errors as the decrease in SNR is not as dramatic. In fact, the significantly reduced performance of the variational EM algorithm on different gender mixtures when using full model adaptation is a result of errors made during target speaker selection, i.e. the process of deciding which voice has spoken the target color (see table 4.4). When these errors are corrected, the difference in performance between the two systems is reduced by more than half. As before, this indicates that the variational EM algorithm is more prone to source permutations or the addition of artifacts than the hierarchical algorithm. The sped up hierarchical algorithm suffers from similar problems, making even more errors due to target speaker selection than the variational algorithm.

Qualitatively, the main difference between the two algorithms is that the EM approach considers all possible paths through the joint state space of the speech models whereas the hierarchical algorithm focuses only on the most likely path. This results in differing convergence behavior of the two algorithms. This is made worse by the fact that our implementation of the variational EM algorithm does not evaluate all possible state paths. As described in section 4.2.3.1, the algorithm is initialized using an approximate computation of the posterior distribution using aggressive pruning of state combinations with low likelihood. Unfortunately, such pruning can result in the state combinations with high posteriors being discarded altogether. This problem is exacerbated because the observation likelihoods tend to vary over a wide dynamic range (a side effect of the high dimensional spectral representation), making it likely that a particular path will be overlooked if it is a poor match for a very short segment, even if it is a good fit for another. If this initialization procedure discards the optimal state combinations, they will never be recovered, ultimately leading to poor separation. While the resulting lattice is generally reasonably accurate, this problem does happen. Increasing the beam pruning threshold would alleviate the problem, at the cost of significantly increasing the computation time, which goes against the very motivation behind this algorithm in the first place.

A second, more minor, difference between these two approaches lies in the way the eigenvoice parameters are updated in each iteration. In the hierarchical algorithm the parameters are learned to match the source reconstructions. In dimensions dominated by the masking signal the reconstructions contain the mean settings corresponding to the previous parameter setting. In contrast, the variational EM algorithm is designed to simply ignore the masker-dominated regions. Because the source masks are held constant, the updates are approximate by nature. This means that it is possible for the parameters to be updated in a way that flips the mask around in some regions. As a result, over the course of multiple iterations, the masks, and thus the parameters, can oscillate around the correct settings. On the other hand, the reconstruction used in the hierarchical updates have a regularizing effect on the parameter updates, encouraging the parameters to remain consistent with the previous settings and therefore the previous mask, avoiding this problem.

The reduced performance of the variational algorithm is a result of the combination of these effects. However, the advantage to this algorithm is that the nature of the approximation allows it to run significantly faster than the hierarchical algorithm, which runs the Viterbi algorithm over the factorial HMM state space for every iteration. Our Matlab implementation of the variational E-step runs about 3-5 times faster than our previous optimized, pruned, C-coded Viterbi search. Based on the significant difference in performance between the variational algorithm and the sped up hierarchical algorithm it is clear that the variational approach is superior when speed is the primary concern. However, despite this speed advantage, the cost in terms of accuracy was significant, so we only use the hierarchical learning algorithm in the remaining experiments.

Recently, Rennie et al. [86] described an extremely efficient method for FHMM separation based on loopy belief propagation that scales linearly in the number of sources. Using this algorithm instead of the Viterbi search would significantly increase the speed of the hierarchical algorithm, eliminating the speed advantages of the variational approach. It is also possible to use this method of inference to develop an efficient EM algorithm similar to the variational approach that does not suffer from the variational parameter initialization problems described in section 4.2.3.1.

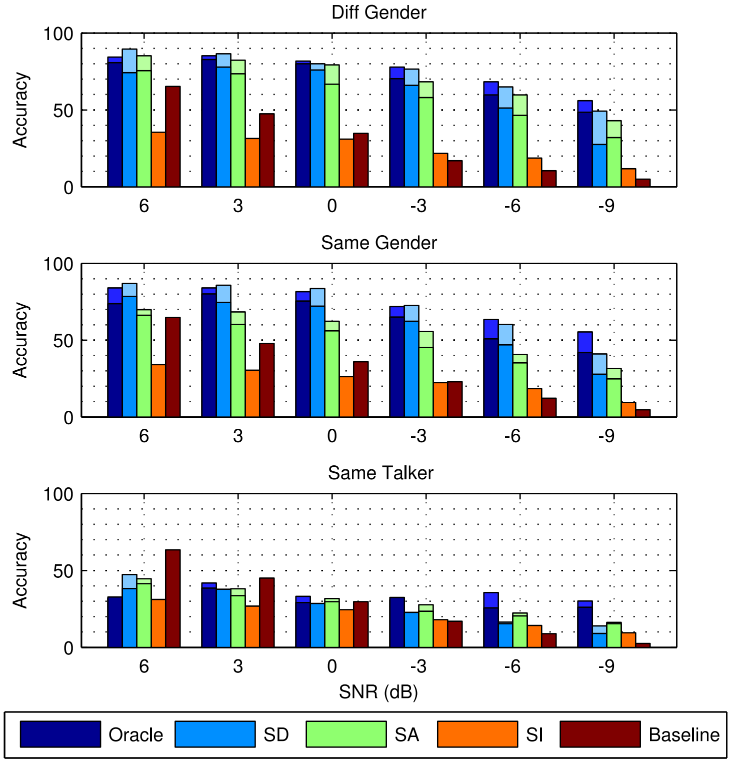


Figure 4.11: Separation performance using speaker-dependent (SD), speaker-adapted (SA), and speaker-independent (SI) models. Also shown is performance of separation using oracle knowledge of speaker identities and relative gains (Oracle) and baseline performance of the recognizer on the mixed signal (Baseline). For oracle, SD, and SA, darker colors correspond to performance using mean only models and lighter colors correspond to performance using full models.

4.3.2 SSC results

In this section we evaluate the proposed adaptation method on the full 2006 Speech Separation Challenge test set. Figure 4.11 compares the performance of the speaker adaptation (SA) system to two comparison systems based on SD and SI models respectively. As before, for both SA and SD systems we used two different sets of models: one using speaker-dependent (or adapted) means only and one using adapted means and covariance. We also compare this to performance when using oracle knowledge of the speaker identities and gains. Finally, we include baseline performance of the recognizer generating a single transcript for the original mixed signal. The overall performance of the SA systems are also listed in tables 4.5 and 4.6 for the mean only and full cases respectively. SD performance is listed in tables 4.7 and 4.8.

SNR	Same Talker	Same Gender	Diff Gender	Avg.
6 dB	41.44	66.20	75.50	60.15
3 dB	33.56	60.34	73.50	54.83
0 dB	29.73	56.15	66.75	49.92
3 dB	23.42	45.25	58.00	41.43
-6 dB	20.50	35.20	46.50	33.53
-9 dB	15.32	24.86	32.00	23.71

Table 4.5: Recognition accuracy (in percent) on the 2006 Speech Separation Challenge data test set using our source-adapted separation system with mean only models.

SNR	Same Talker	Same Gender	Diff Gender	Avg.
6 dB	44.59	69.83	85.25	65.64
3 dB	38.06	68.44	82.25	61.81
0 dB	31.76	62.29	79.25	56.66
-3 dB	27.70	55.59	68.25	49.50
-6 dB	22.30	40.78	59.75	40.27
-9 dB	16.22	31.56	43.00	29.70

Table 4.6: Recognition accuracy (in percent) on the 2006 Speech Separation Challenge data test set using our source-adapted separation system with full models.

Looking at general trends, we see that the SD models perform similarly whether using oracle or Iroquois-style speaker identification. Both of these often perform significantly better than the SA system, which performs better than the SI system and baseline. As before, using fully adapted models further improves performance over the mean only models.

The reduced performance of the SA system in this task is mainly due to its vulnerability to permutations between sources, which reflects the sensitivity of the initial separation to initialization. The adaptation process is able to compensate for limited permutations, as demonstrated in figure 4.7. However when the initialization does not sufficiently separate the sources, the system can get stuck in poor local optima where each of the estimated sources is only a partial match to the ground truth. In contrast, it performs significantly better on the different gender condition because the initial separation tends to be better. In fact, when using full models the SA system's performance is quite close to that of the SD system on different gender mixtures. The errors get worse as SNR decreases because the stage of initialization that adapts to the mixed signal favors the louder source. Fortunately the algorithm is generally able to compensate for poor initialization as it iterates, however, as shown in figure 4.12, this process takes many iterations.

The performance of the SI system is not sensitive to the different speaker conditions because the same model is used for both sources. The other separation systems work best on mixtures of different genders because of the prominent differences between male

SNR	Same Talker	Same Gender	Diff Gender	Avg.
6 dB	38.29	78.49	74.25	62.23
3 dB	37.84	74.58	77.75	62.06
0 dB	28.60	72.07	76.00	57.32
-3 dB	22.75	62.29	66.00	48.92
-6 dB	15.32	46.93	51.25	36.69
-9 dB	9.01	27.93	27.50	20.80

Table 4.7: Recognition accuracy (in percent) on the 2006 Speech Separation Challenge data test set using mean-only speaker dependent models and Iroquois speaker identification.

SNR	Same Talker	Same Gender	Diff Gender	Avg.
6 dB	47.30	86.87	89.50	73.13
3 dB	34.01	85.75	86.50	66.89
0 dB	22.30	83.52	80.00	59.73
-3 dB	19.37	72.63	76.50	54.24
-6 dB	16.44	60.34	65.00	45.67
-9 dB	13.96	41.06	49.25	33.78

Table 4.8: Recognition accuracy (in percent) on the 2006 Speech Separation Challenge data test set using full speaker dependent models and Iroquois speaker identification.

and female vocal characteristics, which means that such sources tend to have less overlap. Conversely, the performance on the same talker task is quite poor. This is because the models used for each source are identical (or close to it in the SA case) except for the gain term, and the models enforce only limited dynamic constraints. The performance of the SI system is quite poor in all conditions for the same reason. The marked difference between the SA and SI systems demonstrates that adapting the source models to match the source characteristics can do a lot to make up for the limited modeling of temporal dynamics.

The source permutation errors on the same talker task are reminiscent of human listening test results reported by Cherry [13]. In these experiments, monaural mixtures were generated from utterances consisting of connected strings of short "cliche" phrases spoken by the same talker. Test subjects were unable to reliably transcribe either of the sources in their entirety. Instead, the transcriptions consisted of a grammatically correct string of cliches composed of equal numbers of phrases spoken by each of the underlying sources. The top down constraints from the listeners' knowledge of the rules of English grammar allowed them to reliably identify individual sources over the short term. However, in the absence of other source-specific information these constraints were ambiguous over the long term, making accurate separation impossible. The phone permutations in the results obtained by our separation systems are consistent with this behavior, albeit at a shorter time scale, due to the lack of a strong language model.

The significance of the adaptation process is demonstrated in figure 4.12 which shows how

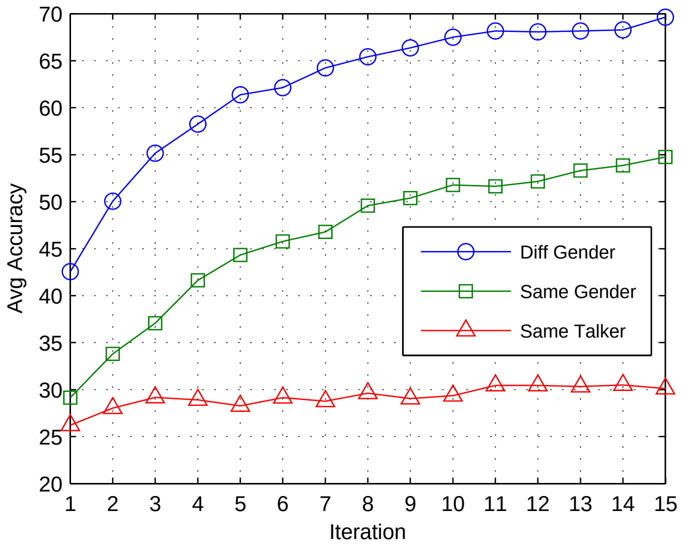


Figure 4.12: Separation performance improvement of the SA system using full models averaged over all SNRs as the source adaptation/separation algorithm iterates.

the recognition accuracy of the SA system improves after each iteration averaged across all SNRs. It is quite clear from this figure that the iterative algorithm helps significantly, increasing the average accuracy by over 20% in both the same gender and different gender conditions after 15 iterations. The performance improvement for the same talker condition is more modest. While the models are able to adapt somewhat to match the specific sounds (i.e. model states) present in each source signal, they still remain very similar. Because of this and the limited temporal dynamics built into the models, the same talker separations suffer from source permutations.

Table 4.9 compares the average performance over all SNRs of all participants in the evaluation to the systems proposed in this chapter. A breakdown of performance as a function of SNR is shown in figure 4.13. Broadly speaking, in mixtures of different talkers our SD systems have performance roughly on par with the other model-based systems. Like Kristjansson et al. [55], the SD (Full) system even outperforms human listeners under some conditions. Despite the fact that they do not rely on precise knowledge of the exact talkers present in the mixture, the SA systems are still able to perform quite well.

As described above, the primary weakness of all of our systems when compared to the top performers lies in its performance on same talker mixtures. This is a direct result of our deliberate attempt to be as general as possible and not use any grammar-specific knowledge in our temporal models. This is in direct contrast to Kristjansson et al. [55] who explicitly use grammar knowledge for separation and Virtanen [112] whose separation system is integrated with the recognizer. Such integration is also present in the CASA-based systems [7, 103]. Another trend in figure 4.13 is that the performance of the proposed systems drops off faster than the top performers at very low SNRs. This is because unlike e.g. [55, 8]

System	Description	ST	SG	DG
Human [17]	N/A	66	81	88
Kristjansson [55]	Source models, FHMM	56	87	87
Virtanen [112]	Source models, FHMM	48	77	75
Ming [70]	Source models	51	60	65
Barker [7]	CASA, Speech fragment decoder	48	62	64
Schmidt [97]	Source models, NMF	42	47	62
Srinivasan [103]	CASA	28	52	61
Deshmukh [22]	Phase Opponency	30	33	32
Every [30]	Pitch tracking	19	23	28
Runquiang [93]	CASA	19	22	24
SA (mean only)	Eigenvoice models, FHMM	27	48	59
SA (full)	Eigenvoice models, FHMM	30	55	70
SD (mean only)	Source models, FHMM	25	60	62
SD (full)	Source models, FHMM	26	72	74

Table 4.9: Performance of all participants in the 2006 Speech Separation Challenge averaged over all SNRs compared to the proposed separation systems.

we do not take advantage of the fact that the mixtures were generated from a limited set of SNRs. Instead we initialize the channel response to zero for all sources in all systems. Once again, this demonstrates how performance suffers when minimal assumptions are made about the mixtures.

The final weakness of our system relative to the top performers is the speech recognizer. Like many of the entrants [97, 22, 30], we utilize the default recognizer provided by the organizers. It is clear that using more advanced ASR methods, such as the speaker-dependent acoustic models used in [55, 112, 7] and incorporating knowledge of the background signal into the recognizer as in [112, 7] would improve recognition performance. Many of the errors made by the proposed systems are caused by common phonetic confusions between letters (e.g. "e" and "v" are confused 15% of the time in SD (Full) separation). However, some of these errors are the result of phone permutations. These are particularly harmful to this task because it requires the accurate identification of isolated phones, i.e. single letters in the grammar, which are sometimes permuted. Finally, we note that the target speaker selection algorithm also depends on the performance of this recognizer, so any errors it makes in this stage are passed through to the final performance metric.

4.3.3 Held out speakers

As described previously, source separation systems based on speaker-dependent models are potentially at a disadvantage when presented with mixtures containing sources that are not represented in the SD model set. We expect that the SA system should be better suited to handling such cases. To evaluate this hypothesis we separated the SSC training data into

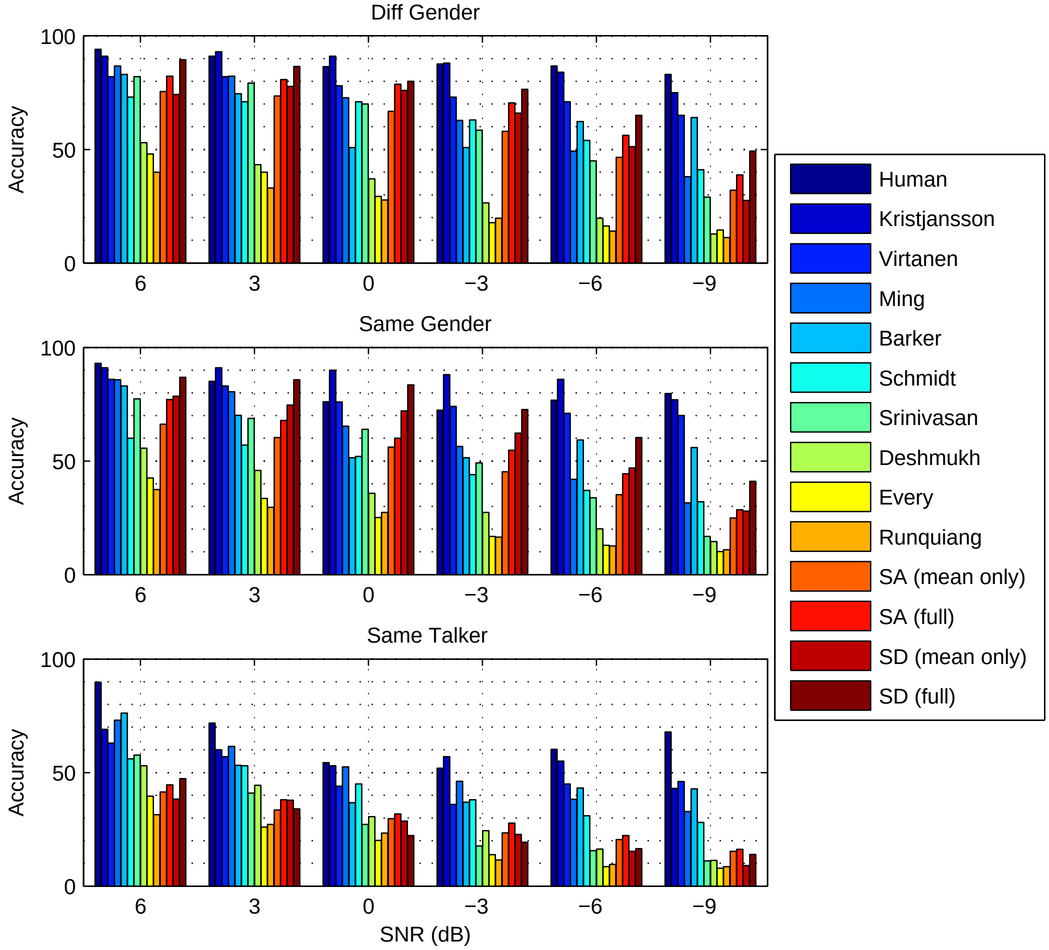


Figure 4.13: Performance of all participants in the 2006 Speech Separation Challenge compared to the proposed separation systems.

random subsets of 10, 20, and 30 speakers and trained new eigenvoice models from each subset. All eigenvoice dimensions were retained for these models, e.g. the model trained from 10 speaker subset used 9 eigenvoice dimensions for adaptation, etc. A new test set was generated from utterances from the four speakers held out of all training subsets. The held out speakers were evenly split by gender. The test set consists of 400 mixtures at 0 dB SNR, broken up into 200 same gender mixtures and 200 different gender mixtures. These experiments only utilize mean only models.

Figure 4.14 compares the performance of the SD system and SA system on this data set. The SD models used were limited to the same speakers as were used to train the new eigenvoice models. Performance using smaller subsets of training speakers is compared to a baseline containing all 34 speakers from the training set. It is important to note that the mixtures in the test set were generated from portions of the clean data used to train the

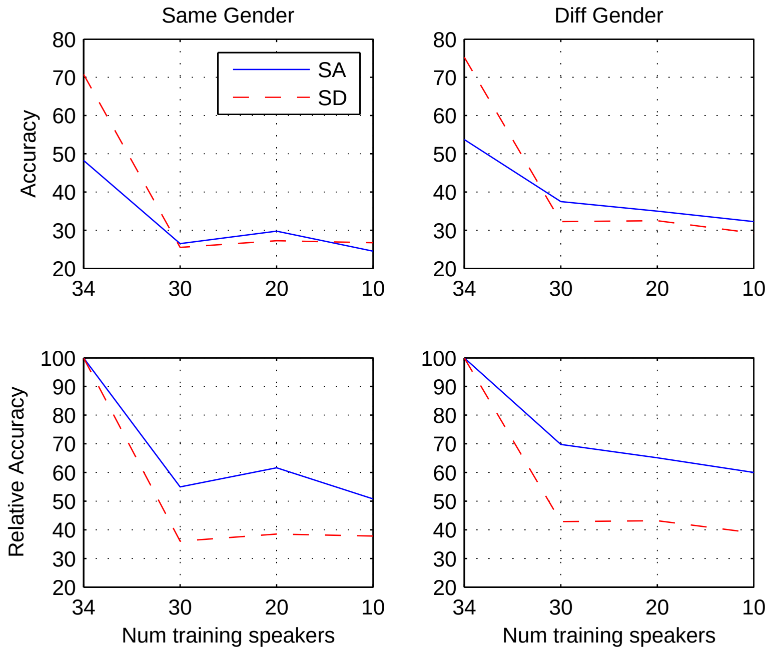


Figure 4.14: Performance on mixtures of utterances from held out speakers using only subsets of 10, 20, and 30 speakers for training compared to models trained on all 34 speakers.

baseline 34 speaker SD and SA models. The performance would likely have been worse had there been enough clean data to properly generate a separate test set, so the accuracy shown for the 34 speaker set should be viewed as an upper bound.

Performance of both the SD and SA systems suffers on held out speakers, but the performance decrease relative to the use of models trained on all 34 speakers shown in the bottom row of figure 4.14 is much greater for the SD models. In fact, the SA system slightly outperforms the SD system in absolute accuracy in most of the held out speaker cases. It is clear that separation using eigenvoice speech models generalizes better to unseen data than separation based on model selection from a set of SD models.

Despite this, the performance drop on held out speakers for both systems is quite significant. We expect that this is because a relatively small set of speakers were used to train the systems. As the number of eigenvoice training examples increases we expect the model to better capture the characteristics of the general speaker space and thus be able to generalize better to unseen speakers. At the same time, as the number of training speakers grows it becomes increasingly likely that one of the models in the set will be a good match for a previously unseen speaker. Still, we expect that the performance of the SD system will not improve as quickly as the SA system as the size of the training set grows. This can be seen to some extent in the fact that the SD system has a flatter slope than the SA system as the number of models decreases in figure 4.14, especially in the different gender case.

Based solely on these experiments it is unclear how a system based on eigenvoice adaptation would compare to a system based on model selection from a large set of speaker-dependent models. The results in [58] indicate that on the order of 100 speaker models are needed to learn eigenvoice models that generalize well. The SSC data set contains training data for relatively few speakers, and based on the results in this section, it is clear that more are needed to learn adequate speaker subspace bases. In the following section we describe a set of experiments that better evaluate the ability of these systems to generalize to previously unseen speakers using training data from hundreds of speakers.

4.4 Experiments: Switchboard

In this section we evaluate the proposed separation systems on a data set similar to the Speech Separation Challenge set used in the previous section but created from a larger set of speakers. We use the Switchboard corpus of conversational telephone speech [36] which contains a significant amount of data from over 500 speakers. The use of such a large data set should allow the estimation of more accurate eigenvoice bases, which would improve adaptation performance. Despite containing data from a large number of speakers, this data set presents a number of challenges that were not present in the previous experiments. It consists of conversational speech recorded under a wide variety of recording conditions, making for a more realistic separation task than the highly constrained setting of the Speech Separation Challenge.

We divided the set of speakers into training and testing sets containing 253 and 290 speakers respectively. All speakers for whom there was more than 30 minutes of data available were assigned to the training set, the remaining speakers were used solely for testing. The training speakers were further subdivided into smaller sets of 250, 128, 64, and 32 speakers to evaluate performance as a function of number of training speakers. The data from each training set was used to a train a set of eigenvoice models as described in chapter 3. In all cases the models used speaker-dependent means and covariances and only retained the 30 eigenvoices with highest variance.

All data from the 32 speaker training subset left over after removing 30 minutes for training was used to create a set of 100 test mixtures, 50 mixtures of speakers of the same gender and 50 different gender. The mixtures were generated by taking randomly selected 5 second clips based on the segmentation from [104]. All test signals were mixed at 0 dB SNR. This "In Train 32" test set was used to evaluate the performance of the different separation systems on matched training and testing speakers. We also created an analogous test meant to evaluate the ability of the different methods to generalize to previously unseen speakers. This "Not In Train" test set based only on data from speakers not used at all for training.

It is important to note that the nature of the Switchboard data makes the task significantly more difficult than the Speech Separation Challenge. First, it has half the bandwidth of the 16kHz data we used in our SSC system, which makes identification of phones dominated by high frequency energy such as fricatives more difficult.

A more significant difference lies in the nature of the speech. The GRID data consists of speech read from a prompt that is generated from a very constrained grammar. In

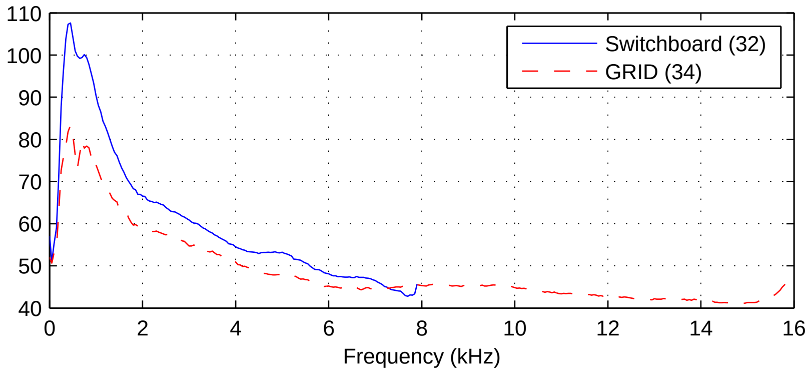


Figure 4.15: Average diagonal covariance parameters of GRID and Switchboard speaker models. The parameters are averaged across all speakers and HMM states.

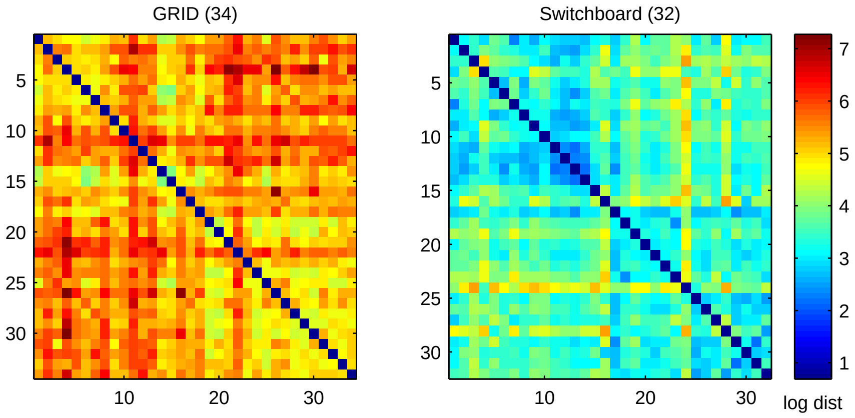


Figure 4.16: Illustration of the pairwise distances between GRID and Switchboard speaker models. The distances between GRID models are higher than those between Switchboard models implying that the Switchboard speaker models are more easily confusable with each other. The distance metric used is the normalized Jensen approximation to the Bhattacharyya divergence between GMMs described in Hershey and Olsen [40].

contrast, the Switchboard utterances come from a completely unconstrained vocabulary and are spontaneous in nature, containing a lot of irregularities such as laughter and various disfluencies common in natural speech. As in ASR applications, an acoustic model appropriate for such data is generally composed of context-dependent phone models containing thousands of HMM states. Despite this, we are forced to use a context independent phone configuration for our speech models composed of only 556 states because of the unfavorable scaling in terms of computational complexity inherent in the FHMM signal model. As a result the Switchboard phone models tend to have higher variances than the GRID models because they are averaged over many different contexts all containing more variation than is present in the more constrained GRID data. This is demonstrated in figure 4.15. A direct consequence of these simplifications is that the Switchboard speaker models must necessarily be more loose and thus less distinguishable from each other than the GRID models. This is demonstrated in figure 4.16 which plots the pairwise distance between the GRID and Switchboard speaker models. The distance metric is the negative logarithm of the normalized Jensen approximation to the Bhattacharyya divergence between GMMs described in Hershey and Olsen [40], an approximation to the Bayes error between GMMs which measures the amount of overlap of the distributions. This metric was originally proposed as a measure of word confusability between ASR acoustic models. The average distance between Switchboard models (3.5) is significantly lower than that between GRID models (5.4), implying that Switchboard speakers will likely be more difficult to isolate than the GRID speakers.

The final difficulty in modeling Switchboard speech is that each conversation was recorded over a potentially unique telephone channel. Therefore, although all mixtures are at 0 dB SNR, there is still no guarantee that a signal from an arbitrary conversation will be a good match to the corresponding speaker model. This makes the channel compensation technique described in section 3.4.3 particularly important for this data set. Similar methods have also proved to be quite important in speaker verification performed on Switchboard data (e.g. [48]). Therefore, we use a channel basis set composed of 10 low order DCT bases (see figure 4.4 for an example). This also complicates training. In order to obtain speaker models with the channel response factored out, the channel adaptation algorithm of section 4.2.1.1 is used to estimate the channel of all training utterances (pooling all utterances belonging to a particular conversation together because they all use the same channel). The channel estimates are then subtracted out of the data before using it to train speaker models. The channel prior distribution $\mathcal{N}(\mathbf{h}; \nu_h, \Xi_\mathbf{h})$ is learned from the estimated channels as well. Note that this channel model is somewhat weak. A better matched channel basis set can be learned directly from the training data, as described in [48], but this requires a more complex training algorithm to factor the signal into portions that depend on the speaker and portions that depend only on the channel.

The eigenvoice adaptation and model selection (SD) systems were evaluated using all different training sets on the two test sets described above. The adaptation system was based on the hierarchical learning algorithm using two initialization schemes: the method described in section 4.2.2.1 (SA) and initialization based on the model selection algorithm (SA (SD init)). Also plotted is separation using the oracle knowledge of the speaker identities and channel responses (Oracle) and separation using models adapted to the clean source signals (Estimated Oracle). Performance is measured using the magnitude signal-to-noise ratio (SNR) of the final source reconstruction: The results are shown in figure 4.17. As predicted above, the results are somewhat worse than on GRID (see table 4.3). This reflects the difficulty of the unconstrained monaural source separation problem when little is known in advance about the source identities and unknown channel differences between training and test data need to be taken into account.

In the matched training and testing set (In Train 32), all adaptation systems perform roughly comparably. As before, performance on different gender mixtures is better than on same gender mixtures. Also as predicted, SD performs best when the training and test speakers match exactly. As the number of training speakers increases, SD performance decreases.

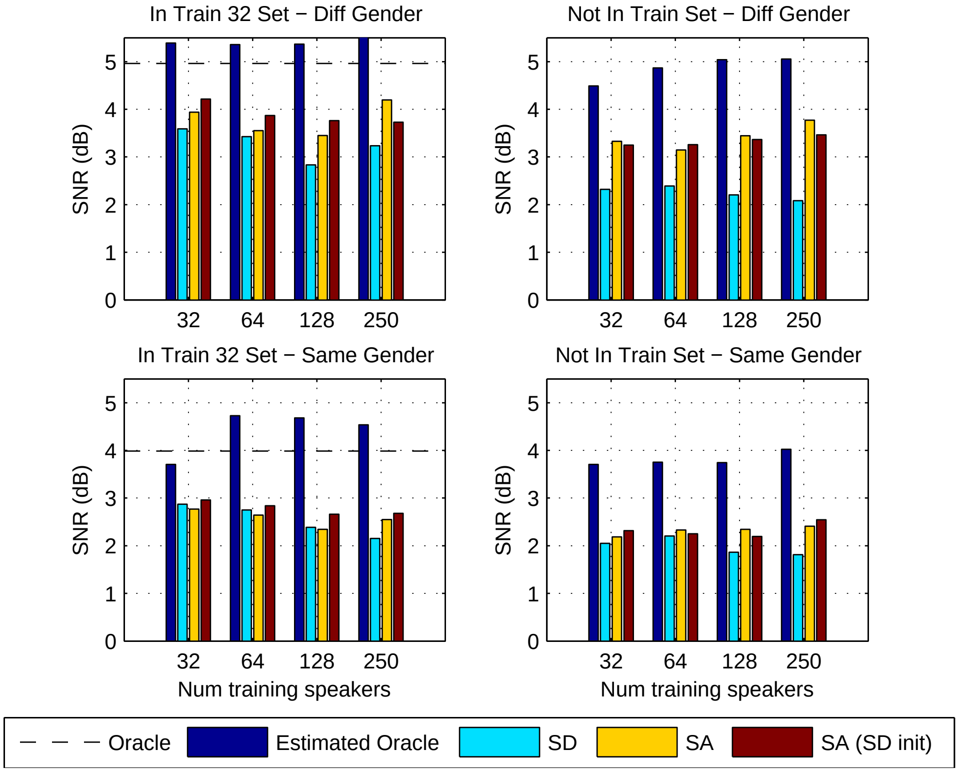


Figure 4.17: Average separation signal-to-noise ratio on the Switchboard data set.

This is because the chance of selecting the wrong model increases. Finally, the SA systems both outperform the SD system in the matched conditions. This is perhaps not surprising for SA (SD init) because it begins from essentially the same point as the SD system but is able to find a closer fit to the particular utterance due to the additional degrees of freedom provided by the eigenvoice model. The fact that this improves performance over the SD system implies that either the channel model alone is insufficient to fit the true channel and the learned eigenvoice bases are able to compensate for this deficiency, or that the model selection algorithm incorrectly identifies the speakers, yet still serves as a superior initialization to the adaptation algorithm. In fact, the model selection algorithm is only correct 49% of the time on the exactly matched conditions (32 training speakers) due to confusion caused by the differences in channel. Still it is likely that the former explanation is true to some extent as well. This reasoning also explains the fact that SA with the default initialization outperforms SD and that Estimated Oracle outperforms Oracle under matched training and testing conditions.

The key result of these experiments is on the mismatched train/test set (Not In Train). The performance of the SD system drops off by an average of more than 1 dB SNR, while the SA systems remain more stable, decreasing by only a small fraction of a decibel. The SA (SD init) system no longer outperforms SA on this data set, however their performance is quite similar. As the number of training speakers increases, performance of the SA systems increase slightly, while SD performance decreases slightly.

Under all conditions, the performance of all SD and SA systems is quite poor relative to the oracle systems. The performance of Estimated Oracle shows similar trends to the SA systems: it performs slightly worse on held out speakers, however it improves as the number of training speakers increases. Once again, the decrease in performance when moving to mismatched training and test speakers is quite modest. This indicates that the subspace adaptation approach is able to accurately capture the speaker-dependent characteristics of held out speakers. The poor performance of the SA systems relative to Estimated Oracle is a result of the difficulty of performing model adaptation when only a mixture is observed, not a shortcoming of the subspace adaptation approach itself.

In conclusion, it is clear from these experiments that the use of the more flexible subspace model has a significant advantage over the model selection approach on more realistic data. The additional confusion caused by the significant channel variation between the training and testing data puts the model selection approach at a disadvantage in these experiments. This could be minimized by using a set of channel bases that is better matched to the data and a better channel prior distribution to improve performance of the selection algorithm. However, even with such improvements it is likely that the SA systems would still perform better under mismatched conditions. In situations where an iterative learning algorithm needs to be applied to the SD system anyway (i.e. when channel compensation is necessary), there is no advantage to not using eigenvoice adaptation as well because the computational costs are approximately equal. Additionally, the SA systems have the advantage of being more compact than the SD systems because only 30 eigenvoice bases are retained, compared to anywhere between 32 to 250 full speaker models in the SD case.

4.5 Summary

In this chapter we described an approach to source model-based monaural source separation based on the speaker subspace model described in chapter 3. We developed two algorithms for estimating the subspace parameters from a mixture and described how the resulting adapted models can be used to reconstruct the sources. These algorithms were compared to an alternate method for source model estimation based on selection from a predefined database of source models in the context of the 2006 Speech Separation Challenge. Although the best performance requires the use of models that capture speaker specific characteristics exactly, we have shown that good results are still possible using minimal prior knowledge about the signal content using speaker adaptation. Finally, we show that although the model selection approach yields superior results when training data is available for the speakers comprising a mixed signal, the approach based on speaker subspace adaptation is better able to generalize to previously unseen speakers, especially if the number of training speakers is sufficiently large. Also, under some conditions using the model selection method to initialize the subspace parameter estimation algorithm can further improve performance.

The greatest weakness of the proposed system is its tendency to permute between sources due to limited modeling of temporal dynamics. This is often alleviated through the iterative eigenvoice re-estimation, however the tendency to fall into local maxima is aggravated by the use of the approximate “max” mixing model. This results in a model with a discontinuous likelihood surface which is difficult to optimize, and necessitates a number of approximations in developing the adaptation algorithms. The use of a continuous mixing model, such as Algonquin [56], which more accurately models the non-linear mixing of equation (4.2) would make these approximations unnecessary and improve the adaptation performance, albeit at increased computational cost. This would also reduce the sensitivity of the adaptation algorithm to initialization. The development of this extension remains future work.

The results of the Switchboard experiments reflect the difficulty of the task of single channel speech separation in general. Although we demonstrated that subspace adaptation can generalize better than simple model selection on this task because it naturally compensates for the effects of speaker and channel variation, the performance of the best adaptation system was still significantly worse than on the Speech Separation Challenge data. Obtaining high quality separation of monaural mixtures of spontaneous speech with varying channel effects requires the use of more complex models that incorporate more detailed knowledge of these sources of variation. However, in situations when more than one channel is observed, good separation is possible using very simple source models. This is the subject of the next chapter.

Chapter 5

Binaural Separation Using Explicit Source Models

In this chapter, we describe a system for separating multiple sources from an underdetermined, reverberant, two-channel recording using the proposed speaker subspace model. The system is based on the model-based expectation maximization source separation and localization (MESSL) system of Mandel et al. [66, 67, 68] which uses a probabilistic model of the interaural spectrogram to localize and separate sources. We extend this model to incorporate the speaker subspace model of source statistics described in chapter 3 and derive an EM algorithm for finding the maximum likelihood parameters of the joint model[1]. We conclude with a series of experiments comparing the performance of different variants of MESSL with other state of the art multichannel separation systems under a variety of conditions.

5.1 Introduction

Signal model based approaches to source separation are common when only a single channel observation is available. However, as seen in previous chapters, such algorithms generally require relatively large, speaker-dependent models to obtain high quality separation. The separation problem becomes considerably easier when multiple channels are observed, as demonstrated by the ability of the binaural system of human listeners to focus on a particular sound source in an environment containing distracting sources. In fact, by leveraging the localization cues present in binaural signals it is possible to separate sources without prior knowledge of their content [125, 67]. However, it is to be expected that incorporating such prior knowledge would be further able to improve separation performance. In this chapter we describe a system for source separation that combines inference of localization parameters with model-based separation methods and show that

[1] The MESSL model described in this chapter was originally developed by Mandel and Ellis [67]. This chapter describes an extension to the MESSL model originally described in Weiss, Mandel, and Ellis [120]

the additional constraints derived from the source model help to improve separation performance.

The MESSL system of Mandel et al. [68] combines a cross-correlation approach to source localization with spectral masking for source separation. The approach utilizes the same localization cues used by the human auditory system. It is based on a model of the interaural phase and level differences derived from the observed binaural spectrograms. This is similar to the DUET algorithm for separating underdetermined mixtures [125] and other similar approaches to source localization [72] which are based on clustering localization cues across time and frequency. Such systems work in an unsupervised manner by searching for peaks in the two dimensional histogram of interaural level difference (ILD) and interaural time, or phase, difference (ITD or IPD) to localize sources. In the case of DUET, this stage is followed up by assigning each time-frequency point of the spectrogram to a particular spatial location to create source specific spectral masks. Harding et al. [39] and Roman et al. [89] take a similar but supervised approach, where training data is used to learn a classifier to differentiate between sources at different spatial locations based on features derived from the interaural cues. Unlike the unsupervised approach of Yilmaz and Rickard [125] and Nix and Hohmann [72], this has the disadvantage of requiring a significant amount of training data. MESSL is most similar to the unsupervised separation algorithms. It is able to jointly localize and separate spatially distinct sources using a parametric model of the interaural parameters estimated from a particular mixture.

A problem with all of these methods is the fact that, as we will describe in the next section, the localization cues are often ambiguous in some frequency bands. Such regions can easily be ignored if the intended application is solely localization, but the uncertainty leads to reduced separation quality when using spectral masking. In this chapter we describe an extension to MESSL which incorporates a prior model of the underlying source signal which does not have the same underlying ambiguities as the interaural observations and therefore is able to better resolve the individual sources in these regions. This approach has the disadvantage of requiring training data to learn the source prior model, but as we will show in section 5.4, such a prior can significantly improve performance even if it is not well matched to the test data.

The idea of combining localization with source models for separation has been studied previously in [122, 82, 87, 3]. Given prior knowledge of the source locations, Wilson [122] describes a complementary method for binaural separation based on a model of the magnitude spectrum of the source signals. This approach combines factorial model separation described in chapter 4 with a model of the IPD based on known source locations. A source-independent (SI) GMM is trained on clean speech and used to model both sources. As described in previous chapters, such a model generally results in very poor separation due to the lack of temporal constraints and lack of source-specific information available to disambiguate the sources. In this case, however, the localization model is able to make up for these shortcomings. Per-source binary masks are derived from the joint IPD and source model. This is shown to improve performance over separation systems based on localization cues alone.

Rennie et al. [82] take a similar approach to combining source models with known spatial locations for separation using microphone arrays. Instead of treating the localization and source models independently, they derive a model of the complex speech spectrum based on a prior on the speech magnitude spectrum that takes into account the effect of phase rotation consistent with a source signal arriving at the microphone array from a particular direction. Like the other systems described above, [82] is able to separate more sources than there are microphones.

These systems have some disadvantages when compared to the extensions to MESSL described in this chapter. The primary difference is that they depend on prior knowledge of the source locations whereas MESSL and its extensions are able to jointly localize and separate sources. Rennie et al. [84] describe an extension to [82] that is able to estimate the source locations as well, bringing it closer to our approach. A second difference is that these systems use a factorial model to model the interaction between different sources. In [122] this leads to inference that scales exponentially with the number of underlying sources. Although the signal model in [82, 84] is similar, they are able to manage this complexity using an approximate variational learning algorithm. In contrast, exact inference in MESSL-SP is linear in the number of sources. This is because each time-frequency cell is assumed to be conditionally independent given the latent variables. Because each frequency band is independent given a particular source and mixture component, the sources decouple and all combinations need not be considered.

In the next section, we describe two closely related extensions to the baseline MESSL algorithm to incorporate a prior distribution over the source signal statistics: MESSL-SP (Source Prior) which uses the same SI model for all sources as in [120], and MESSL-EV (Eigenvoice) which uses eigenvoice adaptation to learn source-specific parameters to more accurately model the source signals. In both cases, the information extracted from the interaural cues and source model serve to reinforce each other. We show that it is possible to obtain significant improvement in separation performance of speech signals in reverberation over a baseline system employing only interaural cues. As in [122, 82], the improvement is significant even when the source models used are quite weak, only loosely capturing the spectral shapes characteristic of different speech sounds. The use of speaker-adapted models in MESSL-EV is able to improve performance even more, a further advantage over the source-independent approach used in [122, 82].

5.2 Binaural mixed signal model

We model the mixture of I spatially distinct source signals $\{\mathrm{x}_i(t)\}_{i=1..I}$ based on the binaural observations $\mathrm{y}^\ell(t)$ and $\mathrm{y}^r(t)$, corresponding to the signals arriving at the left and right ears respectively. For a sufficiently narrowband source in an anechoic environment, the observations will be related to a given source signal primarily by the gain and delay that characterize the direct path from the source location. However, in reverberant environments this assumption is confused by the addition of convolutive noise arising from the room impulse response. In general the observations can be modeled as follows:

$$\mathrm{y}^\ell(t) = \sum_i \mathrm{x}_i(t - \tau_i^\ell) * \mathrm{h}_i^\ell(t) \tag{5.1}$$

$$\mathrm{y}^r(t) = \sum_i \mathrm{x}_i(t - \tau_i^r) * \mathrm{h}_i^r(t) \tag{5.2}$$

where $\tau_i^{\ell,r}$ are the delay characteristic of the direct path for source i and $\mathrm{h}_i^{\ell,r}(t)$ are the corresponding "channel" impulse responses for the left and right channels respectively that approximate the room impulse response and additional filtering due to the head related transfer function (HRTF), excluding the primary delay.

5.2.1 Interaural model

We model the binaural observations in the short-time spectral domain using the interaural spectrogram $X_{IS}(\omega, t)$:

$$\mathrm{X}_{IS}(\omega, t) \triangleq \frac{\mathrm{Y}^\ell(\omega, t)}{\mathrm{Y}^r(\omega, t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)} \tag{5.3}$$

where $\mathrm{Y}^\ell(\omega, t)$ and $\mathrm{Y}^r(\omega, t)$ are the short-time Fourier transforms of $\mathrm{y}^\ell(t)$ and $\mathrm{y}^r(t)$, respectively. For a given time-frequency cell, the interaural level difference (ILD) in decibels between the two channels is $\alpha(\omega, t)$, and the corresponding interaural phase difference (IPD) is $\phi(\omega, t)$.

A key assumption in the MESSL signal model is that each time-frequency point is dominated by a single source. This implies the following approximations for the observed ILD and IPD:

$$\alpha(\omega, t) \approx 20 \log_{10} \frac{|\mathrm{H}_i^\ell(\omega)|}{|\mathrm{H}_i^r(\omega)|} \tag{5.4}$$

$$\phi(\omega, t) \approx \omega(\tau_i^\ell - \tau_i^r) \tag{5.5}$$

where i is the index of the particular source dominant at that cell, and thus depends on ω and t. These quantities have the advantage of being independent of the source signal, which is why the baseline MESSL model does not require any knowledge of the distribution of $x_i(t)$.

A necessary condition for the accurate modeling of the observation is that the interaural time difference (ITD) $\tau_i^\ell - \tau_i^r$ be much smaller than the window function used to calculate $\mathrm{X}_{IS}(\omega, t)$. In the experiments described in section 5.4, we use a window length of 64 ms and a maximum ITD of about 0.75 ms. Similarly, $\mathrm{h}_i^{\ell,r}(t)$ must be shorter than the window. This assumption does not generally hold in reverberation because a typical room impulse response has a duration of at least a few hundred milliseconds. However, we ignore this for the purposes of our model and note that effect of violating this assumption is to increase the variance in the ILD model. We model the ILD for source i as a Gaussian distribution whose mean and variance will be learned from the mixed signal:

$$P\big(\alpha(\omega, t) \,|\, i, \theta\big) = \mathcal{N}\big(\alpha(\omega, t);\, \upsilon_i(\omega),\, \eta_i^2(\omega)\big) \tag{5.6}$$

where θ stands for the otherwise unspecified model parameters.

The model for the IPD requires some additional considerations. It is difficult to learn the IPD for a given source directly from the mixed signal because $\phi(\omega, t)$ is only observed modulo 2π. This is a consequence of spatial aliasing that results at high frequencies if

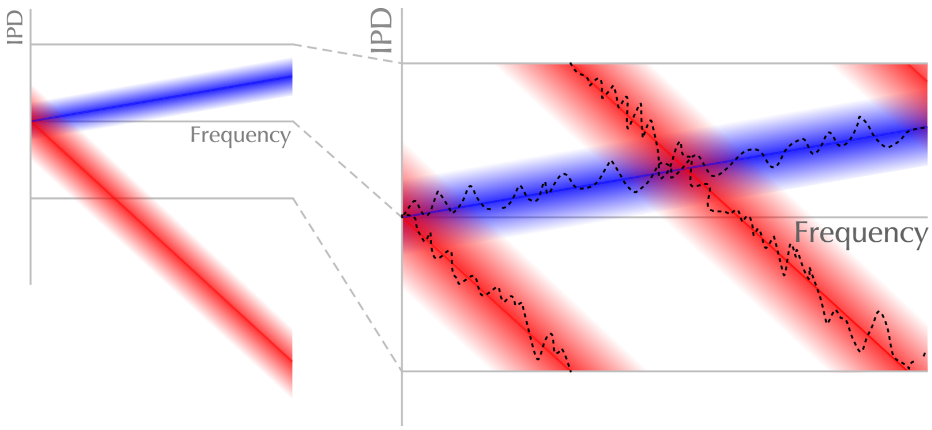


Figure 5.1: Illustration of spatial aliasing in our model of the interaural phase difference (IPD). The left pane shows the predicted IPD distribution for two distinct sources centered on their respective values of ω, t. The right pane demonstrates the observed IPDs for the two sources (dotted lines) with the distributions overlaid. The IPDs are observed modulo 2π due to the periodicity of the complex sinusoid in equation (5.3). For small interaural time difference (blue) this is not a problem. However, if the ITD is large (red), the IPD wraps around from $-\pi$ to π. This is especially problematic in mixtures because the wrapping results in additional ambiguity when the IPDs for the different sources intersect.

the ITD is large enough that $|\omega(\tau^\ell - \tau^r)| > \pi$ [125]. Because of this the observed IPD cannot always be mapped directly to a unique time difference. However, a particular ITD will correspond unambiguously to a single phase difference. This is illustrated in figure 5.1. This motivates a top down approach where the observed IPD will be tested against the predictions of a set of predefined time differences. The difference between the IPD predicted by an ITD of τ samples and the observed IPD is measured by the phase residual:

$$\tilde{\phi}_\tau(\omega, t) = \arg\left(e^{j\phi(\omega,t)}\, e^{-j\omega\tau}\right) \tag{5.7}$$

which is always in the interval $(-\pi, \pi]$. Given a predefined set of τs, the IPD distribution for a given source has the form of a Gaussian mixture model (GMM) with one mixture component for each time difference:

$$P\big(\phi(\omega,t)\,|\,i,\,\tau,\,\theta\big) = \mathcal{N}\big(\tilde{\phi}_\tau(\omega,t);\,0,\,\varsigma_i^2\big) \tag{5.8}$$

$$P\big(\phi(\omega,t),\,i\,|\,\theta\big) = \sum_\tau \psi_{i\tau} P\big(\tilde{\phi}_\tau(\omega,t)\,|\,i,\,\tau,\,\theta\big) \tag{5.9}$$

where $\psi_{i\tau} = P(i,\,\tau)$ are the mixing weights for source i and delay τ.

An example of the ILD and IPD observations used by the interaural model is shown in

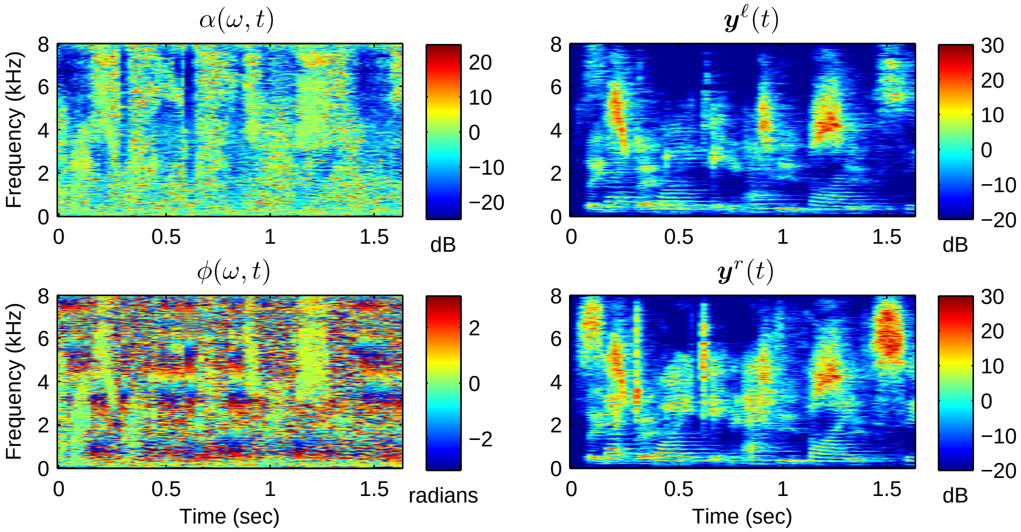


Figure 5.2: Observed variables in the MESSL-EV model derived from a mixture of two sources in reverberation separated by 60 degrees. The left column shows example ILD (top) and IPD (bottom) observations. The right column shows the left (top) and right (bottom) spectrograms modeled using the source model.

figure 5.2. The contributions of the two sources are clearly visible in both the ILD and IPD observations. The target source, which is located at an angle of 0° relative to the microphones, has an ILD close to zero at all frequencies while the ILD of the other source becomes increasingly negative at higher frequencies. This trend is typical of a source off to one side, since the level difference, which results from the "shadowing" effect of the head or baffle between the microphones, increases when the wavelength of the sound is small relative to the size of the baffle. Similarly, the IPD for the target source has an IPD close to zero at all frequencies while the IPD for the other source varies with frequency, with phase wrapping clearly visible at about 1, 3, and 5 kHz.

5.2.2 Source model

We extend the baseline MESSL model described above to incorporate prior knowledge of the source statistics. This makes it possible to capture the source-dependent portions of the observations that are divided out of the interaural spectrogram. We model the binaural observations directly:

$$y^\ell(\omega,t) \approx x_i(\omega,t) + h_i^\ell(\omega) \tag{5.10}$$

$$y^r(\omega,t) \approx x_i(\omega,t) + h_i^r(\omega) \tag{5.11}$$

where $x_i(\omega, t) \triangleq 20 \log_{10} |X_i(\omega, t)|$, and $y^\ell(\omega, t)$, $y^r(\omega, t)$, and $h_i(\omega)$ are defined analogously. An example of these observations derived from a mixture of two sources in reverberation is shown in the right column of figure 5.2.

Because the number of signal points observed is generally very small compared to the amount of data needed to reliably train a prior signal model describing the distribution of $\boldsymbol{x}_i(t)$, we take the subspace model approach and only learn the low dimensional adaptation parameters from the mixture. For simplicity we model this distribution using a Gaussian mixture model with diagonal covariances. The likelihood of a frame of the source signal $\boldsymbol{x}_i(t)$ can therefore be written as follows:

$$P\big(\boldsymbol{x}_i(t)\big) = \sum_c \pi_{ic}\, \mathcal{N}\big(\boldsymbol{x}_i(t);\, \boldsymbol{\mu}_c(\mathbf{w}_i),\, \Sigma_c(\mathbf{w}_i)\big) \tag{5.12}$$

where c indexes the different source mixture components (states), and $\pi_{ic} = P\big(c \,|\, i\big)$ are the mixing weights for source i and component c. As in previous chapters, we assume that the channel responses $h_i^{\ell,r}$ are constant across the entire mixture, i.e. the sources and the sensors remain stationary, and that they ares relatively smooth across frequency. The channel response is parametrized in the DCT domain, giving $\boldsymbol{h}_i^\ell = B\,\mathbf{h}_i^\ell$ where B is a matrix of DCT basis vectors. This allows $\boldsymbol{h}_i^{\ell,r}$ to be modeled using many fewer DCT coefficients than the number of frequency bands Ω. Note that in this chapter we do not assume that the channel response parameters $\mathbf{h}_i^{\ell,r}$ are subsumed into the subspace parameters $\mathbf{w}$. Because the channel responses at each ear are different we model them explicitly.

Combining this model of the channel response with the source model gives the following likelihoods for the left and right channel spectrograms:

$$P\big(y^\ell(\omega, t) \,|\, i,\, c,\, \theta\big) = \mathcal{N}\big(y^\ell(\omega, t);\, \mu_{ic}(\omega) + B(\omega, :)\,\mathbf{h}_i^\ell,\, \sigma_{ic}^2(\omega)\big) \tag{5.13}$$

$$P\big(y^r(\omega, t) \,|\, i,\, c,\, \theta\big) = \mathcal{N}\big(y^r(\omega, t);\, \mu_{ic}(\omega) + B(\omega, :)\,\mathbf{h}_i^r,\, \sigma_{ic}^2(\omega)\big) \tag{5.14}$$

where $B(\omega, :)$ is the row of B corresponding to frequency ω, $\boldsymbol{\mu}_{ic} = \boldsymbol{\mu}_c(\mathbf{w}_i)$, and $\sigma_{ic}^2 = \sigma_c^2(\mathbf{w}_i)$ as defined in section 3.4.3.

Combining the model of the interaural signals with the source model gives the complete likelihood of the model including the hidden variables:

$$\begin{aligned} &P\big(\phi(\omega, t),\, \alpha(\omega, t),\, y^\ell(\omega, t),\, y^r(\omega, t),\, i,\, \tau,\, c \,|\, \theta\big) \\ &\quad = P\big(i,\, \tau\big)\, P\big(\phi(\omega, t) \,|\, i,\, \tau,\, \theta\big)\, P\big(\alpha(\omega, t) \,|\, i,\, \theta\big)\, P\big(c \,|\, i\big)\, P\big(y^\ell(\omega, t) \,|\, i,\, c,\, \theta\big) \\ &\qquad\qquad P\big(y^r(\omega, t) \,|\, i,\, c,\, \theta\big) \end{aligned} \tag{5.15}$$

This equation explains each time-frequency point of the mixed signal as being generated by a single source i at a given delay τ using a particular component c in the source model. The graphical model corresponding to this factorization is shown in figure 5.3. This figure only includes the observations and those parameters that are estimated to match a particular mixture. We describe the estimation of the model parameters in the following section. For simplicity, parameters that remain fixed, including π_c and Σ_c, are omitted. It is also important to note that the figure depicts the full MESSL-EV model. If eigenvoice adaptation

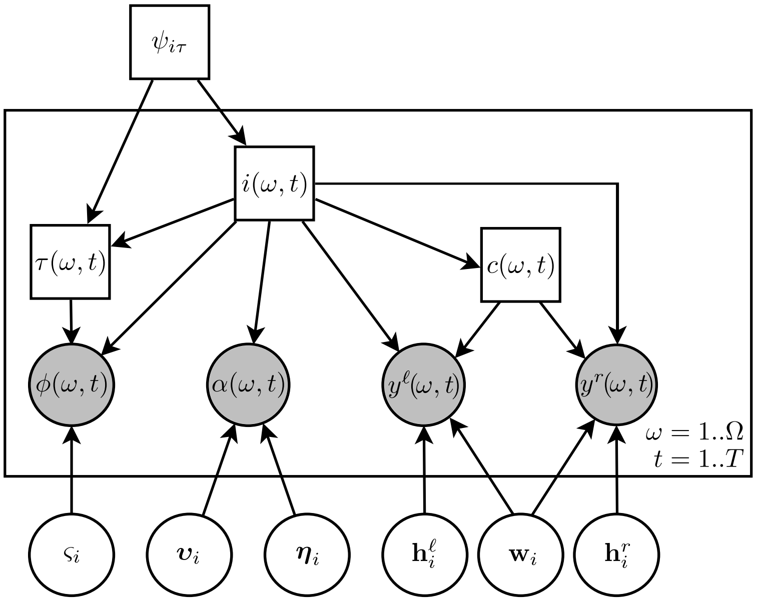


Figure 5.3: MESSL-EV graphical model. Each time-frequency point is explained by a source i, a delay τ, and a source model component c.

is not used then $\mathbf{w}_i$ is clamped to zero, and the model reduces to the original MESSL-SP model described in Weiss et al. [120].

Note that all time-frequency points are conditionally independent given the model parameters. The total likelihood of the observations can therefore be written as follows:

$$P(\boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{y}^\ell, \boldsymbol{y}^r \,|\, \theta) = \prod_{\omega t} \sum_{i\tau c} P(\phi(\omega,t), \alpha(\omega,t), y^\ell(\omega,t), y^r(\omega,t), i, \tau, c \,|\, \theta) \tag{5.16}$$

The combined model is essentially the product of three independent mixtures of Gaussians, corresponding to the IPD, ILD, and source models. For conciseness we will drop the (ω, t) where convenient throughout the remainder of this chapter.

5.3 Parameter estimation and source separation

The model described in the previous section can be used to separate sources because it naturally partitions the spectrogram into regions dominated by different sources. Given estimates of the source-specific model parameters $\theta = \{\psi_{i\tau}, \varsigma_i^2, \boldsymbol{\upsilon}_i, \boldsymbol{\eta}_i^2, \mathbf{w}_i, \mathbf{h}_i^\ell, \mathbf{h}_i^r\}$, the responsibilities at each time-frequency point can be easily computed. Similarly, given knowledge of the responsibilities, it is straightforward to estimate the model parameters. However, because neither of these quantities are generally known in advance, neither can be computed directly. We derive an expectation-maximization algorithm to iteratively learn both the parameters and responsibilities of time-frequency points for each source.

The E-step consists of evaluating the posterior responsibilities for each time-frequency point given θ_j, the estimated parameters for iteration j. We introduce a hidden variable representing the posterior of i, τ and c in a particular time-frequency cell:

$$z_{i\tau c}(\omega, t) = \frac{P\big(\phi,\, \alpha,\, y^\ell,\, y^r,\, i,\, \tau,\, c \,|\, \theta_j\big)}{\sum_{i\tau c} P\big(\phi,\, \alpha,\, y^\ell,\, y^r,\, i,\, \tau,\, c \,|\, \theta_j\big)} \tag{5.17}$$

This is easily computed using the factorization in equation (5.15).

The M-step consists of maximizing the expectation of the total log-likelihood given the current parameters θ_j:

$$\mathcal{L}\big(\theta \,|\, \theta_j\big) = k + \sum_{\omega t} \sum_{i\tau c} z_{i\tau c}(\omega, t)\, \log P\big(\phi,\, \alpha,\, y^\ell,\, y^r,\, i,\, \tau,\, c \,|\, \theta\big) \tag{5.18}$$

where k is a constant that is independent of θ.

The maximum likelihood model parameters are weighted means of sufficient statistics of the data. First, we define the operator

$$\langle x \rangle_{t,\tau} \triangleq \frac{\sum_{t,\tau} z_{i\tau c}(\omega, t) x}{\sum_{t,\tau} z_{i\tau c}(\omega, t)} \tag{5.19}$$

as the weighted mean over the specified variables, t and τ in this case, weighted by $z_{i\tau c}(\omega, t)$. The updates for the interaural parameters can then be written as follows:

$$\varsigma_i^2 = \left\langle \tilde{\phi}_\tau^2(\omega, t) \right\rangle_{\omega,t,\tau,c} \tag{5.20}$$

$$\upsilon_i(\omega) = \langle \alpha(\omega, t) \rangle_{t,\tau,c} \tag{5.21}$$

$$\eta_i^2(\omega) = \left\langle (\alpha(\omega, t) - \upsilon_i(\omega))^2 \right\rangle_{t,\tau,c} \tag{5.22}$$

$$\psi_{i\tau} = \frac{1}{\Omega T} \sum_{\omega t c} z_{i\tau c}(\omega, t) \tag{5.23}$$

Unlike the interaural parameters, the source model parameters are tied across frequency to ensure that each time frame is explained by a single component in the source prior. The updated parameters can be found by solving the following set of simultaneous equations for $\mathbf{w}_i, \mathbf{h}_i^\ell$, and $\mathbf{h}_i^r$:

$$\sum_{tc} U_c^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(2\left(\bar{\boldsymbol{\mu}}_c + U_c \mathbf{w}_i\right) + B\left(\mathbf{h}_i^r + \mathbf{h}_i^\ell\right)\right) = \sum_{tc} U_c^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(\boldsymbol{y}^\ell(t) + \boldsymbol{y}^r(t)\right) \tag{5.24}$$

$$\sum_{tc} B^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(\bar{\boldsymbol{\mu}}_c + U_c \mathbf{w}_i + B\, \mathbf{h}_i^\ell\right) = \sum_{tc} B^T M_{ict}\, \bar{\Sigma}_c^{-1}\, \boldsymbol{y}^\ell(t) \tag{5.25}$$

$$\sum_{tc} B^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(\bar{\boldsymbol{\mu}}_c + U_c \mathbf{w}_i + B\, \mathbf{h}_i^r\right) = \sum_{tc} B^T M_{ict}\, \bar{\Sigma}_c^{-1}\, \boldsymbol{y}^r(t) \tag{5.26}$$

where M_{ict} is a diagonal matrix whose diagonal entries correspond to a soft mask which encodes the posterior probability of component c from source i dominating the mixture at frame t:

$$M_{ict} \triangleq \operatorname{diag}\left(\sum_{\tau} z_{i\tau c}(:,t)\right) \tag{5.27}$$

The updates of equations (5.24) to (5.26) closely resemble the eigenvoice updates derived in section 4.2.3.2. The main differences here are that the posterior depends on frequency, hence the need for the mask term M_{ict}, and that $\mathbf{w}_i$ is tied across two observations, each of which have unique channel responses. The full derivation for all of the parameter updates can be found in section A.3.

This EM algorithm is guaranteed to converge to a local maximum of the likelihood surface, but because the total likelihood in equation (5.16) is not convex, the quality of the solution is sensitive to initialization. We initialize $\psi_{i\tau}$ from a cross-correlation based method while leaving all the other parameters in a symmetric, non-informative state. From those parameters, we compute the first E step mask. Using estimates of τ for each source from PHAT-histogram [1], $\psi_{i\tau}$ is initialized to be centered at each cross-correlation peak and to fall off away from that. Specifically, $P(\tau \mid i)$, which is proportional to $\psi_{i\tau}$, is set to be approximately Gaussian, with its mean at each cross correlation peak and a standard deviation of one sample. The remaining IPD, ILD, and source model parameters are estimated from the data in the M-step following the initial E-step.

It should be noted that initializing models with a large number of parameters requires some care to avoid source permutation errors and other local maxima. This is most important with regards to the ILD parameters $\boldsymbol{v}_i$ and $\boldsymbol{\eta}_i$ which are a function of frequency. To address this problem, we use a bootstrapping approach where initial EM iterations are performed with a frequency-independent ILD model, and frequency-dependence is gradually introduced. Specifically, for the first half of the total number of iterations, we tie all of the parameters across frequency. For the next iteration, we tie the parameters across two groups, the low and high frequencies, independently of one another. For the next iteration, we tie the parameters across more groups, and we increase the number of groups for subsequent iterations until in the final iteration, there is no tying across frequency and all parameters are independent of one another.

Figure 5.4 shows the interaural parameters estimated from the observations in figure 5.2 using the EM algorithm described in this section. The algorithm does a good job localizing the sources, as shown in the plot of $\psi_{i\tau}$. The ILD distribution (bottom left) accurately characterizes the true distribution as well. As described earlier, the ILD of the source facing the microphones head on (source 1), is basically flat across the entire frequency range while that of source 2 becomes more negative with frequency. Similarly, the per-source IPD distributions shown in the right hand column closely match the predictions made earlier. These distributions consist of a mixture of Gaussians calculated by marginalizing over all possible settings of τ as in equation (5.9). Since $\psi_{i\tau}$ contains non-zero probability mass for multiple τ settings near the correct location for each source, there is some uncertainty as to the exact source locations. The mixture components are spaced further apart at high frequencies because of their proportionality to ω. This is why the distributions are quite tight at low frequencies, but get gradually broader with increasing frequency.

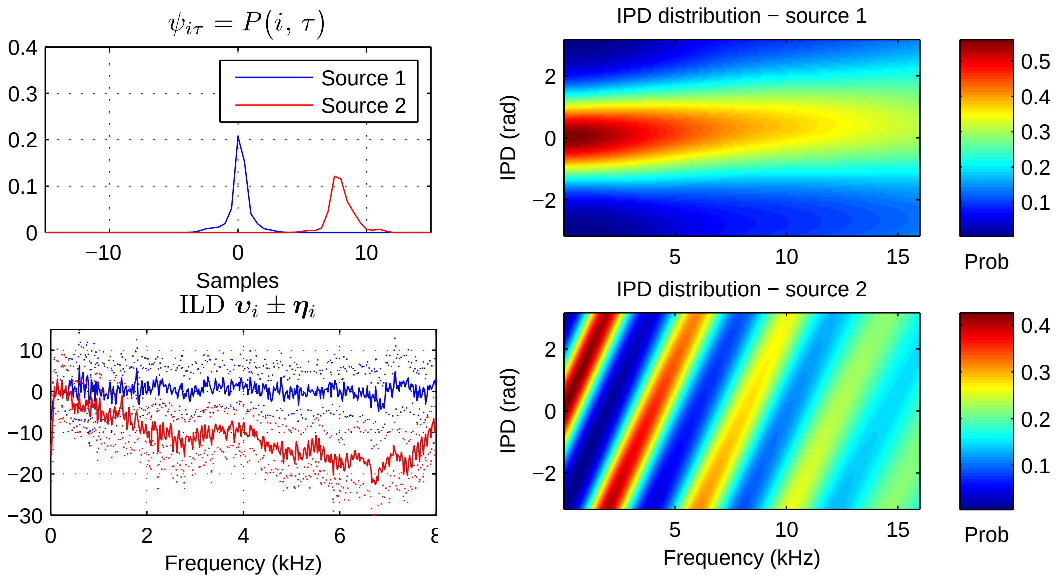


Figure 5.4: Interaural model parameters estimated by the EM algorithm given the observations shown in figure 5.2.

The estimated source model parameters are shown in figure 5.5. As with the ILD and IPD parameters, the source model parameters are initialized to an uninformative state. However, as the binaural cues begin to disambiguate the sources, the learned channel responses and source adaptation parameters help to differentiate the source models. By the time the algorithm has converged, the source models have become quite different, with $\mathbf{w}_i$ learning the characteristics unique to each source under the predefined eigenvoice model that are common to both left and right observations, e.g. the increased energy near 6 kHz in many components for source 2. Similarly, $\mathbf{h}_i^{\ell,r}$ learns the magnitude responses of the filters applied to each channel. Note that the overall shapes of $B\,\mathbf{h}_i^{\ell,r}$ reflect the effects of the HRTFs applied to the source in creating the mixture. These are unique to the particular mixture and were not present in the training data used to learn $\bar{\mu}$ and U. The difference between the channel response at each ear, $B\,\mathbf{h}_i^{\ell} - B\,\mathbf{h}_i^{r}$, reflects the same interaural level differences as the ILD parameters, $\boldsymbol{\upsilon}_i$ in figure 5.4.

Although the parameters learned by the MESSL model are interesting in their own right, they cannot separate the sources directly. After the EM algorithm converges, we derive a time-frequency mask from the posterior probability of the hidden variables for each source:

$$\mathrm{M}_i(\omega, t) = \sum_{\tau c} z_{i\tau c}(\omega, t) \tag{5.28}$$

Estimates of clean source i are then obtained by multiplying the short-time Fourier transform of each channel of the mixed signal by the mask for the corresponding source. This

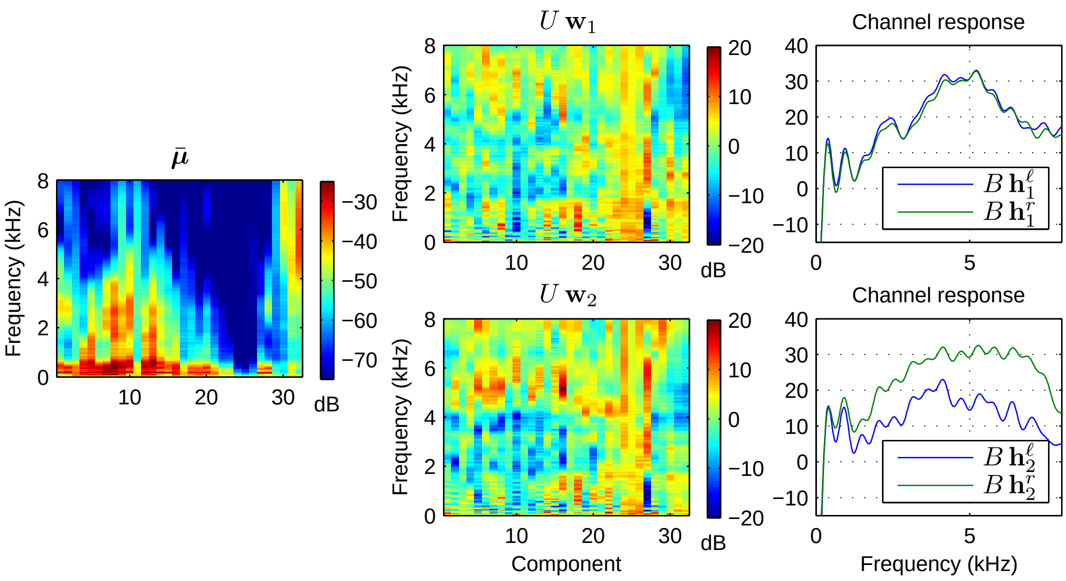


Figure 5.5: Source model parameters estimated by the EM algorithm given the observations shown in figure 5.2. The overall model for source i is the sum of the speaker-independent means, $\bar{\boldsymbol{\mu}}$, the source-adapted term $U\mathbf{w}_i$ based on the eigenvoice model of inter-speaker variability, and the channel response at each ear, $B\,\mathbf{h}_i^{\ell,r}$.

assumes that the mask is identical for both channels.

$$\hat{X}_i^\ell(\omega,t) = M_i(\omega,t)\,Y^\ell(\omega,t) \tag{5.29}$$

$$\hat{X}_i^r(\omega,t) = M_i(\omega,t)\,Y^r(\omega,t) \tag{5.30}$$

Figure 5.6 shows an example mask derived from the proposed algorithm. To demonstrate the contributions of the different types of observations in the signal model to the overall mask, we also plot masks isolating the IPD, ILD, and source models. These masks are found by leaving unrelated terms out of the factored likelihood and computing "marginal" posteriors. The full model is used to learn the parameters, but in the final EM iteration the contributions of each underlying model to the complete likelihood in equation (5.15) are normalized independently to compute three different posterior distributions.

The IPD and ILD masks make qualitatively different contributions to the final mask, so they serve as a good complement to one another. The IPD mask is most informative in low frequencies, and has characteristic subbands of uncertainty caused by the spatial aliasing described earlier. The ILD mask primarily adds information in frequencies above 2 kHz and so it is able to fill in many of the regions where the IPD mask is ambiguous. This poor definition in low frequencies is because the per-source ILD distributions shown in figure 5.4 have significant overlap below 2 kHz. These observations are consistent with the use of the ITD and ILD cues for sound localization in human audition [121].

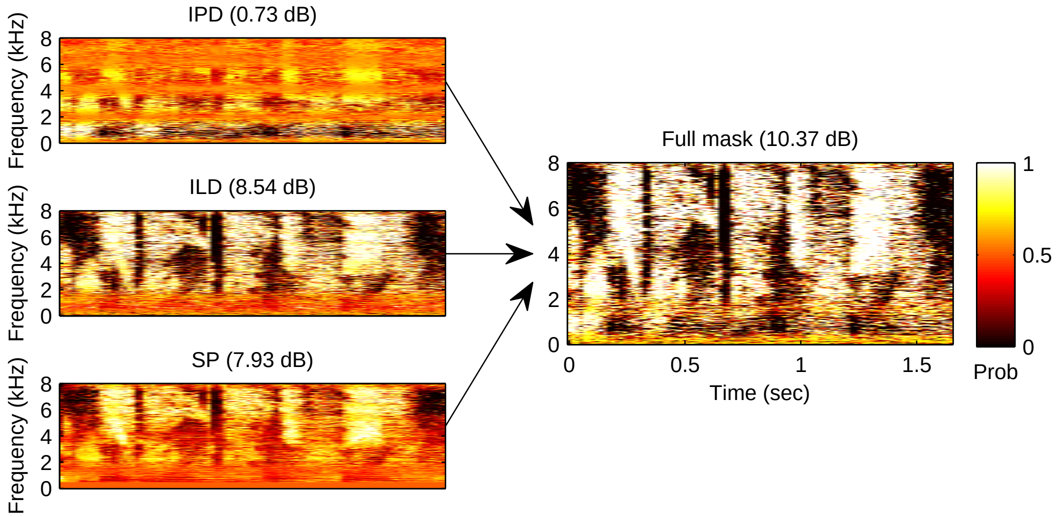


Figure 5.6: Contribution of the IPD, ILD, and source model to the final mask learned using the full MESSL-EV algorithm on the mixtures from figure 5.2. The SNR improvements for each mask computed using equation (5.31) are shown in parenthesis.

Finally, the source model mask is qualitatively quite similar to the ILD mask, with some additional detail below 2 kHz. This is not surprising because both the ILD and source models capture related features of the mixed signal. We expect that the additional constraints from the prior knowledge built into the source model should allow for more accurate estimation than the ILD model alone, however this is not clear from this figure because the ILD and SP masks are estimated jointly.

To better illustrate the contribution of the source model, figure 5.7 shows the mask estimated from the same data using the baseline MESSL algorithm of [67] which is based only on the interaural model. The MESSL mask is considerably less confident (i.e. less binary) than that of the MESSL-EV mask in figure 5.6. The contribution of the IPD mask is quite similar in both cases. The difference in quality between the two systems is a result of the marked difference in the ILD contributions. The improvement in the MESSL-EV case can be attributed to the addition of the source model, which, although not as informative on its own, is able to indirectly improve the estimation of the ILD parameters. This is because the source model introduces correlations across frequency that are only loosely captured by the ILD model during initial iterations. This is especially true in the higher frequencies which are highly correlated in speech signals. By modeling each frame with GMM components with a different spectral shape, the source model is able to decide which time-frequency regions are a good fit to each source based on how well the observations in each frame match the source prior distribution. It is able to isolate the sources based on how speech-like they are, using prior knowledge such as the high-pass shape characteristic of fricatives and characteristic resonance structure of vowels, etc. In contrast, the ILD

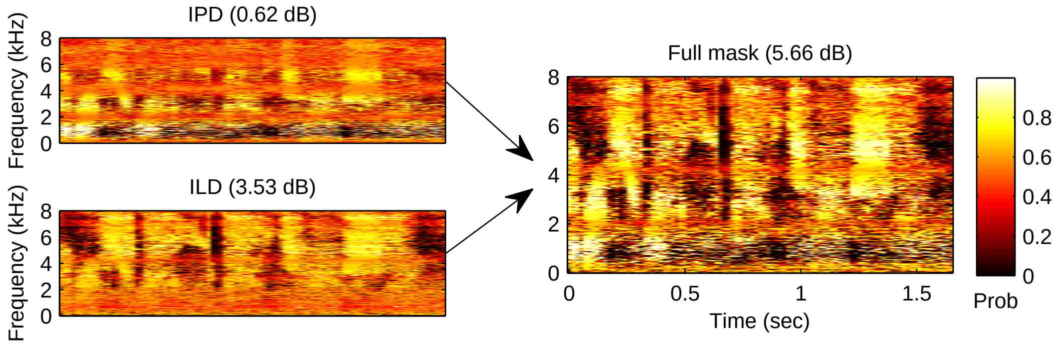


Figure 5.7: Contribution of the IPD and ILD to the final mask learned using the baseline MESSL separation algorithm using only the interaural signal model on the mixtures from figure 5.2. The SNR improvement computed using equation (5.31) is shown in parenthesis.

model treats each frequency band independently and is prone to source permutations if poorly initialized. Although the bootstrapping process described earlier alleviates these problems to some extent, the source model's ability to emphasize time-frequency regions consistent with the underlying speech model further reduces this problem and significantly improves the quality of the interaural parameters and thus the overall separation.

5.4 Experiments

In this section we describe a set of experiments designed to evaluate the performance of the proposed algorithm under a variety of different conditions and compare it to two other well known binaural separation algorithms. We assembled a data set consisting of mixtures of two and three speech signals in simulated anechoic and reverberant conditions. The mixtures were formed by convolving anechoic speech utterances with a variety of different binaural impulse responses. We formed two such data sets, one from utterances from the GRID corpus [16] for which training data was available for the source model, and another using the TIMIT corpus [34] to evaluate the performance on held out speakers using the GRID source models. Although the TIMIT data set contains speech from hundreds of different speakers, it does not contain enough data to adequately train models for each of these speakers. This makes it a good choice for evaluation of the eigenvoice adaptation technique. In both cases, we used a randomly selected subset of 15 utterances to create each test set. Prior to mixing, the utterances were passed through a first order pre-emphasis filter to whiten their spectra to avoid overemphasizing the low frequencies in our SNR performance metric.

The anechoic binaural impulse responses came from [4], a large effort to record head-related transfer functions for many different individuals. We used the measurements for a KEMAR dummy head with small ears, taken at 25 different azimuths at 0° elevation.

The reverberant binaural impulse responses were recorded by Shinn-Cunningham et al. [99] in a real classroom with a reverberation time of around 565 ms. These measurements were also made with a KEMAR dummy head, although a different 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 synthesized mixtures, 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, defined for source i as follows:

$$\mathrm{SNRI}_i = 10\log_{10}\frac{\|\mathrm{M}_i\mathrm{X}_i\|^2}{\|\mathrm{X}_i - \mathrm{M}_i\sum_j \mathrm{X}_j\|^2} - 10\log_{10}\frac{\|\mathrm{X}_i\|^2}{\|\sum_{j\neq i}\mathrm{X}_j\|^2} \tag{5.31}$$

where X_i is the clean spectrogram for source i, M_i is the corresponding mask estimated from the mixture, and $\|\cdot\|$ is the Frobenius norm operator. This measure penalizes both noise that is passed through the mask and signal that is rejected by the mask.

We also evaluate the speech quality of the separations using the Perceptual Evaluation of Speech Quality (PESQ) [64, Sec. 10.5.3.3]. This measure is highly correlated with the Mean Opinion Score (MOS) of human listeners asked to evaluate the quality of speech examples. MOS ranges from -0.5 to 4.5, with 4.5 representing the best possible quality. Although it was initially designed for use in evaluating speech codecs, PESQ can also be used to evaluate speech enhancement systems.

We compare the proposed separation algorithms to the two-stage frequency-domain blind source separation system from [96] (2S-FD-BSS), the Degenerate Unmixing Estimation Technique from [47, 125] (DUET), and the performance using ground truth binary masks derived from oracle knowledge of the clean source signals. We also compare three variants of our system: the full MESSL-EV algorithm described in this chapter, the MESSL-SP algorithm from [120] that uses a speaker-independent source prior distribution (identical to MESSL-EV but with $\mathbf{w}_i$ fixed at zero), and the baseline MESSL algorithm from [67] that does not utilize source constraints at all. The MESSL-SP system uses a 32 mixture component speaker-independent model trained over data from all 34 speakers in the GRID data set. Similarly, the MESSL-EV system uses a 32 component eigenvoice speech model source GMMs trained over all 34 speakers using the procedure described in chapter 3. All 33 eigenvoice bases were retained. Figure 5.8 shows example masks derived from these systems.

DUET creates a two-dimensional histogram of the interaural level and time differences observed over an entire spectrogram. It then smooths the histogram and finds the I largest peaks, which should correspond to the I sources. DUET assumes that the interaural level and time difference are constant at all frequencies and that there is no spatial aliasing, conditions which can be met to a large degree with free-standing microphones close to

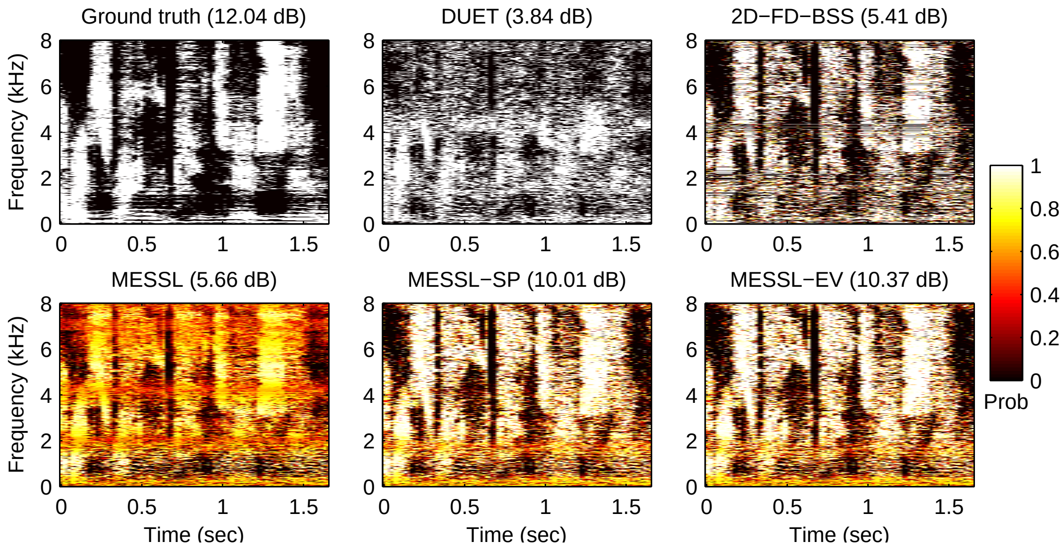


Figure 5.8: Example binary masks found using the different separation algorithms evaluated in section 5.4. The mixed signal is composed of two GRID utterances in reverberation separated by 60 degrees.

one another. With dummy head recordings, however, the interaural level difference varies with frequency and the microphones are spaced far enough apart that there is spatial aliasing above about 1 kHz. Frequency-varying ILD scatters observations of the same source throughout the histogram as does spatial aliasing, making the sources more difficult to isolate. As shown in figure 5.8, this manifests itself as poor estimation in frequencies above 4 kHz which the algorithm overwhelmingly assigns to a single source, and errors due to spatial aliasing in subbands around 2 and 4 kHz.

The 2S-FD-BSS system uses a combination of ideas from model-based separation and independent component analysis (ICA) that can separate underdetermined mixtures. In the first stage, blind source separation is performed on each frequency band of a spectrogram separately using a probabilistic model of mixing coefficients. In the second stage, the sources in different bands are associated with the corresponding signals from other bands using k-means clustering on the posterior probabilities of each source and then further refined by matching sources in each band to those in nearby and harmonically related bands. The first stage encounters problems when a source is not present in every frequency and the second encounters problems if sources' activities are not similar enough across frequency. Permutation errors are visible in the 2S-FD-BSS mask shown in figure 5.8 in subbands around 2 and 4 kHz, regions that are ambiguous due to spatial aliasing. In general, such errors tend to happen at low frequencies, where adjacent bands are less well-correlated. In contrast, the failure mode of the MESSL variants is to pass both sources equally when it is unable to sufficiently distinguish between them. This is clearly visible in the regions of the MESSL mask in figure 5.8 that have posteriors close to 0.5. As a result 2S-FD-BSS is more prone to source permutation errors where significant target energy can

System	A2	R2	A3	R3	Avg
Ground truth	11.83	11.58	12.60	12.26	12.07
MESSL-EV	8.79	**7.85**	**8.20**	**7.54**	**8.09**
MESSL-SP	6.30	7.39	7.08	7.18	6.99
MESSL	7.21	4.37	6.17	3.56	5.33
2S-FD-BSS	**8.91**	6.36	7.94	5.99	7.30
DUET	2.81	0.59	2.40	0.86	1.67

Table 5.1: Average SNR improvement (in dB) across all distractor angles on mixtures created from the GRID data set. 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).

System	A2	R2	A3	R3	Avg
Ground truth	3.41	3.38	3.10	3.04	3.24
MESSL-EV	**3.00**	**2.65**	**2.32**	**2.24**	**2.55**
MESSL-SP	2.71	2.62	2.22	2.22	2.44
MESSL	2.81	2.39	2.15	1.96	2.33
2S-FD-BSS	2.96	2.50	2.28	2.04	2.44
DUET	2.56	2.03	1.85	1.53	1.99
Mixture	2.04	2.04	1.60	1.67	1.84

Table 5.2: Average PESQ score (mean opinion score) across all distractor angles on mixtures created from the GRID data set.

be rejected by the mask.

5.4.1 GRID performance

The average performance of the evaluated algorithms on the GRID data set is summarized in tables 5.1 and 5.2 using the SNR improvement and PESQ metrics, respectively. Under both metrics all algorithms perform better in anechoic conditions than in reverberation and on mixtures of two sources than on mixtures of three sources. In most cases MESSL-EV performs best, followed by MESSL-SP and 2S-FD-BSS. 2S-FD-BSS outperforms MESSL-SP in anechoic conditions, however, in reverberation, this trend is reversed and 2S-FD-BSS performs worse. Both of the MESSL variants perform significantly better than the MESSL baseline for the reasons described in the previous section. The addition of speaker adaptation in MESSL-EV gives an overall improvement of about 1.1 dB over MESSL-SP and 2.8 dB over MESSL in SNR improvement on average. 2S-FD-BSS generally performs better than MESSL, but not as well as MESSL-SP and MESSL-EV. The exception is on mixtures of two sources in anechoic conditions where 2S-FD-BSS performs best overall in terms of SNR improvement. Finally, DUET performs worst, especially in reverberation where the IPD/ILD histograms are more diffuse, making it difficult to accurately localize the sources.

We note that unlike the initial results reported in [120] MESSL-SP does not perform worse than MESSL on anechoic mixtures. The problems in [120] were caused by the channel parameters $\mathbf{h}_i^{\ell,r}$ over-fitting which led to source permutations. To fix this problem in the results reported here, we used a single, flat channel basis for the channel parameters in anechoic mixtures. In reverberant mixtures 30 DCT bases were used.

The poor performance of some of the MESSL systems in table 5.1 on anechoic mixtures is a result of poor initialization at small distractor angles. An example of this effect can be seen in the left column of figure 5.9 where the MESSL systems have very poor performance compares to 2S-FD-BSS when the sources are separated by 5 degrees. However, as the sources get further apart, the performance of all of the MESSL systems improves dramatically. The very poor performance at very small angles heavily skews the averages in table 5.1. This problem did not affect MESSL's performance on reverberant mixtures because the minimum separation between sources on that data was 15 degrees and the initial localization used to initialize MESSL was adequate. Finally, 2S-FD-BSS was unaffected by this problem at small distractor angles because, unlike the other systems we evaluated, it does not directly utilize the spatial locations for separation.

MESSL, 2S-FD-BSS, and DUET all perform significantly better on anechoic mixtures that on reverberant mixtures because the lack of noise from reverberant echoes makes anechoic sources much easier to localize. As described in the previous section, the additional constraints from the source models in MESSL-EV and MESSL-SP help to resolve the ambiguities in the interaural parameters in reverberation so the performance of these systems does not degrade nearly as much. In reverberation MESSL-EV and MESSL-SP both improve over the MESSL baseline by over 3 dB. The added benefit from the speaker adaptation in MESSL-EV is limited in reverberation, but is significant in anechoic mixtures. This is likely a result of the fact that the EV model has more degrees of freedom to adapt to the observation. The MESSL-SP system can only adapt a single parameter per source in anechoic conditions due the limited model of channel variation described above. Finally, we note that the addition of the source model in MESSL-EV and MESSL-SP is especially useful in underdetermined conditions (i.e. A3 and R3) because of the source model's ability to emphasize time-frequency regions consistent with the underlying speech model which would otherwise be ambiguous. In two source mixtures this effect is less significant because the additional clean glimpses of each source allow for more robust estimation of the interaural parameters.

5.4.2 TIMIT performance

Tables 5.3 and 5.4 show the performance of the different separation algorithms on the data set derived from TIMIT utterances. The trends are very similar to those seen in the GRID data set, however performance in general tends to be a bit better in terms of SNR improvement. This is probably because the TIMIT utterances are longer than the GRID utterances, and the additional observations lead to more robust localization which in turn leads to better separation. The main point to note from the results in table 5.3 is that the performance improvement of MESSL-EV over the other MESSL variants is significantly reduced when compared to the GRID experiments. This is because of the mismatch between the mixtures and the data used to train the models. However, despite

System	A2	R2	A3	R3	Avg
Ground truth	12.09	11.86	12.03	11.84	11.95
MESSL-EV	10.08	**8.36**	**8.21**	**7.22**	**8.47**
MESSL-SP	10.00	8.10	7.97	6.96	8.26
MESSL	9.66	5.83	7.12	4.32	6.73
2S-FD-BSS	**10.29**	7.09	6.17	4.86	7.10
DUET	3.87	0.59	3.63	0.62	2.18

Table 5.3: Average SNR improvement (in dB) across all distractor angles on mixtures created from the TIMIT data set. 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).

System	A2	R2	A3	R3	Avg
Ground truth	3.35	3.33	3.06	3.02	3.19
MESSL-EV	2.99	**2.52**	**2.30**	**2.11**	**2.48**
MESSL-SP	2.98	2.50	2.28	2.10	2.47
MESSL	2.92	2.33	2.24	1.96	2.36
2S-FD-BSS	**3.07**	2.36	1.91	1.76	2.28
DUET	2.59	1.85	2.01	1.48	1.98
Mixture	1.96	1.92	1.53	1.62	1.76

Table 5.4: Average PESQ score (mean opinion score) across all distractor angles on mixtures created from the TIMIT data set.

this mismatch, the performance improvement of the MESSL variants that incorporate a prior source model still show a significant improvement over MESSL.

The performance of MESSL-EV relative to the other MESSL variants on both data sets is compared in table 5.5. On the matched data set, MESSL-EV outperforms MESSL by an average of about 2.8 dB. It also outperforms MESSL-SP by an average of 1.1 dB. However, on the mismatched data set the improvement of MESSL-EV is significantly reduced. In fact, the improvement of MESSL-EV over MESSL-SP on this data set is only 0.2 dB on average. This implies that the eigenvoice model of speaker variation is significantly less informative when applied to speakers that are very different from those in the train set. The bulk of MESSL-EV's improvement is therefore due to the speaker-independent portion of the model which is still a good enough model for speech signals in general to improve performance over MESSL, even on mismatched data.

The small improvement in the performance of MESSL-EV when the training and test data are severely mismatched is the result of a number of factors. The primary problem is that a relatively small set of speakers was used to train the GRID eigenvoice bases. As described in chapter 4, in order to adequately capture the full subspace of speaker variation and generalize well to held-out speakers, data from a large number of training speakers, on the order of a few hundred, is typically required. In these experiments, training data was only available for 34 different speakers.

Data set	System 1 – System 2	A2	R2	A3	R3	Avg
GRID	MESSL-EV – MESSL	1.58	3.46	2.03	3.98	2.76
	MESSL-EV – MESSL-SP	2.49	0.48	1.12	0.36	1.10
TIMIT	MESSL-EV – MESSL	0.42	2.53	1.09	2.89	1.74
	MESSL-EV – MESSL-SP	0.08	0.26	0.24	0.26	0.21

Table 5.5: Comparison of the relative performance in terms of dB SNR improvement of MESSL-EV to MESSL-SP and the MESSL baseline on both the GRID data set where the source models are matched to the test data, and on the TIMIT data set where the source models are mismatched to the test data.

This lack of diversity in the training data is especially relevant because of the significant differences between the GRID and TIMIT speakers. The speakers in the GRID data set were all speaking British English while TIMIT consists of a collection of American speakers. There are significant pronunciation differences between the two dialects, e.g. British English is generally non-rhotic, which lead to signification differences in the acoustic realizations of common speech sounds and therefore differences between the corresponding speech models. These differences make it impossible to fully capture the nuances of the both dialects without including some speakers of both dialects in the training set. Finally, the likelihood that the eigenvoice model will generalize well to capture speaker-dependent characteristics across both data sets is further decreased because the models themselves were quite small, consisting of only 32 mixture components.

5.4.3 Performance at different distractor angles

Finally, the results on the TIMIT set are shown as a function of distractor angle in figure 5.9. Performance of all algorithms generally improves when the sources are better separated in space. In anechoic mixtures of two sources the MESSL variants all perform essentially as well as ground truth masks when the sources are separated by more than 40°. None of the systems are able to approach ideal performance under the other conditions. As noted earlier, 2S-FD-BSS performs best on 2 source anechoic mixtures in tables 5.1 and 5.3. As seen in figure 5.9 this is mainly an effect of very poor performance of the MESSL systems on mixtures with small distractor angles. All MESSL variants outperform 2S-FD-BSS when they are separated by more than about 20°. The poor performance of MESSL when the sources are separated by 5° is a result of poor initialization due to the fact that localization is difficult because the parameters for all sources are very similar. This is easily solved by using better initialization. In fact, it is possible to effectively combine the strengths of both of the ICA and localization systems by using the mask estimated by 2S-FD-BSS to initialize MESSL. This would require starting the separation algorithm with the M-step instead of the E-step as described in section 5.3, but the flexibility of MESSL's EM approach allows this. We leave the investigation of the combination of these techniques as future work.

This dependence on spatial localization for adequate source separation highlights a disadvantage of the MESSL family of algorithms, especially as compared to model-based

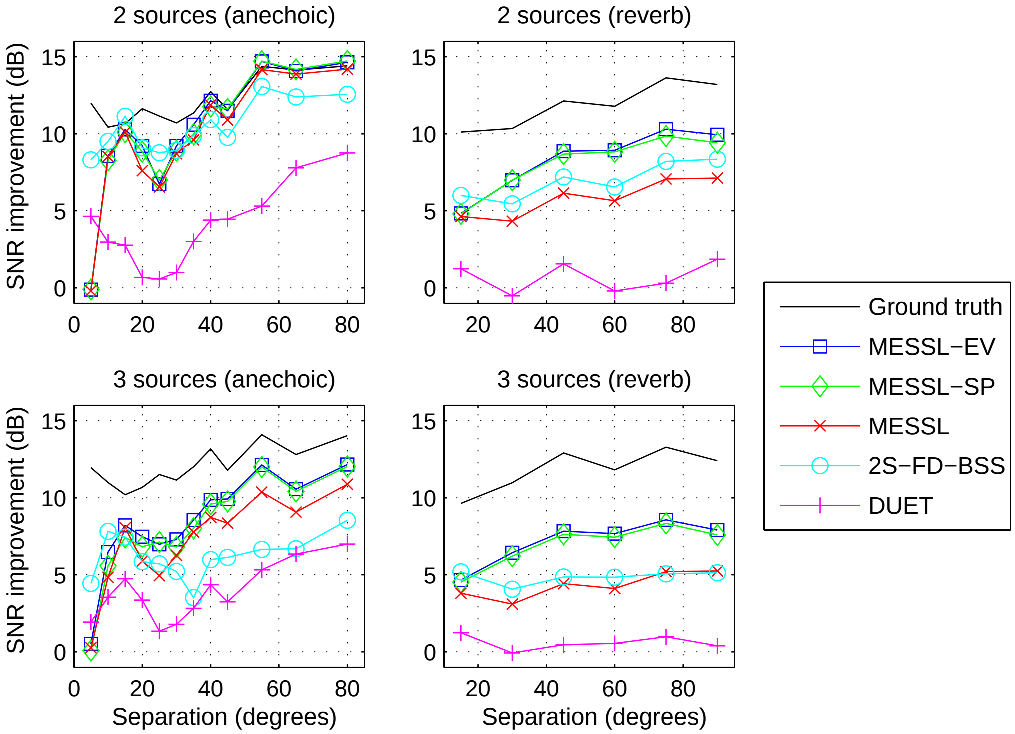


Figure 5.9: Separation performance on the TIMIT data set as a function of distractor angle.

binaural separation algorithms that use factorial model combination [82, 122]. As seen in the examples of figures 5.6 and 5.7, in MESSL-SP and MESSL-EV the source model is used to help disambiguate uncertainties in the interaural localization model. It does not add any new information about the interaction between the two sources and can only offer incremental improvements over the MESSL baseline. Therefore the addition of the source model does not improve performance when the sources are located very close to each other in space.

In contrast, in Rennie et al. [82] and Wilson [122], the factorial source model is used to model the interaction between the sources directly. In these algorithms, the localization cues are used to disambiguate the source model, which, on its own is inherently ambiguous because identical speaker-independent models are used for all sources. This makes it impossible for the models to identify portions of the signal that are dominated by each source without utilizing the fact that they arrive from distinct spatial locations. These algorithms would therefore suffer from similar problems to MESSL at very small distractor angles where the localization cues are similar for all sources. However, this could be overcome by incorporating additional knowledge about the differences between the distributions of each source signal through the use of speaker-dependent models, or model adaptation as described in this thesis. When the sources are close together, the binaural separation problem reduces to that of monaural separation, where factorial model based techniques using source-dependent or -adapted models have been very successful [118]. MESSL-EV, however, still suffers at small distractor angles despite utilizing source-specific models.

The advantage of MESSL-EV over the factorial model approach to combining source models with localization cues is that it enables efficient inference because it is not necessary to evaluate all possible model combinations. This becomes especially important for dense mixtures of many sources. As the number of sources grows, the factorial approach scales exponentially in terms of the number of Gaussian evaluations required [92]. In contrast, the computational complexity of the algorithms described in this paper scale linearly in the number of sources.

5.5 Summary

We have presented a system for source separation based on a probabilistic model of binaural observations. We combine a model of the interaural spectrogram with a prior model of the source spectrogram. This model is able to obtain a significant performance improvement over the algorithm that does not rely on a prior source model and over another state of the art source separation algorithm based on frequency domain ICA. The improvement is substantial even when the prior on the source statistics is quite limited, consisting of a small speaker-independent model. The addition of source adaptation based on a speech subspace model is shown to further improve performance. As seen in the experimental results in previous chapters, the performance improvements when using source adaptation are largest when the test data comes from the same sources as were used to train the model. However, even when the training and test data are mismatched, the addition of source adaptation is able to boost performance by a small amount.

Chapter 6

Conclusion

6.1 Thesis summary

We have presented a novel speaker-adaptive speech signal model and demonstrated its utility in a number of underdetermined source separation applications. The model is based on the idea of eigenvoices, bases for GMM or HMM parameters learned using PCA, first proposed for speaker adaptation in automatic speech recognition. Under this model the set of model parameters specific to a given speaker can be accurately represented as a weighted linear combination of the eigenvoice bases. Therefore, adapting the model to a given source is simply a matter of using the observed signal to estimate these weights. The resulting speaker subspace model is quite extensible, allowing both Gaussian mean and covariance parameters to be adapted to match a particular source as well as naturally allowing for the use of additional bases to compensate for any unknown channel effects that are independent of the speakers present in the mixed signal.

We developed a set of algorithms for joint speaker adaptation and source separation given a monaural speech mixture that are extended to the popular model-based approach to source separation described in chapters 2 and 4. In general the use of the proposed adaptation model allows for a significant improvement in separation performance when compared to separation using unadapted, source-independent models. However, the proposed system does not always perform as well as a system that uses predefined speaker-dependent models, largely because it is prone to permutations between sources. Despite these shortcomings, we show that proposed model can also generalize well to previously unseen speakers. This removes the unrealistic requirement that training data for all possible speakers be available in advance and therefore makes the model-based approach to monaural source separation more useful in practice. Furthermore, it is possible to combine the proposed model adaptation approach with the model selection method described in section 4.2.1 to obtain good performance in both matched and mismatched conditions by using the latter method to initialize the former.

Finally, we presented a system that extends the use of statistical source models to two-channel underdetermined source separation. The system is based on a probabilistic model of binaural observations that combines a model of the interaural spectrogram with a prior model of the source spectrogram. By learning the frequency response of the combined room impulse response and head-related transfer function, i.e. the channel response, applied to each source signal it is possible to obtain a significant performance improvement over the algorithm that does not utilize a prior source model. The improvement is significant even when the prior on the source statistics is quite limited, consisting of a small speaker-independent model. The addition of source adaptation based on the proposed speaker subspace model further improves performance, although the improvement is not as large as it is in the monaural case. Also, as in the monaural case, the use of the speaker subspace model generalizes well to mixtures composed from speakers not present in the training set. However, the improvement over the speaker-independent model is not as large under such conditions due to the mismatch with the training data.

6.2 Future work

The experimental results described in this thesis establish the use of model adaptation as a powerful extension to the popular methods for statistical model-based source separation. The applications of supervised source separation remain somewhat limited due to the requirement that the identities of the sources comprising the mixture be known in advance and that training data be available for them. The adaptation model described in this thesis partially alleviates these requirements, but it still requires that the class of signal (e.g. speech or music) be known and that training data be available for a subset of such sources, even if they do not exactly match the sources in the mixture.

One approach to alleviating this requirement and making the model-based approach still more general would be to use an extension of the model selection idea described in section 3.3.1 to select models of source *classes*. This could be done by training a set of signal subspace models similar to the speaker subspace model utilized in this thesis on a variety of different source types (e.g. speech, tonal musical instruments, percussive instruments, etc.) and then selecting the models that best match the observations from this set prior to separation. This would essentially discretize the space of all possible source signals (i.e. not only speech) into distinct *subspaces* of related sources that are naturally represented using similar model topologies. For example, it would be difficult to create a single subspace model that accurately captures both speech and music due to the significant differences in their structure (e.g. speech signals, even when produced by different talkers, share many similarities in the form of common phoneme structures that have no analogies in most music signals). By coarsely discretizing a broader space of audio signals into different classes, yet still allowing continuous variation within each subspace it should be possible to extend the ideas described in this thesis to work on a broader class of sources.

The performance of such a general model-based separation system still depends on the quality of its constituent signal models. Based on the results reported in section 4.4, it is clear that even the more constrained problem of monaural speech separation is not yet solved in general. While our results on this data set could be improved through the use of a more sophisticated channel model and larger speech models to more accurately cover the full space of signals found in the data set, it is still not clear how far the statistical model-based separation paradigm can realistically be taken. As the source signals become more complex, the supervised approach becomes more impractical because of the increased amount of training data needed to train large models. Even given enough data to train large source-dependent models, at some point the models themselves become large enough that inference becomes infeasible. The experiments reported in this thesis have already approached this threshold using the factorial inference algorithms described in chapter 4. Very recently, Rennie et al. [86] proposed a factorial HMM inference algorithm significantly faster than that described in chapter 4. This could enable separation based on significantly larger source models, potentially on the scale of thousands of states, which would make monaural separation of spontaneous speech possible using the algorithms described in chapter 4.

One approach that has been proposed to limit the complexity of such speech models is to model the high frequency resolution pitch structure (corresponding to the excitation of the vocal folds) independently of the more coarse envelope structure (corresponding to the formant resonances of the vocal tract) [53]. Such a factored model could use fewer states than the equivalent monolithic model. A similar model might also be appropriate for music signals where instrument timbre and pitch can be treated independently. The speaker subspace model used in this work does implicitly decouple the pitch and envelope structure of speech to some extent. This can be seen in figure 3.5 where the first two eigenvoice bases primarily contain information about the course structure, whereas the third contains significant pitch information. However, it does not treat the pitch and envelope as being completely independent and thus does not use the relationship to reduce the number of states in the model. An extension of the subspace model to explicitly take advantage of this structure remains to be explored.

Another direction worth investigating is the extension of the subspace adaptation approach to other types of source models. For example, the supervised NMF approach to monaural source separation [97] suffers from the same problems as the standard VQ/HMM approach in terms of requiring that the source identities and their corresponding NMF bases be known in advance. This requirement can be loosened in a similar manner to that described in this thesis for HMM source models. As described in chapter 2, in supervised NMF source separation, the observed mixture Y is represented as the product of a set of source bases, B_i, and a matrix representing their temporal structure, S:

$$Y = \begin{bmatrix} B_1 & B_2 & \dots & B_I \end{bmatrix} S \tag{6.1}$$

If B_i is known to be contained in a predefined subspace then it can be represented as a linear combination of the subspace bases in the same way that we have constrained HMM parameters, i.e. basis c for speaker i can be written as $B_{ic} = U_c \mathbf{w}_i$ as in equation (3.8). Separation under such a subspace NMF model would require estimating the per-source adaptation parameters $\mathbf{w}_i$ and the temporal structure matrix S. The only predefined portion of the model is the set of subspaces bases which is not specific to any source. An advantage of the NMF approach over the factorial HMM model from chapter 4 is that factorization of equation (6.1) can be found with an algorithm that scales linearly with the number of basis vectors [59].

The NMF approach to source separation has another advantage over the models described in this thesis in that it uses a continuous model of the mixing process. The assumption that sources rarely overlap in time and frequency (see figure 2.1), which is the basis of the models described in this thesis, does not always hold. For example, sources in a typical polyphonic music mixture are, by design, synchronized in time and frequency. This leads to significant overlap in frequency of harmonics of notes with related pitch, and in time due to the underlying rhythmic structure. The use of a continuous mixing model such as NMF or Algonquin [56] would enable more accurate separation of such mixtures.

It is also worth noting that all of the models described in this thesis assume that the number of source models I is known in advance. These systems can be extended to simultaneously estimate I as well as the underlying source signals by hypothesizing multiple settings of I and using the Bayesian information criterion [10, Chapter 4] to select the best setting. An similar assumption made throughout this thesis is that the sources remain stationary, implying that the channel responses are constant for the duration of the signal. This is especially relevant to binaural separation based on source localization described in chapter 5. Extending the proposed models to compensate for time-varying channel effects through the use of dynamic Bayesian networks remains future work.

Finally, it might be possible to derive alternate subspace bases or adaptation algorithms with the explicit goal of making it easier to distinguish between sources. Jang and Lee [46] describe the idea of improving the performance of their supervised ICA separation algorithm on speech mixtures by learning the predefined source bases in a discriminative manner. This idea is worth exploring in the context of our adaptation model as well. For example, it might be possible to more easily estimate the subspace parameters for each source when using speaker subspace bases that were learned in a way that inherently separates them, e.g. using a method like LDA. Similarly, it might be possible to derive alternate adaptation algorithms to those described in section 4.2 that better separates sources. This might involve estimating the subspace parameters using a constrained optimization of an objective function that incorporates a metric of the confusibility of the different source models (e.g. the Kullback-Leibler divergence) into the model log likelihood from equation (4.41).

In conclusion we note that while many of the extensions we have described in this section would be quite useful for monaural separation, the same concerns are not necessarily as important for the multichannel problem. When more channels are available there are many constraints that can be used to separate sources without requiring any prior knowledge. As we showed in chapter 5, such source-independent constraints can be effectively combined with those derived from pretrained source models to improve performance. Of course, the most generally useful source separation algorithm will leverage as much information as it can to improve separation quality. The ideas presented in this thesis make the use of source-specific models to help constrain possible source reconstructions more generally applicable, and represent a step in the direction of enabling source separation systems to utilize similar top-down constraints to those used in human audition in the form of familiarity with a particular speaker.

Appendix A

Appendix

A.1 Notation

This section reviews the notational conventions used throughout this thesis.

Signals

$\mathrm{x}_i(t)$

Time domain signal produced by source i.

$\mathrm{X}_i(\omega, t)$

Point in time-frequency of the short-time Fourier transform (STFT) of $\mathrm{x}_i(t)$.

$\boldsymbol{x}_i(t)$

Frame of the log power spectrum, corresponding to the logarithm of a column of the STFT X_i.

$\hat{\mathrm{x}}(t)$, $\hat{\mathrm{X}}(\omega, t)$, $\hat{\boldsymbol{x}}(t)$

Signals corresponding to the reconstruction of x_i.

$\mathrm{y}(t)$, $\mathrm{Y}(\omega, t)$, $\boldsymbol{y}(t)$

Signals corresponding to the monaural mixture of multiple sources. In the case of binaural mixtures, superscripts of ℓ and r indicate the left and right channels, respectively.

$\mathrm{h}(t)$, $\mathrm{H}(\omega)$

Channel response time domain waveform and Fourier spectrum, respectively. Note that throughout the thesis we assume that the duration of $\mathrm{h}(t)$ is shorter than an STFT frame so the spectrum, H, does not have a time index.

$\phi(\omega, t)$

The interaural phase difference (IPD) between the left and right channels of a binaural signal.

$\alpha(\omega, t)$

The interaural level difference (ILD) between the left and right channels of a binaural signal.

Distributions and parameters

$P(a, b \mid \theta),\ E(a, b \mid \theta)$

Probability and expectation, respectively, of a and b given θ.

$\mathcal{N}(x; \mu, \sigma^2),\ \mathcal{C}(x; \mu, \sigma^2)$

Shorthand for the probability of x under a Gaussian probability distribution and cumulative distribution, respectively, parametrized by mean μ and variance σ^2.

$\bar{\mu},\ \bar{\Sigma}$

Multivariate Gaussian mean vector and covariance matrix of a speaker-independent speech model.

$\boldsymbol{\nu},\ \Xi$

Mean and covariance of the prior distribution over adaptation parameters.

$\mathbf{w}$

Subspace adaptation parameters for the eigenvoice model.

$\mathbf{h}$

Adaptation parameters corresponding to the speaker-independent channel response.

$\boldsymbol{\mu}(\mathbf{w}),\ \Sigma(\mathbf{w})$

Speaker-adapted mean and covariance parametrized by adaptation parameter $\mathbf{w}$.

$U,\ B$

Set of eigenvoice and channel basis vectors, respectively.

$i,\ c,\ s$

Indices corresponding to a particular source, Gaussian mixture model component, and hidden Markov model state, respectively.

N

Number of states in an HMM.

$\gamma_s(t)$

Posterior probability of HMM state s at time t.

τ

Iteraural time difference time delay.

$\Psi_{i\tau}$

Mixing weights of the IPD distribution corresponding to the probability of source i arriving from the direction corresponding to delay τ.

ς^2

Variance of the IPD distribution.

$\boldsymbol{\nu},\ \boldsymbol{\eta}^2$

Gaussian mean and diagonal covariance of the ILD distibution.

$z_{i\tau c}(\omega, t)$

Posterior probability of source i, delay τ, and GMM component c, at the given time-frequency point.

A.2 Derivation of MAP eigenvoice updates

In this section we derive the M-step of the adaptation algorithm used in section 4.2.2.2, which uses an estimate of the signal from a given source, $\hat{x}_i(t)$, to learn the adaptation parameters for that source. The extension to the M-step of the variational learning algorithm described in section 4.2.3 is straightforward.

We begin by deriving the updates under the assumption that the channel parameters $\mathbf{h}_i$ are subsumed into the subspace adaptation parameters $\mathbf{w}_i$ as described in section 3.4.3. The updates are found by maximizing the expected complete log likelihood from equation (4.31):

$$\begin{aligned}
\mathcal{L}(\mathbf{w}_i) &= E\big(\log P\big(\boldsymbol{y}(t),\, \mathbf{w}_i \,|\, s\big)\big) && \text{(A.1)}\\
&= E\big(\log P\big(\boldsymbol{y}(t) \,|\, \mathbf{w}_i,\, s\big) P\big(\mathbf{w}_i\big)\big) && \text{(A.2)}\\
&= E\big(\log \mathcal{N}\big(\boldsymbol{y}(t);\, \mu_s(\mathbf{w}_i),\, \Sigma_S(\mathbf{w}_i)\big) + \log \mathcal{N}\big(\mathbf{w}_i;\, \boldsymbol{\nu},\, \Xi\big)\big) && \text{(A.3)}\\
&= k - \frac{1}{2} \sum_t \sum_s \gamma_s(t)\, \big(\hat{\boldsymbol{x}}_i(t) - U_s \mathbf{w}_i - \bar{\boldsymbol{\mu}}_s\big)^T \bar{\Sigma}_s^{-1} \big(\hat{\boldsymbol{x}}_i(t) - U_s \mathbf{w}_i - \bar{\boldsymbol{\mu}}_s\big)\\
&\quad - \frac{1}{2}\big(\mathbf{w}_i - \boldsymbol{\nu}\big)^T \Xi^{-1} \big(\mathbf{w}_i - \boldsymbol{\nu}\big) && \text{(A.4)}
\end{aligned}$$

Because this auxiliary function is convex, it can be maximized by setting its derivative to zero. The derivative of the Mahalanobis distance in this form can be written as follows:

$$\frac{\partial}{\partial \mathbf{w}} \big(y - U\mathbf{w}\big)^T \Sigma^{-1} \big(y - U\mathbf{w}\big) = -2U^T \Sigma^{-1} \big(\boldsymbol{y} - U\mathbf{w}\big) \qquad \text{(A.5)}$$

By differentiating using equation (A.5), we get:

$$\frac{\partial}{\partial \mathbf{w}_i}\mathcal{L}(\mathbf{w}_i) = \sum_t \sum_s \gamma_s(t)\, U_s\, \bar{\Sigma}_s^{-1} \left(\hat{\boldsymbol{x}}(t) - U_s \mathbf{w}_i - \bar{\boldsymbol{\mu}}_s\right) - \Xi^{-1}\left(\mathbf{w}_i - \boldsymbol{\nu}\right) = 0 \tag{A.6}$$

which implies the following update for $\mathbf{w}_i$:

$$\left(\sum_t \sum_s \gamma_s(t)\, U_s^T\, \bar{\Sigma}_s^{-1}\, U_s + \Xi^{-1}\right) \mathbf{w}_i = \sum_t \sum_s \gamma_s(t)\, U_s^T\, \bar{\Sigma}_s^{-1}\, (\hat{\boldsymbol{x}}_i(t) - \bar{\boldsymbol{\mu}}_s) + \Xi^{-1}\boldsymbol{\nu} \tag{A.7}$$

When implementing this algorithm it is convenient to explicitly differentiate between $\mathbf{w}_i$ and $\mathbf{h}_i$. The solution can then be written as a matrix inversion as follows:

$$\begin{bmatrix} \mathbf{w}_i \\ \mathbf{h}_i \end{bmatrix} = \begin{bmatrix} A_{\mathbf{ww}} & A_{\mathbf{wh}} \\ A_{\mathbf{wh}}^T & A_{\mathbf{hh}} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{b}_{\mathbf{w}} \\ \boldsymbol{b}_{\mathbf{h}} \end{bmatrix} \tag{A.8}$$

where

$$A_{\mathbf{ww}} = \sum_t \sum_s \gamma_s(t)\, U_s^T\, \bar{\Sigma}_s^{-1}\, U_s + \Xi_{\mathbf{w}}^{-1} \tag{A.9}$$

$$A_{\mathbf{wh}} = \sum_t \sum_s \gamma_s(t)\, U_s^T\, \bar{\Sigma}_s^{-1}\, B \tag{A.10}$$

$$A_{\mathbf{hh}} = \sum_t \sum_s \gamma_s(t)\, B^T\, \bar{\Sigma}_s^{-1}\, B + \Xi_{\mathbf{h}}^{-1} \tag{A.11}$$

$$\boldsymbol{b}_{\mathbf{w}} = \sum_t \sum_s \gamma_s(t)\, U_s^T\, \bar{\Sigma}_s^{-1}\, (\hat{x}_i(t) - \bar{\boldsymbol{\mu}}_s) + \Xi_{\mathbf{w}}^{-1}\boldsymbol{\nu}_{\mathbf{w}} \tag{A.12}$$

$$\boldsymbol{b}_{\mathbf{h}} = \sum_t \sum_s \gamma_s(t)\, B^T\, \bar{\Sigma}_s^{-1}\, (\hat{x}_i(t) - \bar{\boldsymbol{\mu}}_s) + \Xi_{\mathbf{h}}^{-1}\boldsymbol{\nu}_{\mathbf{h}} \tag{A.13}$$

Note that we assume that the priors on $\mathbf{w}_i$ and $\mathbf{h}_i$ are independent.

The derivation of the M-step updates of the variational EM algorithm follows the same procedure, except the auxiliary function utilizes the mixed signal observations $\boldsymbol{y}(t)$ and includes the binary masks M_i as described in section 4.2.3.2. The derivation is nearly identical because the source signals decouple given the source masks (see equation (4.44)).

A.3 Derivation of MESSL updates

The parameter updates are found by maximizing the expected complete log likelihood from equation (5.18):

$$\begin{aligned}
\mathcal{L}(\theta \,|\, \theta_j) &= E\big(\log P\big(\boldsymbol{\phi},\, \boldsymbol{\alpha},\, \boldsymbol{y}^{\ell},\, \boldsymbol{y}^{r},\, i,\, \tau,\, c \,|\, \theta\big)\big) \\
&= \sum_{\omega t} E\big(\log P\big(\phi(\omega,t),\, \alpha(\omega,t),\, y^{\ell}(\omega,t),\, y^{r}(\omega,t),\, i,\, \tau,\, c \,|\, \theta\big)\big) \\
&= \sum_{\omega t} \sum_{i\tau c} z_{i\tau c}(\omega,t)\, \log P\big(\phi(\omega,t),\, \alpha(\omega,t),\, y^{\ell}(\omega,t),\, y^{r}(\omega,t) \,|\, i,\, \tau,\, c,\, \theta\big)
\end{aligned} \tag{A.14}$$

Because the total likelihood of equation (5.16) consists of the product of a set of conditionally independent factors, the objective function becomes the sum of independent terms allowing the parameters of each term to be optimized independently:

$$\mathcal{L}(\theta\,|\,\theta_j) = k + \mathcal{L}(\psi_{i\tau}\,|\,\theta_j) + \mathcal{L}(\varsigma_i^2\,|\,\theta_j) + \mathcal{L}(\boldsymbol{\upsilon}_i,\, \boldsymbol{\eta}_i^2\,|\,\theta_j) + \mathcal{L}(\mathbf{w}_i,\, \mathbf{h}_i^\ell,\, \mathbf{h}_i^r\,|\,\theta_j) \tag{A.15}$$

Each of the factored terms is concave in θ so they can each be maximized by differentiating with respect to θ and setting equal to zero.

We will optimize each of these terms in turn to derive the updates given in section 5.3.

A.3.1 Update for $\psi_{i\tau}$

$$\begin{aligned}\mathcal{L}(\psi_{i\tau}\,|\,\theta_j) &= \sum_{\omega t}\sum_{i\tau c} z_{i\tau c}(\omega, t)\ \log P(i,\,\tau) \\ &= \sum_{\omega t}\sum_{i\tau c} z_{i\tau c}(\omega, t)\ \log \psi_{i\tau} + \lambda\Big(1 - \sum_{i\tau}\psi_{i\tau}\Big)\end{aligned} \tag{A.16}$$

where we have introduced a Lagrange multiplier λ to ensure that $\phi_{i\tau}$ sums to one. Differentiating and setting to zero yields the following update:

$$\begin{aligned}\frac{\partial}{\partial\psi_{i\tau}}\mathcal{L}(\psi_{i\tau}\,|\,\theta_j) &= \sum_{\omega t}\sum_{c} z_{i\tau c}(\omega, t)\ \frac{1}{\psi_{i\tau}} - \lambda = 0 \\ \psi_{i\tau} &= \frac{1}{\lambda}\sum_{\omega t}\sum_{c} z_{i\tau c}(\omega, t)\end{aligned} \tag{A.17}$$

where λ is found using the constraint in equation (A.16):

$$\begin{aligned}1 &= \sum_{i\tau}\psi_{i\tau} = \sum_{i\tau}\frac{1}{\lambda}\sum_{\omega t}\sum_{c} z_{i\tau c}(\omega, t) \\ \lambda &= \sum_{\omega t}\sum_{i\tau c} z_{i\tau c}(\omega, t)\ = \Omega T\end{aligned} \tag{A.18}$$

where the last equality holds because each time-frequency point must be explained by a single source, delay, and source model mixture component. Combining this with equation (A.17) yields the update for $\psi_{i\tau}$:

$$\psi_{i\tau} = \frac{1}{\Omega T}\sum_{\omega t}\sum_{c} z_{i\tau c}(\omega, t) \tag{A.19}$$

A.3.2 Update for IPD parameter ς_i^2

$$\begin{aligned}
\mathcal{L}\big(\varsigma_i^2 \,|\, \theta_j\big) &= \sum_{\omega t} \sum_{i\tau c} z_{i\tau c}(\omega, t)\ \log P\big(\phi(\omega, t) \,|\, i,\, \tau,\, \theta\big) \\
&= \sum_{\omega t} \sum_{i\tau c} z_{i\tau c}(\omega, t)\ \log \mathcal{N}\big(\tilde{\phi}_\tau(\omega, t);\, 0,\, \varsigma_i^2\big) \\
&= k - \frac{1}{2} \sum_{\omega t} \sum_{i\tau c} z_{i\tau c}(\omega, t) \left(\frac{\tilde{\phi}_\tau(\omega, t)^2}{\varsigma_i^2} + \log \varsigma_i^2 \right) \qquad \text{(A.20)}
\end{aligned}$$

Differentiating and setting to zero yields the following update:

$$\begin{aligned}
\frac{\partial}{\partial \varsigma_i^2} \mathcal{L}\big(\varsigma_i^2 \,|\, \theta_j\big) &= -\frac{1}{2} \sum_{\omega t} \sum_{c} z_{i\tau c}(\omega, t) \left(-\frac{\tilde{\phi}_\tau(\omega, t)^2}{\varsigma_i^2} \frac{1}{\varsigma_i^2} + \frac{1}{\varsigma_i^2} \right) = 0 \\
\varsigma_i^2 &= \frac{\sum_{\omega t} \sum_c z_{i\tau c}(\omega, t)\, \tilde{\phi}_\tau(\omega, t)^2}{\sum_{\omega t} \sum_c z_{i\tau c}(\omega, t)} \qquad \text{(A.21)}
\end{aligned}$$

A.3.3 Update for ILD parameters υ_i, η_i^2

$$\begin{aligned}
\mathcal{L}\big(\boldsymbol{\upsilon}_i,\, \boldsymbol{\eta}_i^2 \,|\, \theta_j\big) &= \sum_{\omega t} \sum_{i\tau c} z_{i\tau c}(\omega, t)\ \log P\big(\alpha(\omega, t) \,|\, i,\, \theta\big) \\
&= \sum_{\omega t} \sum_{i\tau c} z_{i\tau c}(\omega, t)\ \log \mathcal{N}\big(\alpha(\omega, t);\, \upsilon_i(\omega),\, \eta_i^2(\omega)\big) \\
&= k - \frac{1}{2} \sum_{\omega t} \sum_{i\tau c} z_{i\tau c}(\omega, t) \left(\frac{\big(\alpha(\omega, t) - \upsilon_i(\omega)\big)^2}{\eta_i^2(\omega)} - \log \eta_i^2(\omega) \right) \qquad \text{(A.22)}
\end{aligned}$$

Differentiating with respect to the mean $\upsilon_i(\omega)$ and setting to zero yields the following update:

$$\begin{aligned}
\frac{\partial}{\partial \upsilon_i(\omega)} \mathcal{L}\big(\boldsymbol{\upsilon}_i,\, \boldsymbol{\eta}_i^2 \,|\, \theta_j\big) &= \sum_{t} \sum_{\tau c} z_{i\tau c}(\omega, t) \frac{\alpha(\omega, t) - \upsilon_i(\omega)}{\eta_i^2(\omega)} = 0 \\
\upsilon_i(\omega) &= \frac{\sum_t \sum_{\tau c} z_{i\tau c}(\omega, t)\, \alpha(\omega, t)}{\sum_t \sum_{\tau c} z_{i\tau c}(\omega, t)} \qquad \text{(A.23)}
\end{aligned}$$

Similarly, differentiating with respect to the variance $\eta_i^2(\omega)$ and setting to zero yields the following update:

$$\begin{aligned}
\frac{\partial}{\partial \eta_i^2(\omega)} \mathcal{L}\big(\boldsymbol{\upsilon}_i,\, \boldsymbol{\eta}_i^2 \,|\, \theta_j\big) &= -\frac{1}{2} \sum_{t} \sum_{\tau c} z_{i\tau c}(\omega, t) \left(-\frac{\big(\alpha(\omega, t) - \upsilon_i(\omega)\big)^2}{\eta_i^2(\omega)} \frac{1}{\eta_i^2(\omega)} + \frac{1}{\eta_i^2(\omega)} \right) = 0 \\
\eta_i^2(\omega) &= \frac{\sum_t \sum_{\tau c} z_{i\tau c}(\omega, t)\, \big(\alpha(\omega, t) - \upsilon_i(\omega)\big)^2}{\sum_t \sum_{\tau c} z_{i\tau c}(\omega, t)} \qquad \text{(A.24)}
\end{aligned}$$

A.3.4 Update for source and channel parameters $\mathbf{w}_i, \mathbf{h}_i^r, \mathbf{h}_i^\ell$

$$
\begin{aligned}
&\mathcal{L}\big(\mathbf{w}_i,\, \mathbf{h}_i^\ell,\, \mathbf{h}_i^r \,|\, \theta_j\big) \\
&\quad = \sum_{\omega t} \sum_{i\tau c} z_{i\tau c}(\omega, t)\, \log P\big(y^\ell(\omega, t) \,|\, i,\, c,\, \theta\big)\, P\big(y^r(\omega, t) \,|\, i,\, c,\, \theta\big) \\
&\quad = \sum_{t} \sum_{i\tau c} \Big(\log \mathcal{N}\big(\boldsymbol{y}^\ell(t);\, \boldsymbol{\mu}_c(\mathbf{w}_i, \mathbf{h}_i^\ell),\, M_{ict}\bar{\Sigma}_c\big) + \log \mathcal{N}\big(\boldsymbol{y}^r(t);\, \boldsymbol{\mu}_c(\mathbf{w}_i, \mathbf{h}_i^r),\, M_{ict}\bar{\Sigma}_c\big)\Big) \\
&\quad = k - \frac{1}{2} \sum_{t} \sum_{ic} \Big(\big(\boldsymbol{y}^\ell(t) - U_c\mathbf{w}_i - \bar{\boldsymbol{\mu}}_c - B\mathbf{h}_i^\ell\big)^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(\boldsymbol{y}^\ell(t) - U_c\mathbf{w}_i - \bar{\boldsymbol{\mu}}_c - B\mathbf{h}_i^\ell\right) \\
&\qquad\qquad + \big(\boldsymbol{y}^\ell(t) - U_c\mathbf{w}_i - \bar{\boldsymbol{\mu}}_c - B\mathbf{h}_i^r\big)^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(\boldsymbol{y}^\ell(t) - U_c\mathbf{w}_i - \bar{\boldsymbol{\mu}}_c - B\mathbf{h}_i^r\right)\Big) \qquad \text{(A.25)}
\end{aligned}
$$

where M_{ict} is a soft mask encoding the posterior probability of component c from source i dominating the mixture at frame t:

$$
M_{ict} \triangleq \text{diag}\left(\sum_{\tau} z_{i\tau c}(:, t)\right) \qquad \text{(A.26)}
$$

Unlike the similar objective function in equation (A.4), the posterior varies with frequency due to the assumption that each time-frequency cell is dominated by a particular source, hence the need to incorporate them into a soft mask.

Differentiating and setting to zero yields the following set of simultaneous equations:

$$
\sum_{tc} U_c^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(2\left(\bar{\boldsymbol{\mu}}_c + U_c\, \mathbf{w}_i\right) + B\left(\mathbf{h}_i^r + \mathbf{h}_i^\ell\right)\right) = \sum_{tc} U_c^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(\boldsymbol{y}^\ell(t) + \boldsymbol{y}^r(t)\right) \qquad \text{(A.27)}
$$

$$
\sum_{tc} B^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(\bar{\boldsymbol{\mu}}_c + U_c\, \mathbf{w}_i + B\,\mathbf{h}_i^\ell\right) = \sum_{tc} B^T M_{ict}\, \bar{\Sigma}_c^{-1}\, \boldsymbol{y}^\ell(t) \qquad \text{(A.28)}
$$

$$
\sum_{tc} B^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(\bar{\boldsymbol{\mu}}_c + U_c\, \mathbf{w}_i + B\,\mathbf{h}_i^r\right) = \sum_{tc} B^T M_{ict}\, \bar{\Sigma}_c^{-1}\, \boldsymbol{y}^r(t) \qquad \text{(A.29)}
$$

The updates for $\mathbf{w}_i, \mathbf{h}_i^\ell$, and $\mathbf{h}_i^r$ can be written as a matrix inversion as follows:

$$
\begin{bmatrix} \mathbf{w}_i \\ \mathbf{h}_i^\ell \\ \mathbf{h}_i^r \end{bmatrix} = \begin{bmatrix} A_{\mathbf{ww}} & A_{\mathbf{wh}} & A_{\mathbf{wh}} \\ A_{\mathbf{wh}}^T & A_{\mathbf{hh}} & A_{\mathbf{wh}} \\ A_{\mathbf{wh}}^T & A_{\mathbf{wh}}^T & A_{\mathbf{hh}} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{b}_{\mathbf{w}} \\ \boldsymbol{b}_{\mathbf{h}^\ell} \\ \boldsymbol{b}_{\mathbf{h}^r} \end{bmatrix} \qquad \text{(A.30)}
$$

where

$$A_{\mathbf{ww}} = \sum_{tc} 2\, U_c^T M_{ict}\, \bar{\Sigma}_c^{-1}\, U_c \tag{A.31}$$

$$A_{\mathbf{wh}} = \sum_{tc} U_c^T M_{ict}\, \bar{\Sigma}_c^{-1}\, B \tag{A.32}$$

$$A_{\mathbf{hh}} = \sum_{tc} B^T M_{ict}\, \bar{\Sigma}_c^{-1}\, B \tag{A.33}$$

$$\boldsymbol{b}_{\mathbf{w}} = \sum_{tc} U_c^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(\boldsymbol{y}^\ell(t) + \boldsymbol{y}^r(t) - 2\,\bar{\boldsymbol{\mu}}_c\right) \tag{A.34}$$

$$\boldsymbol{b}_{\mathbf{h}^\ell} = \sum_{tc} B^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(\boldsymbol{y}^\ell(t) - \bar{\boldsymbol{\mu}}_c\right) \tag{A.35}$$

$$\boldsymbol{b}_{\mathbf{h}^r} = \sum_{tc} B^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(\boldsymbol{y}^r(t) - \bar{\boldsymbol{\mu}}_c\right) \tag{A.36}$$

Note the resemblance to the eigenvoice updates derived in section A.2. The main differences here are that the posterior depends on frequency (hence the need for the mask term M_{ict}) and $\mathbf{w}_i$ is tied across two observations, each of which have separate channel responses. Also, we did not include a prior distribution on $\mathbf{w}$ and $\mathbf{h}$, but its addition is straightforward.

Bibliography

[1] P. Aarabi. Self-localizing dynamic microphone arrays. *IEEE Transactions on Systems, Man, and Cybernetics*, 32(4), November 2002. (Cited on page 86.)

[2] F. Abrard, Y. Deville, and P. White. From blind source separation to blind source cancellation in the underdetermined case: a new approach based on time-frequency analysis. In *Proc. International Conference on Independent Component Analysis and Signal Separation (ICA)*, pages 734–739, 2001. (Cited on pages 8 and 9.)

[3] A. Acero, S. Altschuler, and L. Wu. Speech/Noise Separation Using Two Microphones and a VQ Model of Speech Signals. In *Proc. Sixth International Conference on Spoken Language Processing (ICSLP)*. ISCA, 2000. (Cited on page 78.)

[4] 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*, pages 99–102, October 2001. (Cited on page 90.)

[5] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling, 2000. (Cited on page 19.)

[6] F. R. Bach and M. I. Jordan. Blind one-microphone speech separation: A spectral learning approach. In *Advances in Neural Information Processing Systems*, 2004. (Cited on page 12.)

[7] J. Barker, A. Coy, N. Ma, and M. Cooke. Recent advances in speech fragment decoding techniques. In *Proc. Interspeech*, pages 85–88, 2006. (Cited on pages 24, 67, and 68.)

[8] J. Barker, N. Ma, A. Coy, and M. Cooke. Speech fragment decoding techniques for simultaneous speaker identification and speech recognition. *Computer Speech and Language*, In Press, Corrected Proof:–, 2008. ISSN 0885-2308. doi: DOI:10.1016/j.csl.2008.05.003. (Cited on pages 24 and 67.)

[9] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995. (Cited on page 6.)

[10] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. (Cited on pages 25 and 102.)

[11] A. S. Bregman. *Auditory Scene Analysis*. Bradford Books, MIT Press, 1990. (Cited on page 11.)

[12] M.A. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. In *Proc. of the International Computer Music Conference*, August 2000. (Cited on pages 9 and 10.)

[13] E. C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 25(5):975–979, 1953. doi: 10.1121/1.1907229. (Cited on pages 1 and 66.)

[14] S. Choi, A. Cichocki, H.M. Park, and S.Y. Lee. Blind source separation and independent component analysis: A review. *Neural Information Processing-Letters and Reviews*, 6(1):1–57, 2005. (Cited on page 6.)

[15] M. Cooke and D. P. W. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, 35(3-4):141 – 177, 2001. ISSN 0167-6393. doi: DOI:10.1016/S0167-6393(00)00078-9. (Cited on page 12.)

[16] M. Cooke and T.-W. Lee. The speech separation challenge, 2006. URL `http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm`. (Cited on pages iii, 2, 17, 24, 29, 30, 58, 59, and 90.)

[17] M. P. Cooke, M. L. Garcia Lecumberri, and J. Barker. The non-native cocktail party. (in preparation). (Cited on page 68.)

[18] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120:2421–2424, 2006. (Cited on pages v and 58.)

[19] M. E. Davies and C. J. James. Source separation using single channel ICA. *Signal Processing*, 87(8):1819–1832, 2007. ISSN 0165-1684. doi: http://dx.doi.org/10.1016/j.sigpro.2007.01.011. (Cited on page 9.)

[20] J. R. Deller, J. H. Hansen, and J. G. Proakis. *Discrete Time Processing of Speech Signals*. Wiley-IEEE Press, 1999. ISBN 0780353862. (Cited on pages 15 and 21.)

[21] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likehood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977. (Cited on page 51.)

[22] O. D. Deshmukh and C. Y. Espy-Wilson. Modified Phase Opponency Based Solution to the Speech Separation Challenge. In *Ninth International Conference on Spoken Language Processing*. ISCA, 2006. (Cited on page 68.)

[23] S. C. Douglas. Blind separation of acoustic signals. In M. Brandstein and D. Ward, editors, *Microphone Arrays: Techniques and Applications*, chapter 16, pages 355–380. Springer, 2001. (Cited on page 6.)

[24] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, New York, 2nd edition, 2001. (Cited on pages 27 and 34.)

[25] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, 1996. (Cited on page 20.)

[26] D. Ellis. Model-based scene analysis. In D. Wang and G. Brown, editors, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, chapter 4, pages 115–146. Wiley/IEEE Press, 2006. (Cited on pages 7, 11, 21, and 43.)

[27] D. P. W. Ellis. *Prediction–driven computational auditory scene analysis*. PhD thesis, Department of Electrtical Engineering and Computer Science, M.I.T., 1996. (Cited on page 12.)

[28] D. P. W. Ellis and J. Arroyo. Eigenrhythms: Drum pattern basis sets for classification and generation. In *Proc. International Symposium on Music Information Retrieval (ISMIR)*, volume 4, 2004. (Cited on page 19.)

[29] D. P. W. Ellis and R. J. Weiss. Model-based monaural source separation using a vector-quantized phase-vocoder representation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages V–957–960, Toulouse, France, May 2006. (Cited on pages iii, 3, and 15.)

[30] M. R. Every and P. J. B. Jackson. Enhancement of Harmonic Content of Speech Based on a Dynamic Programming Pitch Tracking Algorithm. In *Ninth International Conference on Spoken Language Processing*. ISCA, 2006. (Cited on page 68.)

[31] M. J. F. Gales. Cluster adaptive training of hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, 8(4):417–428, July 2000. (Cited on page 34.)

[32] M. J. F. Gales. Cluster Adaptive Training for Speech Recognition. In *Proc. Fifth International Conference on Spoken Language Processing (ICSLP)*. ISCA, 1998. (Cited on pages 19 and 26.)

[33] M. A. Gandhi and M. A. Hasegawa-Johnson. Source separation using particle filters. In *Proc. Interspeech*, 2004. (Cited on page 15.)

[34] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. DARPA TIMIT acoustic phonetic continuous speech corpus CDROM, 1993. URL `http://www.ldc.upenn.edu/Catalog/LDC93S1.html`. (Cited on page 90.)

[35] Z. Ghahramani and M.I. Jordan. Factorial hidden markov models. *Machine Learning*, 29(2-3):245–273, November 1997. (Cited on pages 38, 42, 51, 52, and 54.)

[36] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: telephone speech corpus for research and development. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, 1992. (Cited on page 71.)

[37] R. M. Gray. Vector quantization. *IEEE ASSP Magazine*, 1(2):4–29, 1984. (Cited on page 13.)

[38] D. Griffin and J. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32:236–242, 1984. (Cited on page 43.)

[39] S. Harding, J. Barker, and G. J. Brown. Mask estimation for missing data speech recognition based on statistics of binaural interaction. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):58–67, 2006. (Cited on page 78.)

[40] J. R. Hershey and P. A. Olsen. Variational Bhattacharyya divergence for hidden Markov models. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4557–4560, 2008. (Cited on pages 72 and 73.)

[41] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson. Super-human multi-talker speech recognition: A graphical modeling approach. *Computer Speech and Language*, In Press, Accepted Manuscript:–, 2009. ISSN 0885-2308. doi: DOI:10.1016/j.csl.2008.11.001. (Cited on pages 13 and 42.)

[42] G. Hu and D. L. Wang. Monaural speech separation. *Advances in Neural Information Processing Systems*, pages 1245–1252, 2003. (Cited on page 12.)

[43] C.-H. Huang, J.-T. Chien, and H.-M. Wang. A new eigenvoice approach to speaker adaptation. In *Proc. International Symposium on Chinese Spoken Language Processing*, pages 109–112, 2004. (Cited on pages 31, 50, and 56.)

[44] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000. (Cited on pages 6 and 34.)

[45] S. Ikeda and N. Murata. A method of ICA in time-frequency domain. In *Proc. International Conference on Independent Component Analysis and Signal Separation (ICA)*, pages 365–371, 1999. (Cited on pages 6 and 7.)

[46] G. J. Jang and T. W. Lee. A maximum likelihood approach to single-channel source separation. *Journal of Machine Learning Research*, 4:1365–1392, 2003. (Cited on pages 11 and 102.)

[47] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 2985–2988, June 2000. (Cited on page 91.)

[48] P. Kenny, M. Mihoubi, and P. Dumouchel. New MAP Estimators for Speaker Recognition. In *Eighth European Conference on Speech Communication and Technology*. ISCA, 2003. (Cited on page 73.)

[49] P. Kenny, G. Boulianne, and P. Dumouchel. Eigenvoice modeling with sparse training data. *IEEE Transactions on Audio, Speech, and Language Processing*, 13(3), May 2005. (Cited on page 19.)

[50] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel. A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5), July 2008. (Cited on page 19.)

[51] S. Kochkin. MarkeTrak V:" Why my hearing aids are in the drawer": The consumers' perspective. *Hearing Journal*, 53(2):34–42, 2000. (Cited on page 1.)

[52] A. Koutras, E. Dermatas, and G. Kokkinakis. Blind speech separation of moving speakers in real reverberant environments. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages II1133–II1136 vol.2, 2000. doi: 10.1109/ICASSP.2000.859164. (Cited on page 6.)

[53] T. Kristjansson and J. Hershey. High resolution signal reconstruction. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 291–296, 2003. doi: 10.1109/ASRU.2003.1318456. (Cited on pages 15 and 101.)

[54] T. Kristjansson, H. Attias, and J. Hershey. Single microphone source separation using high resolution signal reconstruction. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages II–817–820, 2004. (Cited on page 13.)

[55] T. Kristjansson, J. Hershey, P. Olsen, S. Rennie, and R. Gopinath. Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system. In *Proc. Interspeech*, pages 97–100, 2006. (Cited on pages 21, 22, 23, 24, 67, and 68.)

[56] T. T. Kristjansson. *Speech Recognition in Adverse Environments: a Probabilistic Approach*. PhD thesis, Department of Computer Science, University of Waterloo, 2002. (Cited on pages 40, 76, and 102.)

[57] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini. Eigenvoices for Speaker Adaptation. In *Proc. Fifth International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, 1998. ISCA. (Cited on pages 19, 26, and 27.)

[58] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Transations on Speech and Audio Processing*, 8(6):695–707, November 2000. (Cited on pages 26, 31, 47, 48, 50, and 71.)

[59] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, pages 556–562, 2001. (Cited on page 101.)

[60] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. (Cited on pages 10 and 31.)

[61] T. W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 6(4):87–90, 1999. (Cited on page 7.)

[62] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, pages 171–185, 1995. (Cited on page 25.)

[63] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982. (Cited on page 49.)

[64] P.C. Loizou. *Speech enhancement: theory and practice*. CRC press Boca Raton: FL:, 2007. (Cited on page 91.)

[65] S. Lucey and T. Chen. An investigation into subspace rapid speaker adaptation for verification. In *Proc. International Conference on Multimedia and Expo (ICME)*. IEEE, 2003. (Cited on page 19.)

[66] M. Mandel, D. Ellis, and T. Jebara. An EM algorithm for localizing multiple sound sources in reverberant environments. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA, 2007. (Cited on page 77.)

[67] M. I. Mandel and D. P. W. Ellis. EM localization and separation using interaural level and phase cues. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2007. (Cited on pages 2, 77, 89, and 91.)

[68] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis. Model-based expectation maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 2009. in press. (Cited on pages 3, 77, and 78.)

[69] J. C. Middlebrooks and D. M. Green. Sound localization by human listeners. *Annual Review of Psychology*, 42(1):135–159, 1991. (Cited on page 8.)

[70] J. Ming, T. J. Hazen, and J. R Glass. Combining missing-feature theory, speech enhancement and speaker-dependent/-independent modeling for speech separation. In *Ninth International Conference on Spoken Language Processing*. ISCA, 2006. (Cited on page 68.)

[71] A. Nadas, D. Nahamoo, and M. A. Picheny. Speech recognition using noise-adaptive prototypes. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(10):1495–1503, 1989. (Cited on pages 38, 40, and 42.)

[72] J. Nix and V. Hohmann. Sound source localization in real sound fields based on empirical statistics of interaural parameters. *Journal of the Acoustical Society of America*, 119(1):463–479, 2006. (Cited on page 78.)

[73] J. Nix, M. Kleinschmidt, and V. Hohmann. Computational auditory scene analysis by using statistics of high-dimensional speech dynamics and sound source direction. In *Proc. Eurospeech*, pages 1441–1444, Geneva, 2003. (Cited on page 12.)

[74] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot. One microphone singing voice separation using source-adapted models. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 90–93, October 2005. (Cited on page 25.)

[75] T. W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America*, 60(4):911–918, 1976. doi: 10.1121/1.381172. (Cited on pages 5 and 11.)

[76] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra. A survey of convolutive blind source separation methods. In *Springer Handbook of Speech Processing*. Springer Press, November 2007. (Cited on page 6.)

[77] J. Picone and G. R. Doddington. A phonetic vocoder. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 580–583, 1989. (Cited on page 23.)

[78] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989. (Cited on pages 42, 50, and 54.)

[79] M. H. Radfar and R. M. Dansereau. Single-channel speech separation using soft mask filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8): 2299–2310, November 2007. ISSN 1558-7916. doi: 10.1109/TASL.2007.904233. (Cited on pages 16 and 17.)

[80] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan. Performance evaluation of three features for model-based single channel speech separation problem. In *Proc. Interspeech*. ISCA, 2006. (Cited on page 15.)

[81] A. M. Reddy and B. M. Raj. Soft mask estimation for single channel source separation. In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA)*, 2004. (Cited on pages 13 and 14.)

[82] S. Rennie, P. Aarabi, T. Kristjansson, B. J. Frey, and K. Achan. Robust variational speech separation using fewer microphones than speakers. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003. (Cited on pages 78, 79, and 97.)

[83] S. Rennie, P. Olsen, J. Hershey, and T. Kristjansson. The Iroquois model: Using temporal dynamics to separate speakers. In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA)*, Pittsburgh, PA, September 2006. (Cited on pages 24, 42, 45, and 46.)

[84] S. J. Rennie, K. Achan, B. J. Frey, and P. Aarabi. Variational speech separation of more sources than mixtures. In *Proc. Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2005. (Cited on page 79.)

[85] S. J. Rennie, J. R. Hershey, and P. A. Olsen. Efficient model-based speech separation and denoising using non-negative subspace analysis. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1833–1836, April 2008. doi: 10.1109/ICASSP.2008.4517989. (Cited on page 14.)

[86] S. J. Rennie, J. R. Hershey, and P. A. Olsen. Single-Channel Speech Separation and Recognition Using Loopy Belief Propagation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2009. (Cited on pages 63 and 101.)

[87] M. J. Reyes-Gómez. *Statistical graphical models for scene analysis, source separation and other audio applications*. PhD thesis, Department of Electrical Engineering, Columbia University, 2007. (Cited on page 78.)

[88] M. J. Reyes-Gómez, D. P. W. Ellis, and N. Jojic. Multiband audio modeling for single-channel acoustic source separation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages V–641–644 vol.5, May 2004. doi: 10.1109/ICASSP.2004.1327192. (Cited on pages 13 and 42.)

[89] N. Roman, D. Wang, and G. J. Brown. A classification-based cocktail party processor. In *Advances in Neural Information Processing Systems*, 2003. (Cited on page 78.)

[90] J. Le Roux, N. Ono, and S. Sagayama. Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction. In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA)*, 2008. (Cited on page 43.)

[91] S. T. Roweis. One microphone source separation. In *Advances in Neural Information Processing Systems*, pages 793–799, 2000. (Cited on pages 2, 7, 14, and 42.)

[92] S. T. Roweis. Factorial models and refiltering for speech separation and denoising. In *Proc. Eurospeech*, pages 1009–1012, 2003. (Cited on pages 13, 14, 40, 47, and 98.)

[93] H. Runqiang, Z. Pei, G. Qin, Z. Zhiping, W. Hao, and W. Xihong. CASA based speech separation for robust speech recognition. In *Ninth International Conference on Spoken Language Processing*. ISCA, 2006. (Cited on page 68.)

[94] H. Saruwatari, S. Kurita, and K. Takeda. Blind source separation combining frequency-domain ICA and beamforming. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5. IEEE, 2001. (Cited on page 7.)

[95] H. Sawada, R. Mukai, S. Araki, and S. Makino. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *Speech and Audio Processing, IEEE Transactions on*, 12(5):530–538, September 2004. ISSN 1063-6676. doi: 10.1109/TSA.2004.832994. (Cited on page 7.)

[96] H. Sawada, S. Araki, and S. Makino. A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2007. (Cited on page 91.)

[97] M. N. Schmidt and R. K. Olsson. Single-channel speech separation using sparse non-negative matrix factorization. In *Proc. Interspeech*, pages 2614–2617, September 2006. (Cited on pages 10, 16, 17, 23, 24, 68, and 101.)

[98] M. L. Seltzer, B. Raj, and R. M. Stern. A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition. *Speech Communication*, 43(4):379 – 393, 2004. (Cited on page 12.)

[99] B. Shinn-Cunningham, N. Kopco, and T. Martin. Localizing nearby sound sources in a classroom: Binaural room impulse responses. *Journal of the Acoustical Society of America*, 117:3100–3115, 2005. (Cited on page 91.)

[100] P. Smaragdis. Discovering Auditory Objects Through Non-Negativity Constraints. In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA)*. ISCA, 2004. (Cited on page 10.)

[101] P. Smaragdis. Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):1, 2007. (Cited on pages 10 and 14.)

[102] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22(1):21–34, 1998. (Cited on page 6.)

[103] S. Srinivasan, Y. Shao, Z. Jin, and D. Wang. A computational auditory scene analysis system for robust speech recognition. In *Proc. Interspeech*, pages 73–76, September 2006. (Cited on pages 17, 67, and 68.)

[104] Switchboard Resegmentation. URL `http://www.isip.msstate.edu/projects/switchboard/`. Institute for Signal and Information Processing, Mississippi State University. (Cited on page 71.)

[105] O. Thyes, R. Kuhn, P. Nguyen, and J. C. Junqua. Speaker Identification and Verification Using Eigenvoices. In *Proc. Sixth International Conference on Spoken Language Processing (ICSLP)*. ISCA, 2000. (Cited on page 19.)

[106] M. E. Tipping and C. M. Bishop. Mixtures of Probabilistic Principal Component Analyzers. *Neural Computation*, 11(2):443–482, 1999. (Cited on page 33.)

[107] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991. (Cited on page 19.)

[108] P. Varga and R. K. Moore. Hidden Markov model decomposition of speech and noise. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 845–848, 1990. (Cited on page 40.)

[109] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca. First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results. In *Proc. International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2007. (Cited on page 16.)

[110] E. Vincent, S. Araki, and P. Bofill. The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation. In *Proc. International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2009. (Cited on pages 16 and 17.)

[111] T. Virtanen. Separation of sound sources by convolutive sparse coding. In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA)*. ISCA, 2004. (Cited on page 10.)

[112] T. Virtanen. Speech recognition using factorial hidden Markov models for separation in the feature space. In *Proc. Interspeech*, pages 89–92, September 2006. (Cited on pages 24, 67, and 68.)

[113] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, 2007. (Cited on page 10.)

[114] D. L. Wang and G. Hu. Unvoiced speech segregation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, 2006. doi: 10.1109/ICASSP.2006.1661435. (Cited on page 12.)

[115] M. Weintraub. *A theory and computational model of auditory monoaural sound separation*. PhD thesis, Department of Electrical Engineering, Stanford University, 1985. (Cited on page 11.)

[116] R. J. Weiss and D. P. W. Ellis. Estimating single-channel source separation masks: Relevance vector machine classifiers vs. pitch-based masking. In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA)*, pages 31–36, Pittsburgh, PA, September 2006. (Cited on pages 3 and 12.)

[117] R. J. Weiss and D. P. W. Ellis. Monaural speech separation using source-adapted models. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 114–117, October 2007. (Cited on pages 3 and 59.)

[118] R. J. Weiss and D. P. W. Ellis. Speech separation using speaker-adapted eigenvoice speech models. *Computer Speech and Language*, In Press, Corrected Proof:–, 2008. ISSN 0885-2308. doi: DOI:10.1016/j.csl.2008.03.003. (Cited on pages 3, 29, 59, and 98.)

[119] R. J. Weiss and D. P. W. Ellis. A Variational EM Algorithm for Learning Eigenvoice Parameters in Mixed Signals. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2009. (Cited on pages 3, 31, and 52.)

[120] R. J. Weiss, M. I. Mandel, and D. P. W. Ellis. Source separation based on binaural cues and source model constraints. In *Proc. Interspeech*, pages 419–422, Brisbane, Australia, September 2008. (Cited on pages 3, 17, 77, 79, 84, 91, and 94.)

[121] Frederic L. Wightman and Doris J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *Journal of the Acoustical Society of America*, 91(3):1648–1661, 1992. (Cited on page 88.)

[122] K. Wilson. Speech source separation by combining localization cues with mixture models of speech spectra. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages I–33–36, 2007. (Cited on pages 17, 78, 79, and 97.)

[123] P. C. Woodland. Speaker Adaptation for Continuous Density HMMs: A Review. In *Proc. ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition*. ISCA, 2001. (Cited on pages 25 and 29.)

[124] K. Yao, K. K. Paliwal, and T.-W. Lee. Generative factor analyzed hmm for automatic speech recognition. *Speech Communication*, 45(4):435 – 454, 2005. ISSN 0167-6393. doi: DOI:10.1016/j.specom.2005.01.002. (Cited on page 26.)

[125] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, July 2004. (Cited on pages 7, 9, 77, 78, 81, and 91.)

[126] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK, 2006. (Cited on pages 28 and 58.)

[127] M. Zibulevsky and B.A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13(4):863–882, 2001. (Cited on page 7.)