Model-based Single-microphone Speech Separation Using Conditional Random Fields

YEUNG, Yu Ting

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Electronic Engineering

The Chinese University of Hong Kong
April 2014

To my family

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Prof. Tan Lee, for his guidance, support and advices. I would like to sincerely thank Dr. Chung-Chi Leung for his suggestions. I would like to thank Dr. Feng Huang, Mr Shing Yu for their suggestions and discussions. I would like to thank Mr. Lawrence Man Wai Un for preparing the scripts for speech recognition training under my guidance during his participation in the undergraduate summer research program. I would like to thank Mr. Wang Kong Lam and Mr. Hoi To Wai for the coffee time we enjoyed. I would like to thank all my colleagues in the Digital Signal Processing and Speech Technology Laboratory. They helped me a lot in different ways.

The implementation of the computer algorithms in this thesis is based on the UGM Matlab toolkit. I would like to thank the authors Mark Schmidt *et. al.* for their kindness to release the software under the open-source FreeBSD-style license. The toolkit is available at: http://www.di.ens.fr/~mschmidt/Software/UGM.html.

Finally, I would like to thank my family and friends for the continuous support during this period.

Abstract of thesis entitled:

Model-based Single-microphone Speech Separation Using Conditional Random Fields

Submitted by **YEUNG, Yu Ting**

for the degree of **Doctor of Philosophy**

in **Electronic Engineering**

at **The Chinese University of Hong Kong**

in April 2014.

Single-microphone speech separation requires to reconstruct two or more sources from only one speech mixture. It can serve as the front-end for speech applications that demand for robustness against interfering signals, such as information extraction from sound streams of multimedia. As an extreme case of under-determined source separation problem, a unique solution for source reconstruction is unlikely to be achieved, but the most probable source observations can be obtained through statistical inference given their prior information in a statistical model-based setting.

The performance of statistical model-based methods has been progressively improved by the use of graphical models to organize the prior information. In this thesis, the performance of the exact and the approximated statistical inference algorithms on single-microphone speech separation with factorial Hidden Markov models (HMM) are evaluated in terms of speech quality and computational complexity. The important role of state transitions in the source models is also investigated.

Model mis-specification is a major problem in model-based speech separation. These mis-specifications are caused by various factors, including limited amount of training data and finite number of acoustic states. Compared with generative ap-

proach such as factorial HMM, direct models like conditional random fields (CRF) are considered to be more robust to model mis-specification due to the inherent discrimination ability. In this thesis, the application of conditional random field (CRF) for single-microphone speech separation is investigated. The posterior probabilities of acoustic states given the mixture, which are essential to minimum mean-square error estimation of the sources, are modeled in a maximum entropy probability distribution. The performance of CRF formulations is further improved with a large-margin approach of parameter estimation.

Experimental results confirm that CRF formulations achieve the improved objective quality measures and automatic speech recognition accuracy of the reconstructed sources, especially when the sources are competing with similar signal-to-signal ratio. Even with a simplified CRF formulation, the performance is still comparable to factorial HMM.

摘要

單麥克風語音分離的目標是從一個語音混合 (speech mixture) 中重建兩個或更多的語音源 (source)。這技術可作為語音應用的前置處理，例如從多媒體音軌中抽取資訊。雖然作為欠定 (under-determined) 語音分離的極端例子，基本上沒可能確切地還原語音源，但透過語音源的統計模型，仍可重構出最有可能的語音源。

語音分離的性能藉著圖模式 (graphical modeling) 的應用而得以提升。本論文比較了因子隱馬爾可夫模型(factorial Hidden Markov Model (HMM)) 的精確算法和近似算法的複雜度和對語音分離性能的影響，並且調查語音源統計模型中的狀態轉移機率 (state transition probabilities) 對語音分離性能的影響。

統計模型錯配在語音分離中時有發生。有限的訓練資料和使用有限的狀態空間 (acoustic states) 對語音源建模都會導致錯配。本論文研究了使用條件隨機域 (conditional random field (CRF)) 來對語音源狀態空間的後驗概率直接建模。計算語音源的最小均方差估計 (minimum mean-square error) 時，這後驗概率是必須的。條件隨機域是一種判別模型 (discriminative model)，比生成模型 (generative model) 例如隱馬爾可夫模型對模型錯配有更高的耐受性。使用大間隔 (large-margin) 參數估計更進一步提升語音分離的效能。

實驗結果證明當不同語音源的功率比 (signal-to-signal ratio) 相近時，使用條件隨機域作語音分離可以獲得更好的語音音質客觀測量參數 (objective quality measures) 和語音識別結果。即使使用簡化了的條件隨機域，結果仍和使用因子隱馬爾可夫模型相當。

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Single-microphone speech separation

Speech is a fundamental and an important form of daily communication for human beings. Speech is produced by human articulatory system and propagates across an open space via vibration of air molecules. The acoustic vibration may be picked up by human auditory system, or by electronic devices such as microphones which convert the acoustic signals into electrical signals, referred to as speech signals.

It is unavoidable to deal with corrupted speech signals. For example, in mobile voice communication, the telephone handset may pick up additional audio sources from background, such as babble sound of a crowd, engine sound of a vehicle. The interfering sources may also be from other speakers. If the interfering sources are perceptually undesirable, they are referred to as noise [1].

The interfering sources usually corrupt the target speech source by superposition. Computational processes can be applied to improve the listening experience on the corrupted speech signals. Listening experience accounts for the comfort of listening and the understanding of the spoken content, which are measured in terms of perceptual speech quality and speech intelligibility respectively. Single-microphone speech separation is one of the computational processes, aiming at reconstructing all the speech sources from a single corrupted speech signal with multiple sources.

Single-microphone speech separation is a challenging problem. It operates under highly non-stationary interfering condition. It also tries to reconstruct at least two

sources with only one corrupted signal, which is an extreme case of under-determined source separation. Although unique source reconstruction is unlikely to be achieved, different approaches have been proposed to improve perceptual speech quality and speech intelligibility of the reconstructed sources. A recent development is the incorporation of machine learning techniques to the problem, such as the application of hidden Markov models (HMM) in statistical model-based methods [2][3] and the pattern classification approach in computational auditory scene analysis (CASA) [4].

1.2 Motivations

The motivation of this work comes from the variety of applications of speech separation. A speech separation algorithm can serve as the front-end for speech applications that demand for robustness against interfering signals, e.g., speech recognition on mobile hand-held devices, audio information retrieval from live recordings. Experiments show that the performance of a conventional automatic speech recognition system degrades significantly when there is an interfering source [5]. A speech separation algorithm helps to reconstruct the acoustic features of the target signal, and serves as a preliminary step for robust speech processing such as pitch tracking in noise [6] and periodicity enhancement [7]. The recent advances in graphical models [8] lead to significant improvement on the state-of-the-art speech separation algorithms. It is shown that factorial hidden Markov model (HMM) with loopy belief propagation are able to produce "super-human" performance [9][10][11], i.e., fewer recognition errors than human subjects in the Speech Separation Challenge (SSC) [5][12]. The task is to compare the speech separation and speech recognition performance of computer algorithms and human subjects.

Speech separation is also useful in cochlear implant systems for improving the quality of listening of hearing-impaired people [13][14]. A person with normal hearing is good at extracting information from a specific speech source under a noisy environment with multiple sources, but it is not the case for hearing-impaired people. Incorporating single-microphone noise reduction algorithms into cochlear implants helps to improve the target speech intelligibility [13][14]. It is also becoming more

practical to implement sophisticated machine learning algorithms into speech processors, due to their improvement on the computational power and energy efficiency. The cochlear implant users can build statistical models for their friends or relatives, and select the corresponding models to perform statistical model-based speech separation for the improved hearing condition.

Single-microphone speech separation is a hard problem for computers, but human performance is surprisingly good. It is a good application for which powerful machine learning techniques are required. The recent advances of machine learning techniques have led to significant performance improvement on many speech and language applications. For example, discriminative graphical models like conditional random fields (CRF) [15][16] have been successful in the tasks such as phone recognition [17][18] and part-of-speech tagging [19]. However, there are not many works for speech separation with CRF [20]. Motivated by the success in other language applications, the application of discriminative models may lead to improved speech separation performance over the state-of-the-art factorial HMM formulations.

1.3 Research objectives

This thesis focuses on separation of two speech sources from different speakers with single corrupted speech signal. We follow the statistical model-based approach and consider single-microphone speech separation as a classification problem. We adopt a soft-decision scheme to minimize the effect of mis-classification. It is equivalent to minimum mean square error (MMSE) estimation of the sources. The prior statistical information of the sources are assumed to be available. All the necessary training data are prepared for statistical model training. The speaker identity and signal-to-signal ratio of the sources are assumed to be known.

The main theme of the thesis is to explore the application of conditional random fields (CRF) [15] on single-microphone speech separation. We apply CRF for the statistical inference of the MMSE estimator of the sources. Several CRF formulations have been developed. They are based on different graphical structures, parameter estimation criteria and methods, and the choice and organization of the

observations. We have also built up the factorial HMM baseline systems for the evaluation of CRF formulations. Various factorial HMM configurations are evaluated, including different source interaction models and inference algorithms. We also justify the application of graphical models by emphasizing the importance of modeling speech dynamics for single-microphone speech separation.

1.4 Outline of the thesis

In the next Chapter, there is a review on existing single-microphone speech enhancement and separation algorithms. The metrics for performance evaluation, including the objective metrics of perceived speech quality and machine intelligibility are introduced. The design of the training and evaluation data, and the experiments are described. The procedures for preparing the required statistical models are discussed.

Statistical model-based approach is described in Chapter 3. The derivation of the minimum mean square error estimator of the sources is presented. The statistical relationship between the sources and corrupted speech signal, referred to as an interaction model, is described and verified with experimental data. The application of Gaussian mixture models (GMM) in modeling the interaction between the observed corrupted speech signal and the source statistics is also discussed.

The application of graphical models for single-microphone speech separation is discussed in Chapter 4. After a review on graphical modeling approach for speech applications, the importance of modeling speech dynamics for single-microphone speech separation is discussed. The exact and approximated inference algorithms of factorial HMM are reviewed and evaluated. We also compare the separation performance of factorial HMM with analytically derived interaction model and the GMM modeling approach, in which the latter one becomes our baseline system.

Conditional random field (CRF) for single-microphone speech separation is investigated in Chapter 5. While CRF formulations can be considered as the discriminative counterpart of factorial HMM, their relationship and fundamental difference are discussed. The application of averaged stochastic gradient descent and approximated inference for conditional maximum likelihood parameter estimation are de-

scribed. We also discuss the design of feature functions for CRF formulations. The performance of CRF formulations is further compared with that of factorial HMM baselines.

Extensions of CRF formulations are discussed in Chapter 6. A large-margin criterion for parameter estimation is introduced and evaluated. A simplified CRF formulation is proposed. Exact statistical inference can be performed effectively in the simplified CRF formulation. We also evaluate the CRF formulations and factorial HMM on corrupted speech signals under different signal-to-signal ratios.

Towards the conclusion in Chapter 7, we compare our best experimental results with the oracle results. The oracle results assume the availability of the model parameters of the underlying sources. The conclusion of this thesis is drawn and the potential future directions are discussed.

Chapter 2

Background of research

2.1 Established methods

The development of single-microphone speech separation algorithms can be traced back to the studies of single-microphone speech enhancement. In this Chapter, we review the major approaches. Model-based methods for speech enhancement will be discussed separately in Chapter 3. Unless otherwise specified, a corrupted speech signal is represented by the following instantaneous linear additive mixing model,

$$y(t) = \sum_k a_k x_k(t) \tag{2.1}$$

where $y(t)$ is a noisy speech or speech mixture, $x_k(t)$ is the k^{th} source and a_k is the gain factor. We refer the corrupted speech signal to as noisy speech when the interfering source is noise, and to as speech mixture when the interfering sources are other speech sources.

2.1.1 Stationary noise suppression algorithms

Spectral subtraction

The basic idea of spectral subtraction [21] is to remove the noise spectrum from the spectrum of noisy speech. To perform spectral subtraction, the speech and non-speech segments of a given utterance are first identified, e.g., by voice activity detection [22]. The noise spectrum is estimated from the non-speech segments and

subtracted from the spectra of the speech segments. Spectral subtraction was commonly used in robust speech recognition [23].

Spectral subtraction may introduce low-frequency tones, known as musical noise, into the enhanced speech signals. The musical noise is perceptually even more annoying than the additive noise itself [24][25]. If the musical noise problem is treated properly, the perceived quality of the enhanced speech signals can be improved [24]. Improved spectral subtraction algorithms, such as using multi-band [26], spectral harmonics [23], and geometric approach [24] were developed to minimize the musical noise. Spectral subtraction is designed to deal with stationary noise. It is not suitable for single-microphone speech separation, since interfering speech source is non-stationary.

Subspace methods

Subspace method assumes that the Euclidean space of noisy speech is composed of a clean signal subspace, and a noise subspace [27]. Decomposition in subspace is performed by applying singular value decomposition (SVD) to the time-domain noisy signal [28] or eigenvalue decomposition (EVD) to the covariance matrix of the noisy signal [27]. The EVD approach is equivalent to Karhunen-Loève transform (KLT) in a second-order stationary process. The noise subspace is removed from the noisy speech to obtain an enhanced speech signal. The noise subspace is identified from the covariance matrix of the noise. Spectral subtraction can be considered as a special case of subspace method, in which discrete Fourier transform (DFT) instead of KLT is applied [27].

The subspace method requires that the noise is stationary and uncorrelated with the signal. If the noise is not white, whitening process can be applied. For system implementation, a voice activity detector is used to gather the noise statistics for updating the noise covariance matrix and other parameters. The dimension of signal subspace is not known in advance. Methods such as the minimum description length (MDL) principle [29] were proposed to estimate the model complexity.

Subspace method does not cause signal distortion since only the subspace containing interfering sources is identified and nulled [27]. In practice, the “clean

subspace" may contain residual components of the interfering source. Moreover, subspace method is unable to identify the subspace components of different speakers. These difficulties make the subspace method not suitable for single-microphone speech separation.

2.1.2 Algorithms based on dictionary learning

Independent component analysis (ICA)

Although the spectral subtraction and subspace approaches are not suitable for single-microphone speech separation, it is generally a good idea to decompose a noisy speech into clean and noise components. If an orthogonal basis is not required, an over-completed dictionary can be prepared for the reconstruction of speech sources. The dictionary entries referred to as the basis functions are learned from clean training data and labeled with the speaker identities, resolving the problem of the subspace method.

Several techniques have been developed for learning the basis functions for single-microphone speech separation. An attempt of dictionary learning was reported to apply independent component analysis (ICA) for time-domain basis functions [30]. ICA is originally proposed for over-determined blind source separation (BSS) [31][32]. By assuming that the sources are statistically independent and follow non-Gausssian distribution, ICA aims at finding the mixing matrix that minimizes the mutual information between the sources. For BSS with linear additive model, ICA usually operates in the time domain. Moreover, if the assumptions are fulfilled, exact solution for an over-completed BBS problem is possible [33].

In [30], a source signal is modeled as a linear combination of time-domain basis functions a_i with coefficient s_i, i.e, $x = \sum_i a_i s_i$. By defining the basis functions as the column vectors of a matrix A and rewriting $x = As$, ICA algorithm learns the inverse of the matrix, $W = A^{-1}$, which is referred to as the ICA basis filter. The coefficients are then obtained by matrix multiplication of the basis filter and the source vectors, i.e., $s = Wx$. Single-microphone speech separation problem is then formulated as a maximum a *posteriori* (MAP) problem to determine the most

probable sources that generate the speech mixture.

There is a major problem of using ICA for single-microphone speech separation. Speech signals are non-stationary in time-domain. Short-time analysis is applied for quasi-stationary speech frames. The speech frames have to be sufficiently short to satisfy the quasi-stationary assumption. However, source independence is questionable when the analysis frame is getting shorter [34]. Other dictionary learning methods, such as non-negative matrix factorization (NMF) [35][36][37], has since been investigated.

Non-negative matrix factorization (NMF)

Let N be the observation dimension and T be the number of frames in speech mixture Y. Single-microphone speech separation with NMF requires $Y \in \mathbb{R}^{N\times T}$ to be non-negative. The condition can be fulfilled in spectral domain approximately in the minimum mean square error sense. The observed speech mixture spectrum is modeled as the summation of the basis functions of the source spectra, i.e., $Y \approx BW$, where $B \in \mathbb{R}^{N\times J}$ is an over-complete dictionary containing J non-negative basis functions of the sources, and $W \in \mathbb{R}^{J\times T}$ is a matrix of non-negative weighting coefficients. The basis functions are learned from the spectral domain (either power or magnitude spectrum) of the source training data.

In the reconstruction process, only a few basis functions are dominated in the reconstructed signals. Most of the coefficients are close to zero. This promotes the formulation of single-microphone speech separation as a sparse signal recovery problem [38]. The prior assumption of sparseness is useful to improve the separation performance [37]. Sparsity constraints are applied to minimize the number of non-zero coefficients. However, the cardinality problem is generally NP-hard. The weighting coefficient matrix W is typically approximated from a convex ℓ_1-norm minimization problem with regularization factor c and the learned dictionary B,

$$\min_{W} \left\|Y - BW\right\|_F^2 + c\sum_{ij} |W_{ij}|\,, \forall W_{ij} \geq 0 \tag{2.2}$$

where $\left\|X\right\|_F$ is the Frobenius Norm. Equation 2.2 is solved efficiently by algorithms such as matching pursuit [39][40], or genetic convex optimization solvers [41].

Separation performance is further improved by applying temporal continuity constraints. In [37], temporal continuity is imposed as a cost function of neighbour frames. The cost function can be integrated into the sparse signal recovery problem for determining the weighting coefficient matrix W.

2.1.3 Computational auditory scene analysis

Computational auditory scene analysis (CASA) aims at recovering speech sources in an attempt of mimicking human auditory neural processing [42]. For speech separation, CASA operates on time-frequency representations such as a cochleagram or correlogram [43]. The goal of CASA is to estimate the binary time-frequency masks for the target sources [44]. The frequency components under the time-frequency mask are preserved while those outside the mask are considered as interference and removed. It is showed that when the ideal binary time-frequency masks are obtained, the reconstructed speech sources has much better intelligibility than the corrupted speech [45][46][47].

Traditionally, the estimation of ideal binary mask is done by grouping low-level acoustic cues of the time-frequency representation [42]. Recently it is suggested to consider binary mask estimation as a classification problem [4]. Machine learning and pattern recognition techniques such as supporting vector machine (SVM) [48][49], spectral clustering [50], and graphical models [20] have been applied to determine the binary masks. In [51], N-best outputs from speech recognition were utilized to the estimate ideal binary mask.

Multi-pitch tracking plays an important role in the CASA approach of single-microphone speech separation. It facilitates the retaining of spectral and temporal continuity of a source. A major problem of CASA is the assignment of time-frequency units to specific speakers. One of the solutions is to apply multi-pitch tracking to identify the fundamental frequencies of different speakers, and assign the time-frequency units according to the harmonic patterns corresponding to different fundamental frequencies. An early effort of incorporating multi-pitch tracking into single-microphone speech separation was proposed by Weintraub [52]. This work set up the foundation of CASA. In this study, the pitch period of each speaker at

each time frame was tracked by dynamic programming. The source spectra were estimated based on the periodicity information from multiple-pitch tracking results. A Markov model was applied to determine the voiced or unvoiced status, which helped to determine the spectral continuity of a source signal.

In [53], text transcription of the sources was assumed to be available. HMM forced alignment was performed on the speech mixture to determine the statistics of the sources based on phone-based acoustic models. Wiener filtering was performed with these source statistics. Comb filters derived from multi-pitch tracking were applied to further remove interfering harmonics. The impulse responses of the comb filters are derived from the fundamental frequencies of target speakers.

More recently, multi-pitch estimation was done with hidden Markov Models (HMM) [54][55][56][57]. In [54], a simple HMM with only three states, namely pitch decrement, constant pitch and pitch increment, was proposed. After the fundamental frequency of an individual speaker was obtained, the frequency components located at this fundamental frequency and its harmonics were assigned to the corresponding speaker. A more comprehensive algorithm of HMM based multi-pitch tracking was proposed in [55]. The paper also described a signal processing algorithm to determine the multiple hypotheses on pitch candidates. The posterior probability of the number of candidates was determined by an HMM with $K + 1$ states, which represented the state spaces of K pitches plus zero pitch. The pitch tracking results were integrated into a CASA system for speech separation. However, this algorithm did not resolve the speaker identity of an estimated pitch value [57]. A statistical method was proposed to model the interaction of the sources with factorial HMM . A 200-state HMM was used to represent a range of F0 from 80 Hz to over 1 kHz. The multi-pitch tracking algorithm was integrated with a model-based speech separation system based on source-filter model of speech generation [58]. In [56], prosodic observations of each speaker was modeled by HMM with multi-dimensional observations, including voiced or unvoiced pattern, pitch value, harmonics in voiced portion and constant amplitude for unvoiced portion. Discriminative training was applied to further improve the pitch tracking accuracy. The pitch tracking results were used with the spectral clustering algorithm for speech separation [50].

2.2 Performance evaluation of speech separation algorithms

The effectiveness of speech separation or speech enhancement algorithms can be measured by the improvement of perceived speech quality and speech intelligibility. An improvement on one criterion does not necessarily imply an improvements on the other [59]. Perceived speech quality focuses on the ease and comfort of listening. Speech intelligibility can be indicated by the recognition accuracy on the spoken content [59]. Subjective evaluation by human subjects is a well-recognized method. It is however very costly and time consuming. Objective performance metrics are designed for evaluating the reconstructed sources. The commonly used objective metrics are described below.

2.2.1 Perceptual Evaluation of Speech Quality (PESQ)

Perceptual Evaluation of Speech Quality (PESQ) is designed to predict perceived quality of speech from acoustic signals [60]. It was proposed for evaluation of different speech codecs and network distortions over telecommunication channels [61]. It is accepted as a suitable performance metric for speech enhancement [62] and speech separation [63]. The procedures of computing PESQ for reconstructed speech are described in [61]. The reconstructed speech signals and their references are first aligned to a suitable listening level. After time alignment, auditory transform is performed to obtain a set of distortion parameters in the time-frequency domain. The perceived speech quality is computed from a cognitive model with the distortion parameters. The PESQ score lies in the range of -0.5 to 4.5. A higher score indicates better perceived speech quality.

2.2.2 Blind Source Separation Evaluation Metrics (BSS_EVAL)

Blind Source Separation Evaluation Metrics (BSS_EVAL) were designed to be generic measure for different mixing conditions and separation algorithms [25]. They

are applicable to either over-determined, under-determined or single-microphone speech separation. The metrics require the availability of the reference source signal. A reconstructed source signal are compared with the reference source signal. As suggested in [25], the recovered source $\hat{x}_k$ is expressed as $\hat{x}_k = \hat{x}_{k,target} + e_{k,interf} + e_{k,artif} + e_{k,noise}$, where $\hat{x}_{target}$ is the target signal, $e_{k,interf}$, $e_{k,artif}$ and $e_{k,noise}$ are the interferences, signal artifacts and noise error terms respectively. These terms are obtained by applying orthogonal projections on the reference signal and the reconstructed signal.

In single-microphone speech separation, we assume the noise error is zero, i.e., $e_{k,noise} = 0$, and the source-to-noise ratio (SNR) tends to ∞. The other three evaluation metrics, source-to-distortion ratio (SDR), signal-to-interferences ratio (SIR) and source-to-artifacts ratio (SAR) are defined as,

$$\text{SDR} := 10\log_{10}\frac{||\hat{x}_{k,target}||^2}{||e_{k,interf} + e_{k,artif} + e_{k,noise}||^2} \tag{2.3}$$

$$\text{SIR} := 10\log_{10}\frac{||\hat{x}_{k,target}||^2}{||e_{k,interf}||^2} \tag{2.4}$$

$$\text{SAR} := 10\log_{10}\frac{||\hat{x}_{k,target} + e_{k,interf} + e_{k,noise}||^2}{||e_{k,artif}||^2}. \tag{2.5}$$

SDR can be regarded as a global performance measure of the output signal. SIR measures the ability of a separation or enhancement algorithm in suppressing interfering signals. For speech separation algorithms operating in the frequency domain, SIR evaluates the degree of suppressing on the frequency components of interfering sources. SAR quantifies the burbling artifacts introduced by a separation or enhancement algorithm. These artifacts are caused by spectral attenuation or amplification of the reconstructed sources. Spectral attenuation occurs when the magnitude of enhanced frequency component is smaller than that of the reference component. Spectral amplification occurs when the magnitude of enhanced component is greater than that of the reference component [64][65].

2.2.3 Speech recognition accuracy

A few objective measures were proposed to evaluate speech intelligibility of reconstructed speech [66][65]. PESQ is considered representative indication of speech in-

Table 2.1: Sentence structure of the GRID Corpus. Adapted from [67]

command	colour	preposition	letter	digit	adverb
bin	blue	at	A-Z	1-9	again
lay	green	by	excluding W	zero	now
place	red	in			please
set	white	with			soon

telligibility [66]. Speech separation performance can also be evaluated from the perspective of machine understanding of reconstructed speech sources. In fact, speech separation is often used as a preprocessing step of many speech applications, including speech recognition.

Word Error Rate (WER) is a standard measure of speech recognition performance. It measures the Levenshtein distance (the minimum numbers of insertion, deletion and substitution) between the recognized word sequence and the reference one. A speech recognition system that makes fewer word errors is considered to have better performance. In speech separation, automatic speech recognition can be performed on a reconstructed source. In this way, WER becomes a performance metric for speech separation system.

2.3 Corpus and experiment setting

In this thesis, single-microphone speech separation experiments are performed with speech data from the GRID Corpus [67]. This corpus consists of speech materials from 34 speakers, including 18 male and 16 females. It was created as a controlled corpus for small-vocabulary command recognition. The grammar and the vocabulary are listed in Table 2.1.

Three sets of speech mixtures, namely *Male+Male*, *Male+Female* and *Female+Female,* are prepared from the speech of 3 male and 3 female speakers according to the instantaneous additive model at the desired signal-to-signal ratios. The speaker combinations are shown in Table 2.2. The signal-to-signal ratio (SSR) is

defined as the power ratio between the target source and the masking sources,

$$\text{SSR (dB)}=10\log_{10}\frac{\text{averaged power of the target source}}{\text{averaged power of the masking sources}}. \tag{2.6}$$

All speech data are sampled at 16 kHz. For each speaker, 450 clean utterances are used as training data, and 50 unseen utterances are used for evaluation. Each set of mixture data consists of 2500 speech mixtures for evaluation, and over 200,000 speech mixtures for model training.

Table 2.2: Configuration and speaker ID of 3 sets of speech mixtures

	Male+Male	**Male+Female**	**Female+Female**
Speaker 1	1 (Male)	17 (Male)	24 (Female)
Speaker 2	2 (Male)	18 (Female)	25 (Female)

The choice of speaker pairs is made with considerations on pitch range differences. Pitch range is defined as the range from the lowest fundamental frequency to the highest fundamental frequency produced by a speaker. Figure 2.1 shows the histograms of pitch ranges of the three speaker pairs. The *Male + Male* pair represents the case that the pitch ranges of the two speakers are overlapped. The *Male + Female* pair represents the case with distinct pitch ranges. The *Female + Female* pair lies in the middle of the two extremes. For single-microphone speech separation, speech sources with overlapped pitch range are expected to be the most difficult for separation.

Two tasks are designed for evaluating speech separation algorithms. The first task is to reconstruct individual speech sources from a speech mixture of two different speakers. The speaker identities and the signal-to-signal ratio are assumed to be known. Speaker-dependent acoustic models are also available. The speech quality of reconstructed sources is evaluated in terms of the objective quality measures, including PESQ [60] and BSS_EVAL metrics [25]. The PESQ values of the speech mixtures at 0 dB signal-to-signal ratio are given in Table 2.3. The results are obtained by averaging the PESQ values of the speech mixtures with respect to the two reference sources over the entire evaluation set. Bad speech quality with possibly very

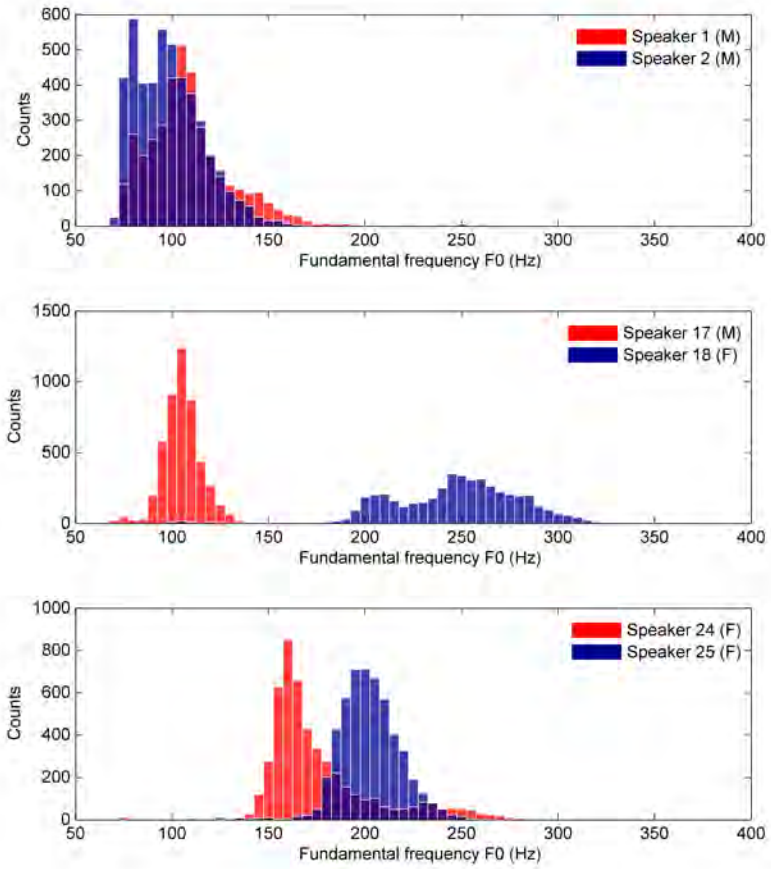


Figure 2.1: Fundamental frequency distribution of the speaker pairs

Table 2.3: The PESQ of the speech mixtures at 0 dB signal-to-signal ratio

Speaker pair	**M (1) + M (2)**	**M (17) + F (18)**	**F (24) + F (25)**	**Overall**
PESQ	1.68	1.64	1.55	1.62

annoying impairment of the speech mixtures is indicated by PESQ values below 2 [68].

The second task is a speech recognition experiment with reconstructed sources. A standard automatic speech recognition system is used. The system is prepared with the HTK [69]. The acoustic models are trained with clean speech. The acoustic features are standard Mel-frequency cepstral coefficients. A grammar network specifying the command recognition task of GRID Corpus is used for speech recognition. For clean speech, the word error rate is less than 1% with standard implementation.

2.4 Acoustic model training

Two types of acoustic models are prepared to support speech separation and speech recognition applications. For each speaker, 450 clean utterances are used for training. Short-time feature extraction is done with Hamming window of 32 ms and frame shift of 10 ms.

2.4.1 Speaker-dependent acoustic models for speech separation

For model-based speech separation, speaker-dependent acoustic models are developed for all of the 6 speakers. The acoustic features are 257-dimension log-magnitude spectrum, which results from 512-point fast Fourier transform. The speech data of an individual speaker are clustered. Each cluster corresponds to a state in the acoustic model. The emission probability of each acoustic state therefore follows a multivariate Gaussian distribution with diagonal covariance matrix. The prior probabilities of the acoustic states are estimated according to the size of the clusters. For each speaker, 3 acoustic models with 16, 128 and 512 states are prepared. Due to the large feature dimension of the speech data, each frame is typically dominated by only one acoustic state [70].

The transition between acoustic states is unrestricted as shown in Figure 2.2. The transition probabilities are estimated by Viterbi decoding procedure. The initial state sequences are first obtained by decoding the source training data with the prior probabilities. The transition probabilities are updated by computing the conditional probability of the current state given the previous state. The state sequences are then updated with the new HMM model parameters. This process is repeated several times. The final acoustic state sequences of the sources are used as the reference labels for subsequent model training tasks.

2.4.2 Speaker-independent models for speech recognition

Standard GMM-HMM based speaker-independent acoustic models are developed for speech recognition. Speech data from all 34 speakers in the GRID corpus are

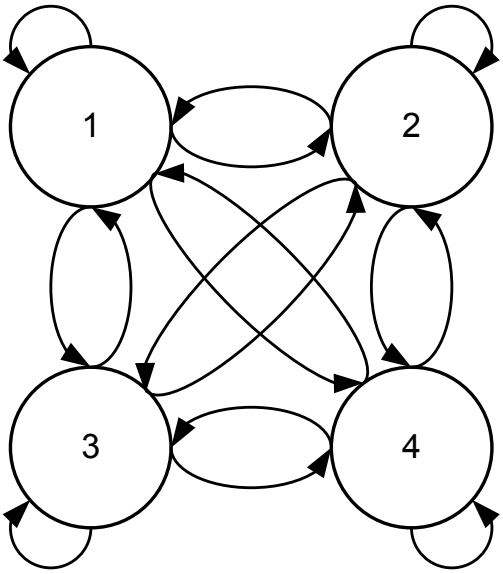


Figure 2.2: An HMM model of 4 states with unrestricted transition

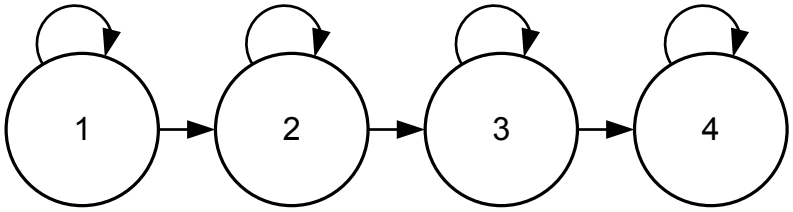


Figure 2.3: An conventional left-to-right HMM topology for speech recognition

used for training the speaker-independent models. Word models are trained with 39-dimensional Mel-frequency cepstral coefficients (MFCC) features (12 coefficients + log-energy + delta and delta-delta coefficients). Conventional left-to-right HMM topology as shown in Figure 2.3 is employed. There are 4 to 8 acoustic states in each word model depending on the number of phonemes in the word. There are 32 Gaussian components with diagonal covariance matrices at each state.

2.5 Terminology and notations

There are three basic elements in statistical model-based single-microphone speech separation. They are the speech mixture y, the sources x, and the acoustic states s that generate the sources. The speech mixture y and the sources x are vectors and s

are labels to index the states. The subscript k is used to index the source and K to denote the number of sources, the subscript t to index the frame, the subscript f to denote the component of a feature vector. For examples, $x_{k,t,f}$ denotes the frequency component f of source k at frame t, and $y_{t,f}$ represents the frequency component f of the speech mixture at frame t.

We use bold-face symbol to represent a time sequence. For example, $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ is an observation sequence of length T and $\mathbf{y_{1\ldots t}}$ is a sub-sequence from 1 to t. Indices in sets are omitted for brevity, e.g., $\{x_k\}_{k=1}^{K} = \{x_1, x_{2,} \ldots x_K\}$ abbreviated as $\{x_k\}$ by assuming that the total number of the members of the set is known. A constant index is parenthesized, i.e., $\{x_{k,(t)}\} = \{x_{1,t}, x_{2,t}, \ldots, x_{K,t}\}$. Other short-hand notations includes $\mathrm{d}x_{1\ldots K} = \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_K$.

We use $p(\cdot)$ to denote probability distribution, $p(\cdot|\cdot)$ to denote conditional distribution and $\mathbb{E}(\cdot)$ to denote expectation. For indexing of operators, we use the following conventions, for example $\sum_{s_k} p(s_k) = \sum_{s_k=1}^{S_k} p(s_k)$ where S_k is the total number of acoustic states for speaker k, and $\sum_{\{s_k\}} p(\{s_k\}|\mathbf{y}) = \sum_{s_1=1}^{S_1} \sum_{s_2=1}^{S_k} \ldots \sum_{s_K=1}^{S_K} p(s_1, s_{2,} \ldots, s_K|\mathbf{y})$ for the operation over a sequence. We use the notation $\setminus$ to represent exclusion of indices, for example, $\sum_{k\setminus k=2} p(s_k) = p(s_1) + p(s_3) + p(s_4) + \ldots + p(s_K)$.

Chapter 3

Statistical model-based methods

3.1 Overview

Speech enhancement and source separation are classical problems of statistical signal processing. Minimum mean square estimation (MMSE) is a common criterion for source estimation. It can be achieved by Wiener filter [71] or Kalman filter [72]. In [73], an MMSE source estimator was derived by measuring the noise statistics from non-speech portion of noisy speech. This work also showed that the phase spectrum of noisy speech is the MMSE phase estimator given the MMSE magnitude estimator.

A statistical model-based method is designed to recover speech sources by modeling the relation between the sources and the mixture. The method assumes the availability of prior knowledge, including but not limited to, the spectral and temporal characteristics of each speech source. These characteristics are specified in the form of probability distributions at the states of acoustic models. Use of HMM for speech enhancement was proposed in [2] and further developed in [3]. An HMM can be unfolded in time as a linear Markov chain, leading to the graphical modeling approach to single-microphone speech separation. Factorial HMM [74] was proposed in [9][10][11]. The linear Markov chains of individual sources are coupled to produce the observed speech mixture. In log-spectral domain, the process is usually modeled by approximations such as ALGONQUIN [75] based on the log-sum-exp expression or the *mixture-maximization* (MIXMAX) model [76][77][78]. Other graphical models investigated include restricted Boltzmann machine [79], and conditional random

fields (CRF) [80][81]. For single-microphone speech separation, statistical model-based methods do not require multi-pitch tracking. This is an advantage over the CASA approaches.

The use of HMM makes a close connection between speech separation and other speech processing problems. HMM-based speech synthesis can be used to reconstruct the speech sources [82]. The noisy spectrum was first analyzed with the missing data approach [83]. The frequency components that were dominated by only one source were identified as the "reliable part". The unreliable parts of the reconstructed sources were re-synthesized with the statistics of the acoustic models and the constraints from the reliable parts.

Various techniques of robust speech recognition are also based on statistical speech models. The missing data approach [83] was shown to resemble a mean-field approximation of the MIXMAX model [84]. In cesptral domain, the Vector Taylor Series (VTS) [85] and parallel model combination (PMC) [86] were based on a log-sum-exp expression of mixing speech and noise. PMC assumed that the power spectra of the speech mixture, the source and the noise all followed log-normal distribution (or Gaussian distributions in log-spectral domain). The speech mixture distribution were derived by moment matching from the distributions of the source and the noise. VTS approaches did not assume a specific distribution of speech mixture, but attempted to linearize the log-sum-exp expression with a Taylor series approximation.

Figure 3.1 illustrates the flow diagram of a typical model-based single-microphone speech separation system. Given prior information of the sources, the sources can be reconstructed with the minimum mean square error (MMSE) or the maximum a *posteriori* (MAP) criteria. In the MAP criterion, the sources are obtained by filtering the speech mixture with the statistics of the most probable sources. The most probable acoustic state sequences are required. In the MMSE criterion, the reconstructed source is the mean of the statistical filter outputs. The posterior probabilities of the acoustic states are obtained by statistical inference.

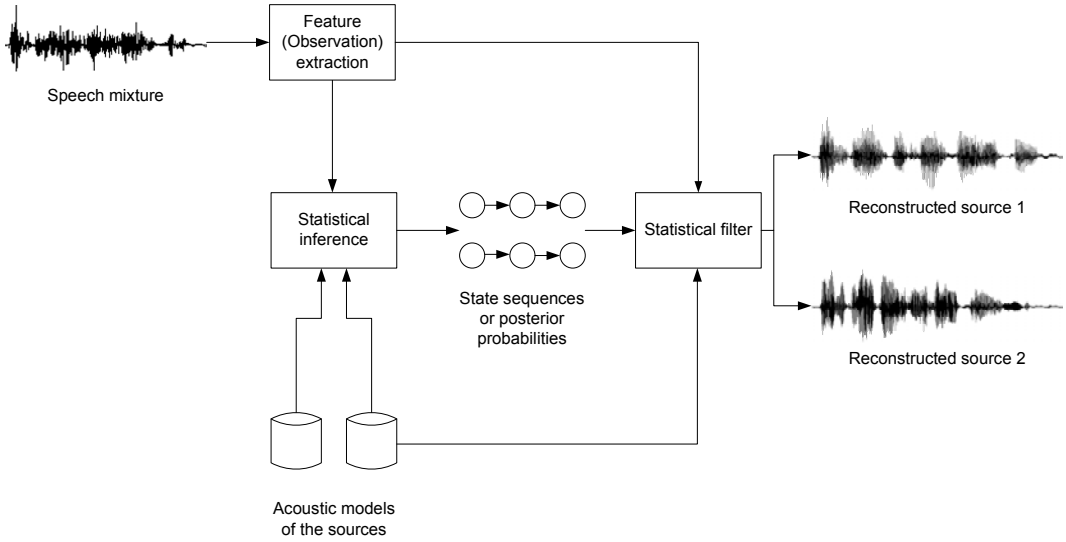


Figure 3.1: A flow diagram of a typical single-microphone speech separation system based on statistical model-based methods

3.2 Conditional MMSE estimation

Let $(\mathbf{y}, \{\mathbf{x}_k\})$ be the set of observation sequences from the speech mixture and K speech sources. The sources $\{\mathbf{x}_k\}$ are unobserved, but their characteristics are described by the acoustic models. A statistical model-based method aims at computing $\mathbb{E}(\mathbf{x}_k|\mathbf{y})$ by modeling $p(\{\mathbf{x}_k\}|\mathbf{y})$ as,

$$p(\{\mathbf{x}_k\}|\mathbf{y}) = \sum_{\{\mathbf{s}_k\}} p(\{\mathbf{x}_k\}|\mathbf{y}, \{\mathbf{s}_k\}) p(\{\mathbf{s}_k\}|\mathbf{y}) \,. \tag{3.1}$$

At each frame, we have

$$p(\{x_{k,(t)}\}|\mathbf{y}) = \sum_{\{s_{k,(t)}\}} p(\{x_{k,(t)}\}|\mathbf{y}, \{s_{k,(t)}\}) p(\{s_{k,(t)}\}|\mathbf{y}) \,. \tag{3.2}$$

After marginalizing $p(\{x_{k,(t)}\}|\mathbf{y}, \{s_{k,(t)}\})$ into $p(x_{k,(t)}|\mathbf{y}, \{s_{k,(t)}\})$, the sources $\hat{x}_{k,t}$ are reconstructed by either MMSE estimator

$$\mathbb{E}(x_{k,t}|\mathbf{y}) = \sum_{\{s_{k,(t)}\}} p(\{s_{k,(t)}\}|\mathbf{y}) \times \mathbb{E}(x_{k,t}|\mathbf{y}, \{s_{k,(t)}\}) \,. \tag{3.3}$$

or the MAP estimator

$$\hat{x}_{k,t} = \mathbb{E}(x_{k,(t)}|\mathbf{y}, \{s^*_{k,(t)}\}) \,, \tag{3.4}$$

from the most probable state sequences $\{\mathbf{s}^*_\mathbf{k}\} = \arg\max_{\{\mathbf{s_k}\}} p(\{\mathbf{s_k}\}|\mathbf{y})$. For HMM-based acoustic models, the observations are conditionally independent. By the

Bayes' theorem, $p(\{x_{k,(t)}\}|\mathbf{y}, \{s_{k,(t)}\})$ is expressed as

$$p(\{x_{k,(t)}\}|\mathbf{y}, \{s_{k,(t)}\}) = \frac{p(y_t|\{x_{k,(t)}\}) \prod_k p(x_{k,t}|s_{k,t})}{p(y_t|\{s_{k,(t)}\})}\,, \tag{3.5}$$

where $p(x_{k,t}|s_{k,t})$ is the emission probability from the acoustic model of x_k. The likelihood $p(y_t|\{x_{k,(t)}\})$ is referred to as an interaction model [11], which is the conditional probability of the mixture observations given the sources. The state-level likelihood $p(y_t|\{s_{k,(t)}\})$ is also derived from the interaction model as

$$p(y_t|\{s_{k,(t)}\}) = \int\cdots\int_{-\infty}^{\infty} p(y_t|\{x_{k,(t)}\}) \prod_k p(x_k|s_k) \mathrm{d}x_{1,t} \cdots \mathrm{d}x_{K,t}\,. \tag{3.6}$$

We refer $p(y_t|\{s_{k,(t)}\})$ to as *state-level interaction model.*

The computation of $\mathbb{E}(x_{k,t}|\mathbf{y}, \{s_{k,(t)}\})$ and $p(\{s_{k,(t)}\}|\mathbf{y})$ are the key problems in speech separation. In this chapter, we review the methods of computing $\mathbb{E}(x_{k,t}|\mathbf{y}, \{s_{k,(t)}\})$, especially when the observations are in log-spectral domain. The computation of the frame-level posterior probability $p(\{s_{k,(t)}\}|\mathbf{y})$ is the main focus of the subsequent chapters.

3.3 Interaction model

3.3.1 Derivation of exact interaction model

Instantaneous additive mixing model in the time domain, i.e., $y(t) = \sum_k x_k(t)$ is assumed. After short-time frame processing, $\{X_k\}$ and Y are the Fourier transform of the sources and the mixture at a specific time frame. Let $|Y|$, $\{|X_k|\}$ denote the magnitude spectra of the speech mixture and the sources. The power spectrum of the speech mixture is given as (the frame index t and frequency index f are dropped for brevity),

$$|Y|^2 = \sum_k |X_k|^2 + \sum_{jk\backslash j=k} |X_j||X_k| \cos(\theta_j - \theta_k)\,, \tag{3.7}$$

where θ_k and θ_j are the phase spectrum of sources k and j. It is noted that the phase of a speech source is uniformly distributed [87]. The probability density function of the phase difference $\Delta\theta_{jk} = \theta_j - \theta_k$ is therefore assumed to be symmetric at $\Delta\theta_{jk} = 0$.

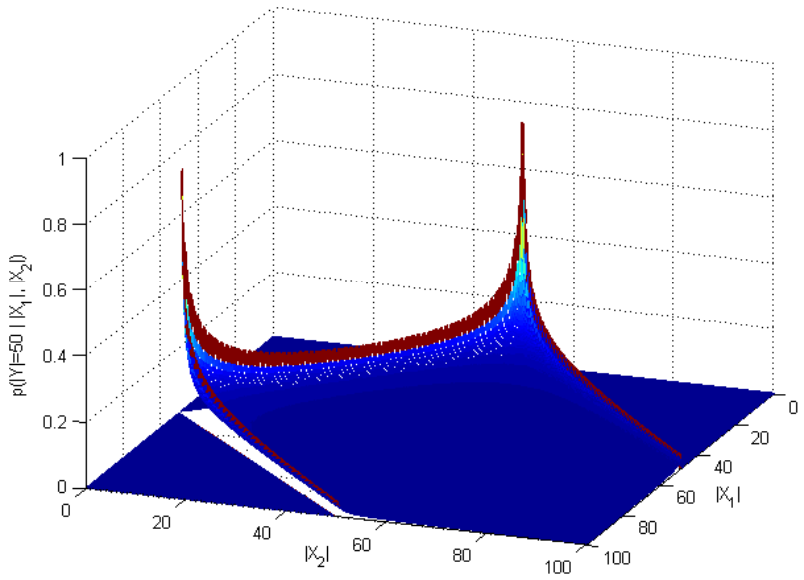


Figure 3.2: The shape of the "two-wave envelope pdf" $p(|Y|\big||X_1|,|X_2|)$ at $|Y|=50$

For the exact interaction model, the effect of superposition is considered. The general form of the interaction model $p\Big(|Y|\Big|\{|X_k|\}\Big)$ is derived by [88],

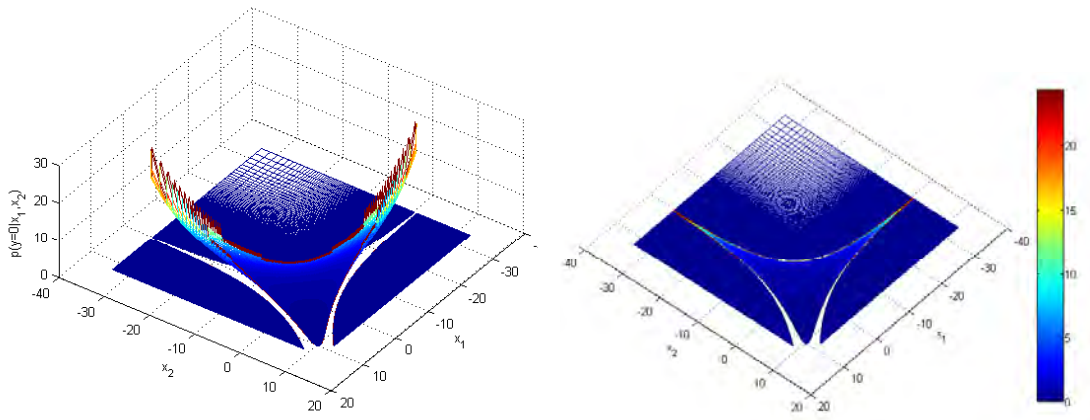
$$p(\Big(|Y|\Big|\{|X_k|\}\Big) = |Y|\int_0^\infty J_0(|Y|q)\Big[\prod_k J_0(|X_k|q)\Big]qdq\,, \tag{3.8}$$

where $J_0(u)=\frac{1}{2\pi}\int_{-\pi}^{\pi}e^{-j(-u\sin(\tau))}\mathrm{d}\tau$ is the zeroth-order Bessel function of the first kind. The derivation is provided in Appendix A. In the two-source case, the exact interaction model is given by [88],

$$p\Big(|Y|\Big||X_1|,|X_2|\Big) = \begin{cases} \frac{2|Y|}{\pi\sqrt{4|X_1|^2|X_2|^2-(|X_1|^2+|X_2|^2-|Y|^2)^2}} & \text{for } \Big||X_1|-|X_2|\Big| < |Y| < |X_1|+|X_2| \\ 0 & \text{otherwise}\,. \end{cases} \tag{3.9}$$

This function is referred to as "two-wave envelope probability density function (pdf)" in communication [88]. Figure 3.2 illustrates the shape of this function for a specific frequency component.

The function has a geometric interpretation. The square root term is the area of a triangle with $|X_1|$, $|X_2|$ and $|Y|$ being the three edges. The function is singular when the area of the triangle becomes arbitrarily small. This happens when the two



(a) (b)

Figure 3.3: Illustrations of the "devil function" $p(y|x_1, x_2)$ at $y = 0$ dB ; a) the 3D-viewpoint, horizontal axes are the power of the sources (in dB), vertical axis is the probability densities, b) the projected view of the function

sources are perfectly in phase ($\Delta\theta_{1,2} = 0$), i.e., $|Y| = |X_1| + |X_2|$ or totally out of phase ($\Delta\theta_{1,2} = \pi$), i.e., $|Y| = \Big||X_1| - |X_2|\Big|$, or when there exists only one source, i.e., the amplitude of the other source is zero.

The interaction model may also be expressed in the log-power domain. By substituting $x_k = \log(|X_k|^2)$, $y = \log(|Y|^2)$ and applying change of variables for probability density functions, the following distribution is obtained,

$$p\Big(y\Big|x_1, x_2\Big) = \begin{cases} \dfrac{e^{y-\frac{x_1+x_2}{2}}}{\pi\sqrt{1-\frac{1}{4}(e^{y-\frac{x_1+x_2}{2}}-e^{\frac{x_1-x_2}{2}}-e^{-\frac{x_1-x_2}{2}})^2}} & \text{for } \Big|e^{\frac{x_1}{2}}-e^{\frac{x_2}{2}}\Big| < e^{\frac{y}{2}} < e^{\frac{x_1}{2}}+e^{\frac{x_2}{2}} \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

which is referred to as the "devil function" in [89], as illustrated in Figure 3.3. Note that in [89], the "devil function" is derived by assuming uniform distribution of the phase difference $\Delta\theta_{1,2}$. However, since the phases are uniformly distributed between $[-\pi, \pi]$ [87], the phase difference should follow a triangular distribution,

$$p(\theta) = \begin{cases} \frac{1}{2\pi}(1 - |\frac{\theta}{2\pi}|) & \text{for } |\theta| \leq 2\pi \\ 0 & \text{otherwise}\,. \end{cases} \tag{3.11}$$

The derived "devil function" is the same.

Similar to the "two-wave envelope pdf", there are singularity points in the "devil function", making it computationally intractable. Different deterministic approximations are made on the exact interaction model. In the log-spectral domain, the MIXMAX model [76][77], ALGONQUIN [75] and its phase-sensitive variation [90] are the most common approximations.

3.3.2 The mixture-maximization interaction model

If the speech sources have distinct pitch ranges, their spectra are not expected to have much overlap. The terms $|X_j||X_k|$ in Equation 3.7 become negligible. This is referred to as *W-disjoint orthogonality* [91]. It is the foundation that supports the use of ideal binary time-frequency masks in CASA [44]. The observation that a frequency component of the speech mixture is usually dominated by the stronger source also motivates the *mixture-maximization* (MIXMAX) model for robust speech applications [77]. In log-spectral domain, let $x_k = \log(|X_k|^2)$ and $y = \log(|Y|^2)$, Equation 3.7 is re-written as,

$$e^y = \sum_k e^{x_k} + \sum_{jk \setminus j=k} \exp(\frac{x_j + x_k}{2}) \cos(\theta_j - \theta_k)\,. \tag{3.12}$$

Since $\cos(\cdot)$ is an even function, we have $\mathbb{E}\{\cos(\Delta\theta_{jk})\} = 0$. By taking expectation on Equation 3.12, a non-linear MMSE estimator of the speech mixture given the sources is obtained [87],

$$\hat{y} = \log(\mathbb{E}(e^y|\{e^{x_k}\})) = \log(\textstyle\sum_k e^{x_k})\,. \tag{3.13}$$

The MIXMAX model is obtained by applying the soft-maximum approximation [77],

$$\hat{y} = \log(\textstyle\sum_k e^{x_k}) \approx \max(\{x_k\})\,. \tag{3.14}$$

The corresponding interaction model is

$$p(y|\{x_k\}) = \delta(y - \max(\{x_k\})) \tag{3.15}$$

where $\delta(\cdot)$ is the Dirac delta function. The MIXMAX model assumes that there is only one dominating source at each frequency component in the speech mixture. The amplitude of the dominating source is the same as the observed value of the

frequency component. Following [77][78][11], $p(y_f|\{s_k\})$ is derived as (t is omitted for brevity),

$$p(y_f|\{s_k\}) = \sum_k g_{x_k}(y_f|s_k) \prod_{j\backslash j=k} G_{x_j}(y_f|s_j) \tag{3.16}$$

where

$$\begin{aligned} g_{x_k}(y_f|s_k) =& p(x_{k,f} = y_f|s_k)\,, \\ G_{x_k}(y_f|s_k) =& p(x_{k,f} < y_f|s_k) \\ =& \int_{-\infty}^{y_f} p(x_{k,f} = u|s)\mathrm{d}u\,. \end{aligned}$$

Assuming independence among frequency components, $p(y|\{s_k\})$ is expressed as,

$$p(y|\{s_k\}) = \prod_f \sum_k g_{x_k}(y|s_k) \prod_{j\backslash j=k} G_{x_j}(y_f|s_j)\,. \tag{3.17}$$

By observing that $p(y_f|\{s_k\}) = p(y_f, x_{k,f} = y_f|\{s_k\}) + p(y_f, x_{k,f} < y_f|\{s_k\})$, $p(x_{k,f}|y_f, \{s_k\})$ is obtained as,

$$\begin{aligned} & p(x_{k,f}|y_f, \{s_k\}) \\ =& \begin{cases} \frac{g_{x_k}(y_f|s_k)\prod_{j\backslash j=k} G_{x_j}(y_f|s_j)}{\sum_k g_{x_k}(y_f|s_k)\prod_{j\backslash j=k} G_{x_j}(y_f|s_j)} & \text{if } x_{k,f} = y_f \\ 1 - \frac{g_{x_k}(y_f|s_k)\prod_{j\backslash j=k} G_{x_j}(y_f|s_j)}{\sum_k g_{x_k}(y_f|s_k)\prod_{j\backslash j=k} G_{x_j}(y_f|s_j)} & \text{otherwise}\,, \end{cases} \end{aligned} \tag{3.18}$$

where $p(x_{k,f} = y_f|y_f, \{s_k\})$ represents the probability that $x_{k,f}$ dominates in the speech mixture given $\{s_k\}$ and y_f, $p(x_{k,f} < y_f|y_f, \{s_k\})$ is the probability that $x_{k,f}$ is masked by other sources given $\{s_k\}$ and y_f. According to the MIXMAX model, the maximum value of $x_{k,f}$ is y_f, and we have $p(x_{k,f} < y_f|y_f, \{s_k\}) + p(x_{k,f} = y_f|y_f, \{s_k\}) = 1$. The conditional expectation $\mathbb{E}\{x_{k,f}|y_f, \{s_k\}\}$ is expressed as [78][11],

$$\mathbb{E}(x_{k,f}|y_f, \{s_k\}) = \rho_{k,f} y_f + (1 - \rho_{k,f})\mathbb{E}(x_{k,f}|x_{k,f} < y_f, \{s_k\}) \tag{3.19}$$

where $\rho_{k,f} = p(x_{k,f} = y_f|y_f, \{s_k\})$.

When the emission probabilities of the acoustic states follow multivariate Gaussian distributions with mean $\mu_{k,f}$ and variance $(\sigma_{k,f})^2$, the expectation $\mathbb{E}(x_{k,f}|x_{k,f} < y_f, \{s_k\}) = \mu_{k,f} - \frac{(\sigma_{k,f})^2 g_{x_{k,f}}(y_f|s_k)}{G_{x_{k,f}}(y_f|s_k)}$ follows a truncated Gaussian distribution [11]. The corresponding linear spectrum of the sources follows a log-normal distribution. The mean in log-spectral domain corresponds the median in linear spectral domain, which is a robust estimator.

By relaxing Equation 3.15 as a zero-mean Gaussian distribution with an arbitrary small variance σ^2, there is an approximation for $\mathbb{E}(x_{k,f}|y_f, \{s_k\})$ referred to as soft-mask filtering [92]. Soft-mask filtering approach approximates

$$\begin{aligned}&\mathbb{E}(x_{k,f}|y_f, \{s_k\}) \\ &= \begin{cases} \frac{\sigma_{k,f}^2}{\sigma_{k,f}^2+\sigma^2} y_f + \frac{\sigma^2}{\sigma_{k,f}^2+\sigma^2}\mu_{k,f} & \text{for } k = \arg\max_{\mu_{k,f}}(\{\mu_{k,f}\}) \\ \mu_{k,f} & \text{otherwise}\,. \end{cases}\end{aligned} \tag{3.20}$$

with mean vector μ_k and diagonal covariance $\Sigma_k = \text{diag}[\sigma_{k,1}^2 \ldots \sigma_{k,f}^2 \ldots \sigma_{k,F}^2]$. Soft-mask filtering is generally inferior to the MMSE estimators of Equation 3.19.

3.3.3 GMM modeling approach to state-level interaction model

As discussed in the previous section, the state-level interaction $p(y_t|\{s_{k,(t)}\})$ plays an important role in computing the conditional mean $\mathbb{E}(x_{k,t}|\mathbf{y}, \{s_{k,(t)}\})$. In addition to analytical derivation, an empirical $p(y_t|\{s_{k,(t)}\})$ can also be learned from training data. If we model $p(y_t|\{s_{k,(t)}\})$ for each pair of $\{s_{k,(t)}\}$ with multivariate Gaussian distribution, the probability $p(y_t)$ of the current frame can be represented by Gaussian mixture models (GMM),

$$p(y_t) = \sum_{\{s_{k,(t)}\}} p(\{s_{k,(t)}\}) p(y_t|\{s_{k,(t)}\})\,, \tag{3.21}$$

where the prior probabilities $p(\{s_{k,(t)}\}) = \prod_k p(s_k)$ can be regarded as the weights on the Gaussian components. The maximum number of Gaussian components is S^K. A large amount of training data for a well-trained statistical model is expected.

3.4 Empirical statistics from experimental data

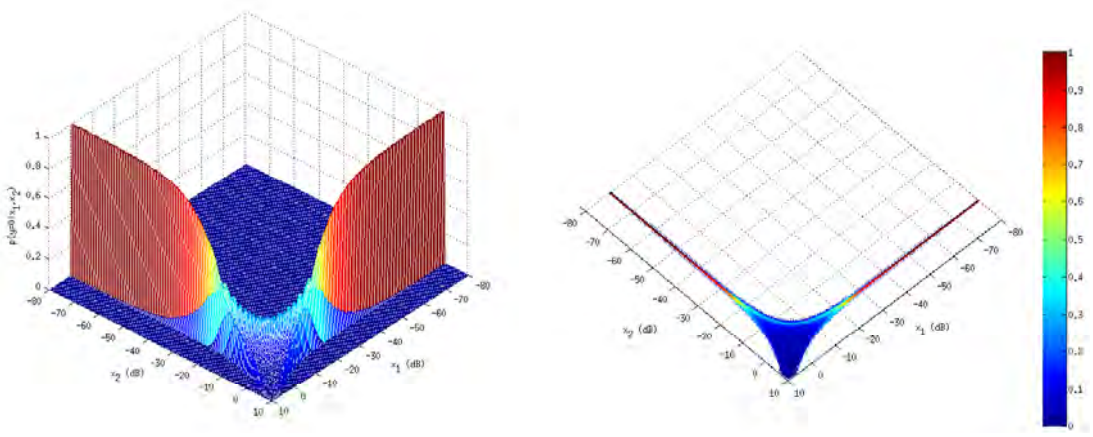
In this section, we study the empirical interaction model with speech data from the GRID Corpus [67]. We compare the empirical interaction model with the exact interaction model and the MIXMAX model. Following [89], we create 6000 speech mixtures from the source signals of 3 speaker pairs (*Male* + *Male*, *Male* + *Female*

and *Female* + *Female*). The mixing process is based on the linear additive model at 0 dB signal-to-signal ratio. Assuming the frequency components are independent, the empirical likelihood $p(y_f|x_{1,f}, x_{2,f})$ is shown as in Figure 3.4, at 0 dB log-power for speech mixtures.

The result from Figure 3.4 indicates that when two sources have sufficiently different power level (over 25 dB in our data) at a given frequency component, the source with higher power dominates the mixture. The empirical probability that the log-power of the mixture is equal to the dominating source approaches to one, i.e., the MIXMAX model is a good and practical approximation of the sources in this situation. However, when the power of the two sources are close, the MIXMAX model deviates significantly from the empirical distribution. In this case, the shape of empirical probability density function reflects the exact interaction model.

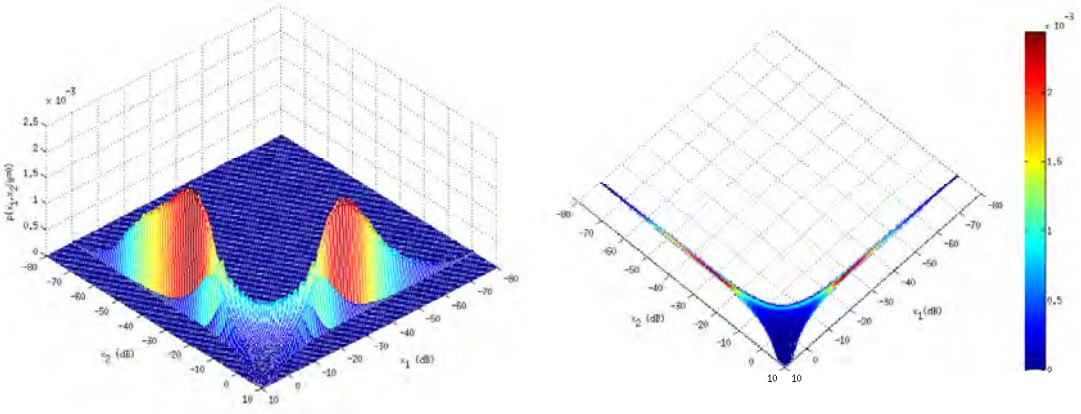
The next question is whether the log-amplitudes of the sources always differ significantly, such that the condition of the MIXMAX model is satisfied. The frame-level posterior probability $p(x_{1,f}, x_{2,f}|y_f)$ is shown in Figure 3.5. It is shown that $p(x_{1,f}, x_{2,f}|y_f)$ is peaked at the position where the log-powers of the sources are significantly different ($\sim$30 dB). These empirical results indicate that the MIXMAX model is a practical model with reasonable accuracy.

The short-time log-power spectra of a speech mixture at 0 dB signal-to-signal ratio and the corresponding sources (*Male* + *Female*) are shown in Figure 3.6. The red circle indicates a frequency component dominated by source 1. At this frequency, the power of source 1 is higher than that of source 2 by 37 dB. Similarly, the square in brown indicated a frequency component dominated by source 2. The power of source 2 is higher than source 1 by about 21 dB. By inspection, the MIXMAX model is quite accurate in these two cases. However, in the third case marked with purple triangle, the MIXMAX model becomes inaccurate. The power of the two source are close to each other (only about 1 dB difference). The power of the resultant speech mixture is 5 dB greater than the stronger source.



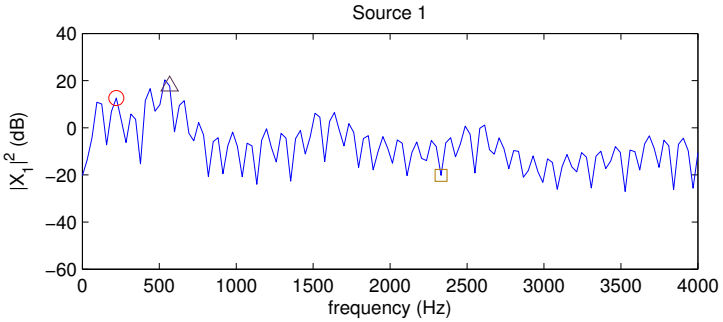
(a) (b)

Figure 3.4: 3D plot of empirical likelihood $p(y_f|x_{1,f}, x_{2,f})$ at 0 dB log-power for speech mixtures; a) the 3D-plot of the density function; b) projected view of the probability density function

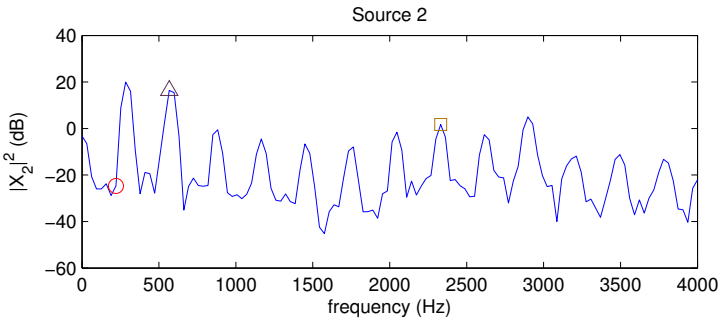


(a) (b)

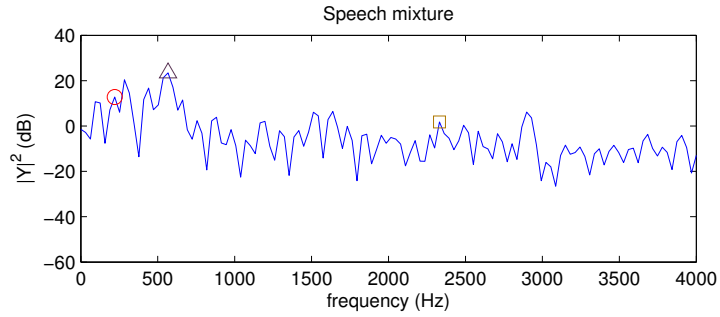
Figure 3.5: a) 3D plot of empirical posterior probability $p(x_{1,f}, x_{2,f}|y_f)$ with speech mixture $y_f = 0$ dB log-power; b) projected view



(a) Source 1 (Male)



(b) Source 2 (Female)



(c) Speech mixture

Figure 3.6: Speech frames in log-power spectra of (a) the speech mixture and the corresponding (b) male and (c) female sources

Chapter 4

Single-microphone speech separation with factorial HMM

4.1 Posterior probabilities of source states

Recall that the MMSE estimator $\mathbb{E}(x_{k,t}|\mathbf{y}) = \sum_{\{s_{k,(t)}\}} p(\{s_{k,(t)}\}|\mathbf{y})\mathbb{E}(x_{k,t}|\mathbf{y}, \{s_{k,(t)}\})$, the posterior probability $p(\{s_{k,(t)}\}|\mathbf{y})$ can be considered as a non-negative weight on $\mathbb{E}(x_{k,t}|\mathbf{y}, \{s_{k,(t)}\})$. Let $(\mathbf{y}, \{\mathbf{s_k}\})$ be the observed mixture and the underlying state sequences. The Bayesian approach attempts to compute $p(\{s_{k,(t)}\}|\mathbf{y})$ by conditioning and marginalizing the joint probability

$$p(\{\mathbf{s}_k\}, \mathbf{y}) = p(\mathbf{y}|\{\mathbf{s_k}\})p(\{\mathbf{s_k}\}) \tag{4.1}$$

in a generative modeling setting, where $p(\mathbf{y}|\{\mathbf{s_k}\})$ and $p(\{\mathbf{s_k}\})$ are the likelihood and the prior respectively. The computation of $p(\mathbf{y}|\{\mathbf{s}_k\}) = \prod_t p(y_t|\{s_{k,(t)}\})$ is known as acoustic inference, where $p(y_t|\{s_{k,(t)}\})$ was derived in Chapter 3. The computation of $p(\{\mathbf{s_k}\})$, referred to as temporal inference, is the main focus of this chapter.

In the simplest case, the acoustic states within the same source are assumed to be statistically independent. With this assumption, $p(\{\mathbf{s_k}\})$ is given by

$$p(\{\mathbf{s_k}\}) = \prod_t \prod_k p(s_{k,t}) \tag{4.2}$$

where $p(s_{k,t}) = p(s_k)$ is the state prior probability in the acoustic models. The

posterior probability is

$$p(\{s_{k,(t)}\}|\mathbf{y}) = \frac{p(y_t|\{s_{k,(t)}\})\prod_k p(s_{k,t})}{\sum_{\{s_{k,(t)}\}} p(y_t|\{s_{k,(t)}\})\prod_k p(s_{k,t})}\,. \tag{4.3}$$

However, it is poor to assume that the states within the same source are independent. The coming section will explain the reasons.

4.2 The importance of speech dynamics

A speech utterance is composed of continuously produced speech units such as phonemes. The duration of a speech unit typically spans over a number of short-time frames. These frames are overlapped in short-time frame processing. The assumption that the acoustic states are independent is obviously invalid. The state dependency within the same source represents the temporal continuity of a speech signal, which is commonly known as speech dynamics [11].

Single-microphone speech separation is an ill-posed problem. Resolving the ambiguities is a major problem in the separation process. For a specific frequency component, the ambiguity refers to the difficulty in determining the correct amplitude of each individual source. Figure 4.1 illustrates an example with the MIXMAX model. A source cannot be determined uniquely if it is masked by another source.

The speech dynamics is determined by the spoken content. It is unlikely to be affected by the mixing process. The temporal continuity is an important cue to be exploited by many speech separation algorithms [5], including non-negative matrix factorization [37][93] and computational auditory scene analysis [43][94].

With hidden Markov models (HMM) [95], speech dynamics is reflected by the state transition probabilities $p(s_{k,t}|s_{k,t-1})$. A recent study confirms the usefulness of transition probabilities in improving speech recognition performance [96]. For single-microphone speech separation, speech dynamics can be modeled by graphical models such as factorial HMM [74]. The posterior probability $p(\{s_{k,(t)}\}|\mathbf{y})$ are computed accordingly from the graphical structures.

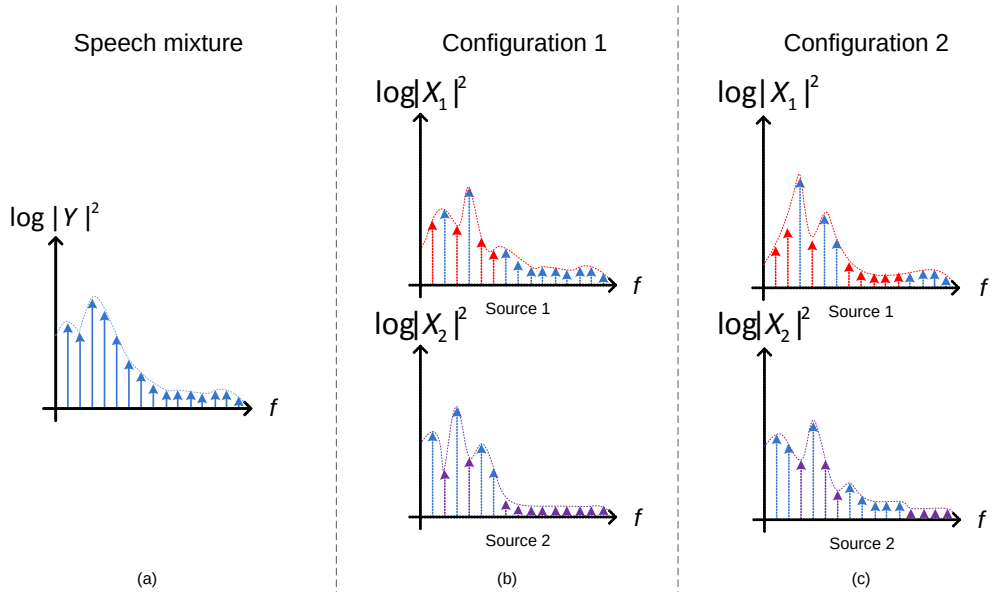


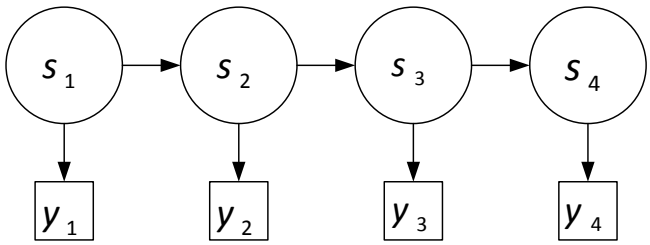
Figure 4.1: An illustration showing the ambiguity of the speech sources given only one speech mixture; a) a speech mixture; b) a possible configuration of the sources, frequency components which are dominated in the speech mixture are in blue, while those are masked by the other source are in red or purple; c) another configuration of the sources which are also resulted in the same speech mixture of (a) according to the MIXMAX model

4.3 Graphical models for speech processing

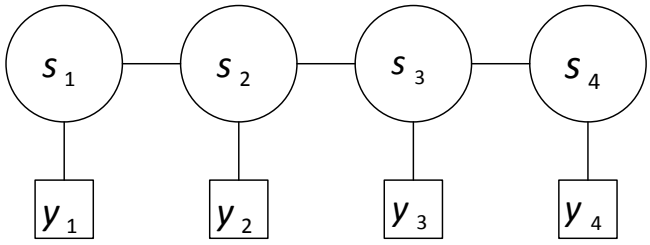
4.3.1 Overview of graphical model

A graph $(\mathcal{V}, \mathcal{E})$ is defined as the collection of a set of M nodes, $\mathcal{V} = \{1, 2, \ldots, M\}$, and a set of edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. An edge connects a pair of nodes $a, b \in \mathcal{V}$. A graphical model is defined by associating each node with a random variable [8]. For a directed graph, each edge is defined with a specific direction. The edge (a, b) is different from (b, a). For undirected graph, (a, b) and (b, a) refer to the same edge. A directed graphical model based on directed acyclic graph (DAG) is also known as a Bayesian network [97].

In a Bayesian network, a directed edge connected from node a to node b represents the conditional probability densities $p(s_b|s_a)$, where s_a and s_b are the random



(a)



(b)

Figure 4.2: (a) A linear-chain directed graphical model which represents the state sequence generated by a hidden Markov model (HMM); (b) A linear-chain undirected graphical model which is referred to as a Markov random field (MRF)

variable on node a and b respectively. An undirected graphical model is also known as a Markov random field (MRF). The edge (a, b) is associated with potential function $\Phi(s_a, s_b)$. Figure 4.2a shows a linear-chain directed graphical model with node variable s_t and observation variables y_t at time t. The model represents a random process generated from a hidden Markov model (HMM), where $\mathbf{s} = (s_1, s_2 \cdots, s_T)$ denotes the hidden state sequence for generating the observation $\mathbf{y} = (y_1, y_2 \cdots, y_T)$. The joint probability corresponding to the linear-chain HMM is given as,

$$p(\mathbf{s}, \mathbf{y}) = \prod_{t=1}^{T} p(y_t|s_t) \times \Big[\prod_{t=1}^{T} p(s_t|s_{t-1}) \Big], \tag{4.4}$$

where $p(s_t|s_{t-1})$ is the transition probability, $p(s_1|s_0) = p(s)$ is the prior probability of the states and $p(y_t|s_t)$ is the emission probability of observation given a state. The corresponding posterior probability $p(\mathbf{s}|\mathbf{y}) = \frac{p(\mathbf{s},\mathbf{y})}{\sum_{\{\mathbf{s}\}} p(\mathbf{s},\mathbf{y})}$ is obtained by the Bayes' rule. A linear-chain Markov random field is illustrated as in Figure 4.2b, with the joint probability

$$p(\mathbf{s}, \mathbf{y}) = \frac{1}{Z} \prod_{t=1}^{T} \Phi(s_t, y_t) \times \Big[\prod_{t=2}^{T} \Phi(s_{t-1}, s_t) \Big], \tag{4.5}$$

where Z is a normalizing factor known as the partition function, $\Phi(y_t, s_t)$ is the joint potential function of the observation and the state and $\Phi(s_t, s_{t-1})$ is the potential function for co-occurrence of the states. One may notice that by setting

$$\begin{aligned} \Phi(s_1, y_1) &= p(y_1|s_1)p(s_1) \\ \Phi(s_t, y_t) &= p(y_t|s_t) \\ \Phi(s_{t-1}, s_t) &= p(s_t|s_{t-1}) \end{aligned}$$

and $Z = 1$, the two graphical models are equivalent.

Given the observation sequence $\mathbf{y}$, the graphical model in Figure 4.2b represents a conditional random field (CRF), which is a special case of Markov random field. Instead of joint probability, the graphical model represents a conditional probability

$$p(\mathbf{s}|\mathbf{y}) = \frac{1}{Z(\mathbf{y})} \prod_{t=1}^{T} \Phi(s_t, y_t) \times \Big[\prod_{t=2}^{T} \Phi(s_{t-1}, s_t) \Big], \tag{4.6}$$

where $Z(\mathbf{y}) = \sum_{\mathbf{s}} \prod_{t=1}^{T} \Phi(s_t, y_t) \times \Big[\prod_{t=2}^{T} \Phi(s_{t-1}, s_t) \Big]$ is the partition function.

4.3.2 Moralization of directed graphical model

A directed graphical model is moral when there is an edge connecting the parent nodes that have the same child node. Moralization is a process of converting a directed graphical model into an undirected one, by adding the edges over or "marrying" the parent nodes and dropping the arrows in an undirected graph representation [98]. It is an essential step to maintain the joint distribution between the directed and the converted undirected graphical models.

Figure 4.3 shows the moralization rules. For the chain and the parent with two children, the joint probability $p(W, X, Y)$ is expressed as

$$\begin{aligned} p(W, X, Y) =& p(Y|X)p(X|W)p(W) \\ =& \frac{1}{Z}\Phi(W, X)\Phi(X, Y), \end{aligned}$$

in which each conditional probability involves at most two objects. Hence a maximal clique (complete subgraph) of two is sufficient to express the join potentials in the undirected graphical model. The term Z is the normalizing term to ensure the sum-to-one property of probability. For the case of two parents with single child as in

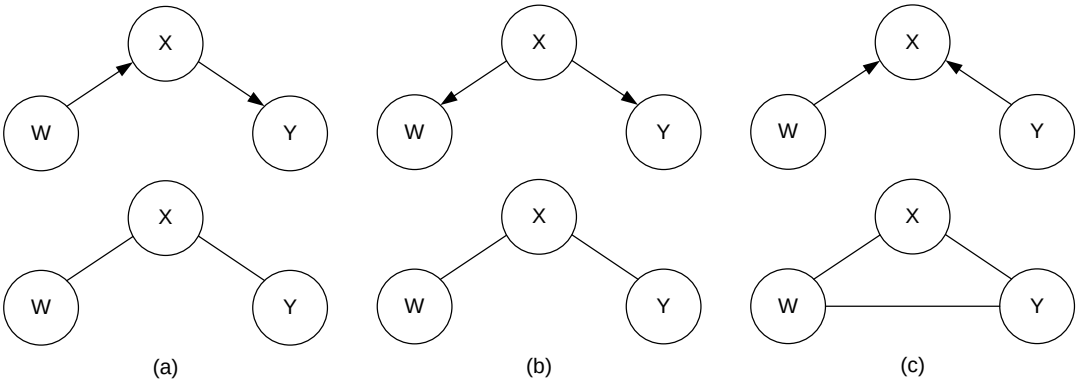


Figure 4.3: The rules for moralization; a) a chain; b) X is the parent with 2 children, W and Y; c) two parents W and Y with single child X

Figure 4.3c, $p(W, X, Y)$ is expressed as

$$\begin{aligned} p(W, X, Y) =& p(X|W, Y)p(W)p(Y) \\ =& \frac{1}{Z}\Phi(W, X, Y) \end{aligned} .$$

The term $p(X|W, Y)$ involves three objects. A maximal clique of three is required in the undirected graphical model. An edge is added to fulfill this requirement.

Figure 4.2b is a moral graph of Figure 4.2a. There exists some parameters for these two graphical models to produce the same joint probability $p(\mathbf{s}, \mathbf{y})$. Note that if the arrows are ignored, their graphical structures are the same. It is a special characteristic of tree-structured graphical models.

4.3.3 Representing speech with graphical model

A speech signal can be represented by a graphical model. Let y_t be the observation vector at frame t, e.g., the short-time log-spectrum. A speech unit, which may be a phoneme, a syllable or a word, is modeled by an HMM models. Each state is associated with an emission probability distribution. The transition from the current state to the next state is described by a probability. Figure 4.4 illustrates how a speech unit is modeled by an hidden Markov model (HMM) represented by a directed graphical model. The directed graphical model is interpreted as follows. At time t, the observation y_t is generated by state i with emission probability $p(y_t|s_t = i)$. State i of node S_t at time t can make transition to state j of node S_{t+1} at time $t + 1$ with

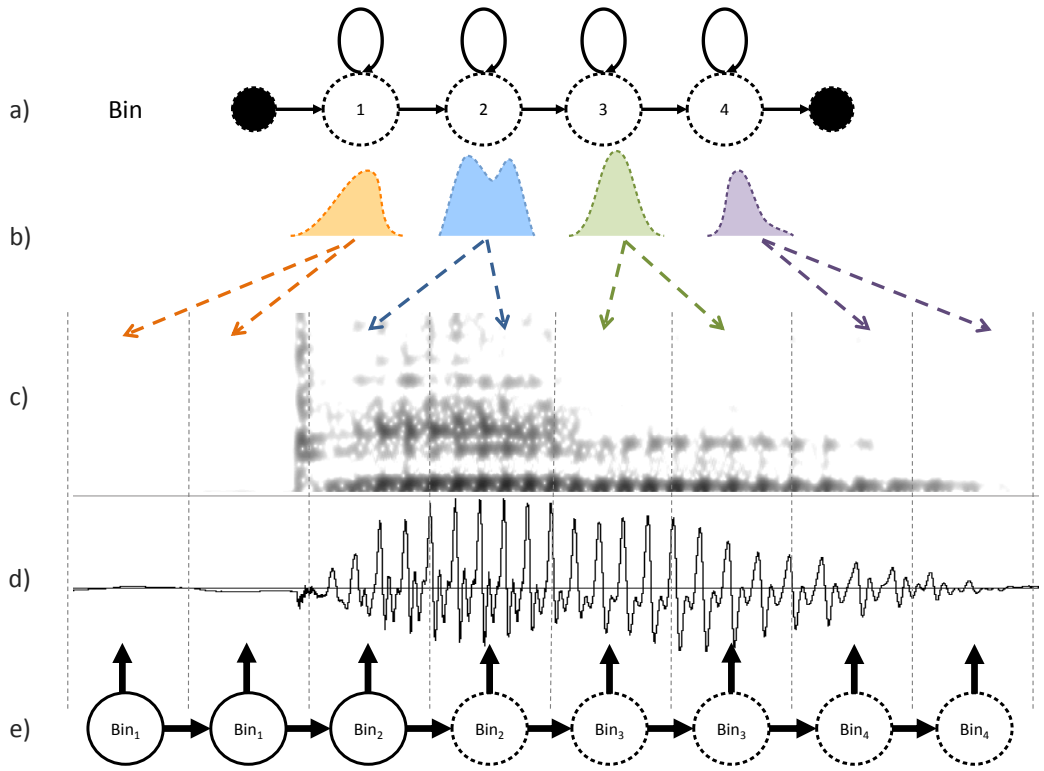


Figure 4.4: An illustration of applying graphical model on speech processing (not in scale); a) a 4-state HMM acoustic model for the word "Bin"; b) emission probabilities of the spectral observations of c) the spectrogram of the speech signal; d) the time-domain waveform of the speech signal; e) a linear-chain graphical model representing the acoustic state sequence which generates the speech signal

transition probability $p(s_{t+1} = j|s_t = i)$.

Undirected graphical model (Figure 4.2b) is also applied to speech processing. For example, conditional random fields (CRF) were used for speech recognition [15][17][18]. The conditional probabilities of the label sequences given the observations are obtained from the joint potential in association with the states s_t and the observation y_t of the node S_t, and the co-occurrence of the state pair (s_t, s_{t+1}).

4.4 Modeling speech dynamics

Speech sources are represented by Markov chains generated by the corresponding acoustic models. The speech sources are interacted to generate the observed mixture

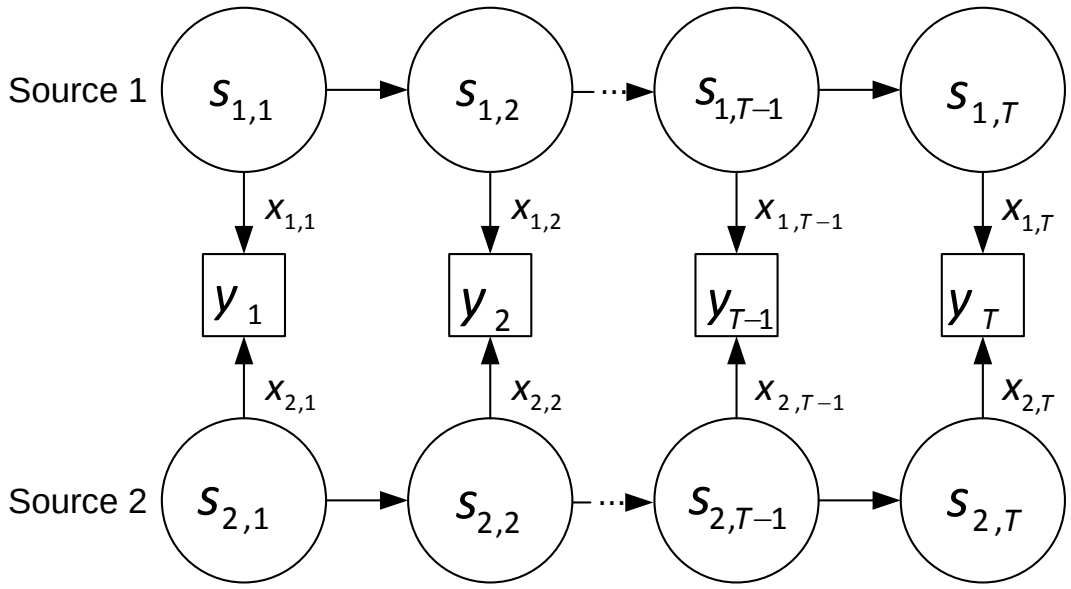


Figure 4.5: A factorial HMM for single-microphone speech separation with two speech sources.

via a mixing process. Since the mixing process does not affect the speech dynamics, the sources are independent to each other. The temporal inference is expressed as,

$$p(\{\mathbf{s_k}\}) = \prod_t \prod_k p(s_{k,t}|s_{k,t-1}) \,, \tag{4.7}$$

where $p(s_{k,1}|s_{k,0}) = p(s_k)$ is the prior probability of the corresponding acoustic state. The joint probability of source state sequences and the mixture observation is

$$p(\{\mathbf{s_k}\}, \mathbf{y}) = \prod_{t=1}^{T} p(y_t|\{s_{k,(t)}\}) \times \left[\prod_{t=1}^{T} \prod_{k=1}^{K} p(s_{k,t}|s_{k,t-1}) \right], \tag{4.8}$$

which can be represented by a directed graphical model referred to as factorial HMM [74] (Figure 4.5). Factorial HMM is a promising approach to single-microphone speech separation [9][99][10][58]. In the Speech Separation Challenge [5], factorial HMM system achieved the best target word recognition results in the mixtures [10]. The accuracies were even better than human recognition accuracy.

The conditional probability $p(\{\mathbf{s_k}\}|\mathbf{y})$ is computed by Bayes' rule. The frame-level posterior probability $p(\{s_{k,(t)}\}|\mathbf{y})$ is obtained by marginalizing the posterior probabilities $p(\{\mathbf{s_k}\}|\mathbf{y})$ with dynamic programming known as the message-passing algorithm. Forward-backward algorithms on HMM is an example. It is tempting to group a pair of state variables $\{s_{k,(t)}\}$ into a single state variable s_t to form a

linear-chain HMM and compute the posterior probability $p(s_t|\mathbf{y})$ as in [3]. However, scalability is the main concern. The exponential growth of the number of grouped states imposes a challenge in statistical inference. With K sources and S states per source, the number of the states for linear-chain HMM is S^K. The number of transitions is as large as $S^K \times S^K$. The complexity of forward-backward procedure would become $O(TS^{2K})$, which is the same as a naive message-passing algorithm on factorial HMM. This may not be a major problem if the interfering signal is noise, because a few acoustic states are sufficient to describe the noise statistics, as in [3]. The problem becomes critical in speech separation, in which highly variable content of speech requires a large number of acoustic states.

4.5 Exact inference of factorial HMM

Exact inference can be performed with dynamic programing when a generalized graphical model is represented in a tree-structured graph. Junction tree algorithm is exactly the algorithm based on this idea [100][97]. The mapping, which is referred to as tree decomposition, can be performed if the graph is chordal [101]. A graph is chordal if every cycle of four nodes or more has a chord. A chord is an edge connected to the nodes that are not adjacent to each other. In junction tree algorithm, the procedure to convert a graph into a chordal graph is referred to as triangulation. The result of tree decomposition is referred to as a junction tree. A junction tree is a maximum spanning clique tree which satisfies the junction tree property, which states [102]:

> "For each pair $\mathcal{U}, \mathcal{V}$ of cliques with intersection S, all cliques on the path between $\mathcal{U}$ and $\mathcal{V}$ contains S."

The forward messages $\alpha_t = p(s_{1,t}, s_{2,t}, \mathbf{y_{1\ldots t}})$, $\alpha_t^\star = p(s_{1,t}, s_{2,t-1}, \mathbf{y_{1\ldots t-1}})$, and the backward messages $\beta_t = p(\mathbf{y_{t+1\ldots T}}|s_{1,t}, s_{2,t})$, $\beta_t^\star = p(\mathbf{y_{t+1\ldots T}}|s_{2,t}, s_{1,t+1})$ are derived from the junction tree updates [74]. The detailed derivation of junction tree algorithm is given in Appendix B. The complexity of junction tree algorithm is exponential to the size of the largest clique. For single-microphone speech separation

of K sources, the complexity is $\mathcal{O}(TKS^{K+1})$ [10][74]. For a two-source case, the update rules for forward messages and backward messages are as follows:

- Forward messages:

$$\alpha_t^\star = \sum_{s_{1,t-1}} p(s_{1,t}|s_{1,t-1})\alpha_{t-1} \tag{4.9}$$

$$\alpha_t = p(y_t|s_{1,t}, s_{2,t}) \sum_{s_{2,t-1}} p(s_{2,t}|s_{2,t-1})\alpha_t^\star \tag{4.10}$$

- Backward messages:

$$\beta_t^\star = \sum_{s_{2,t+1}} p(y_{t+1}|s_{1,t+1}, s_{2,t+1})p(s_{2,t+1}|s_{2,t})\beta_{t+1} \tag{4.11}$$

$$\beta_t = \sum_{s_{1,t+1}} p(s_{1,t+1}|s_{1,t})\beta_t^\star \,. \tag{4.12}$$

Once the forward message and the backward message are obtained, $p(s_{1,t}, s_{2,t}|\mathbf{y})$ is computed as

$$p(s_{1,t}, s_{2,t}|\mathbf{y}) = \frac{\alpha_t\beta_t}{\sum_{\{s_{1,t},s_{2,t}\}} \alpha_t\beta_t} \,. \tag{4.13}$$

Empirically, also observed in [11], $p(s_{1,t}, s_{2,t}|\mathbf{y_{1\ldots t-1}}, \mathbf{y_{t+1\ldots T}})$ seems to achieve a slightly better objective quality than $p(s_{1,t}, s_{2,t}|\mathbf{y})$ on reconstructed sources, and similar speech recognition accuracy. A possible reason is the duplicated information of y_t in y_{t-1} and y_{t+1} due to frame overlapping. The current state is predicted from the previous state and the next state, which prevents the abrupt change of reconstructed source spectrum. The continuity is hence further improved. The heuristic posterior probability $p(s_{1,t}, s_{2,t}|\mathbf{y_1^{t-1}}, \mathbf{y_{t+1}^T})$ is expressed as,

$$p(s_{1,t}, s_{2,t}|\mathbf{y_1^{t-1}}, \mathbf{y_{t+1}^T}) = \frac{\hat{\alpha}_t\beta_t}{\sum_{\{s_{1,t},s_{2,t}\}} \hat{\alpha}_t\beta_t} \,, \tag{4.14}$$

where $\hat{\alpha}_t = \sum_{s_{2,t-1}} p(s_{2,t}|s_{2,t-1})\alpha_t^\star$.

4.6 Approximated statistical inference

Approximated inference algorithms are preferred for non-tree-structured graphical models. Exact computation of posterior probabilities requires a complexity that is exponential to the size of the largest clique of a junction tree. For factorial HMM, the

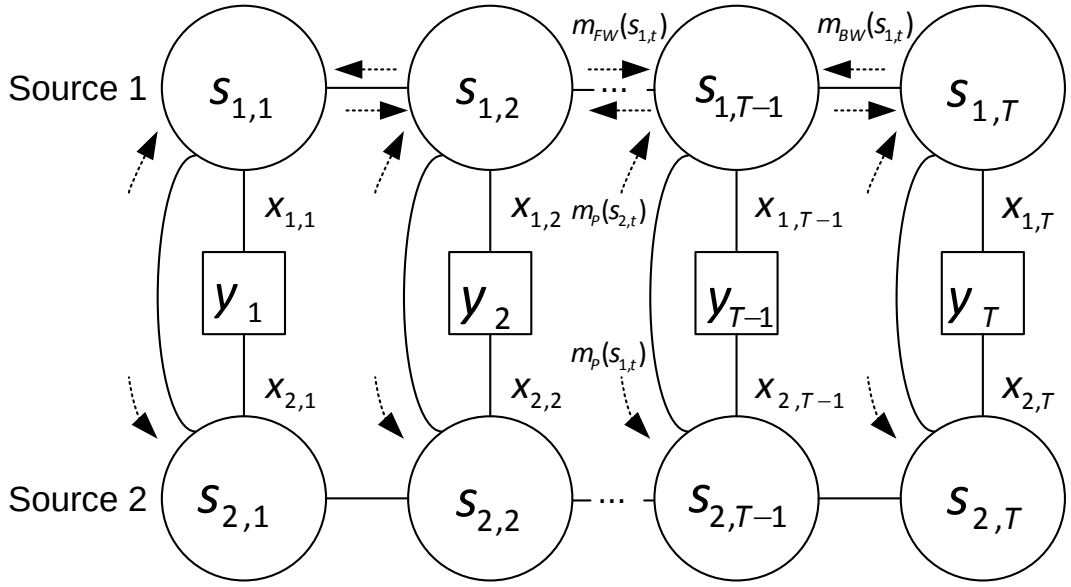


Figure 4.6: The message-passing paths for factorial HMM

size of the largest clique is $K+1$. When the graphical model becomes more complex, e.g., with increased number of sources, exact inference would become intractable. In fact, exact solution is not always necessary, and an approximated solution can be sufficiently accurate in many cases.

4.6.1 Loopy belief propagation

Loopy belief propagation (LBP) is able to reduce the complexity of temporal inference of factorial HMM to quadratic time [103]. Loopy belief propagation ignores the potential loops in a generalized graphical model and performs message-passing as in a tree-structured graphical model. LBP attains satisfactory results in many graphical modeling problems [103], including single-microphone speech separation [99].

For a directed graphical model, LBP is performed on the corresponding moral graph. The messages $m(\cdot)$ and the frame-level posterior probability $p(\{s_{k,(t)}\}|\mathbf{y})$ of factorial HMM are defined according to the moral graph as shown in Figure 4.6,

$$\begin{aligned} m_{FW}(s_{k,t}) &\leftarrow \kappa \sum_{s_{k,t-1}} p(s_{k,t}|s_{k,t-1}) m_{FW}(s_{k,t-1}) m_P(s_{k,t-1}) \\ m_{BW}(s_{k,t}) &\leftarrow \kappa' \sum_{s_{k,t+1}} p(s_{k,t+1}|s_{k,t}) m_{BW}(s_{k,t+1}) m_P(s_{k,t+1}) \\ m_P(s_{k,t}) &\leftarrow \kappa'' \sum_{\{s_{j,(t)\backslash j=k}\}} p(y_t|\{s_{k,(t)}\}) \prod_j m_{FW}(s_{j,t}) m_{BW}(s_{j,t}) \end{aligned} \tag{4.15}$$

$$p(\{s_{k,(t)}\}|\mathbf{y}) \approx \kappa''' p(y_t|\{s_{k,(t)}\}) \prod_k m_{FW}(s_{k,t}) m_{BW}(s_{k,t}) \,. \tag{4.16}$$

where $\kappa, \kappa', \kappa'', \kappa'''$ are normalization constants. The heuristic posterior probability is approximated as $p(s_{1,t}, s_{2,t}|\mathbf{y_{1\ldots t-1}}, \mathbf{y_{t+1\ldots T}}) \approx \kappa''' \prod_k m_{FW}(s_{k,t}) m_{BW}(s_{k,t})$.

A message-passing algorithm may not compute the exact solution due to the potential loops in the graphical model after moralization [104]. Moreover, there is no guarantee on the convergence of the solution with LBP, suggesting that the algorithm may fail in some occasions.

4.6.2 Structured mean field method

From graphical modeling point of view, the difficulty of inferring factorial HMM is due to the coupling of the Markov chains by $p(y_t|\{s_{k,(t)}\})$ from the mixing process. As discussed earlier, the speech dynamics of a source is unlikely to be affected by the mixing process. In designing an approximated inference algorithm, we aim to to keep the internal structure of each source. As shown in Figure 4.7, we can decouple the chains by replacing the posterior probability distribution $p(\{\mathbf{s_k}\}|\mathbf{y})$ with a tractable distribution $Q(\{\mathbf{s_k}\}) = \prod_k Q(\mathbf{s_k})$. This tractable distribution should be as close as the original $p(\{\mathbf{s_k}\}|\mathbf{y})$. The similarity between the distributions can be measured in terms of Kullback-Leibler (KL) divergence $\mathcal{D}$, as it is the difference between $\log p(\mathbf{y})$ and its lower bound $\mathcal{L}$ by Jensen's inequality, i.e. $\mathcal{D}(Q(\{\mathbf{s_k}\})||p(\{\mathbf{s_k}\}|\mathbf{y})) = \log p(\mathbf{y}) - \mathcal{L}$ [105],

$$\begin{aligned} \log p(\mathbf{y}) &= \log \sum_{\{\mathbf{s_k}\}} Q(\{\mathbf{s_k}\}) \frac{p(\{\mathbf{s_k}\}, \mathbf{y})}{Q(\{\mathbf{s_k}\})} \\ &\geq \sum_{\{\mathbf{s_k}\}} Q(\{\mathbf{s_k}\}) \log \frac{p(\{\mathbf{s_k}\}, \mathbf{y})}{Q(\{\mathbf{s_k}\})} \quad = \mathcal{L} \,. \end{aligned} \tag{4.17}$$

Maximizing $\mathcal{L}$ is equivalent to minimizing the KL divergence $\mathcal{D}$. By differentiating $\mathcal{L}$ with respect to $Q(\mathbf{s_k})$ and evaluating from its zero-gradient point, an expression of $Q(\mathbf{s_k})$ is obtained. Note that the solution is only locally optimal. The mean field approximation is in general a non-convex problem,

$$\begin{aligned} Q(\mathbf{s_k}) \propto \prod_{t=1}^{T} \Bigg[\sum_{\{s_{j\setminus j=k}\}} \prod_j p(s_j) p(y_t|\{s_{k,(t)}\}) \Bigg] \\ \times \prod_{t=1}^{T} p(s_{k,t}|s_{k,t-1}) \,. \end{aligned} \tag{4.18}$$

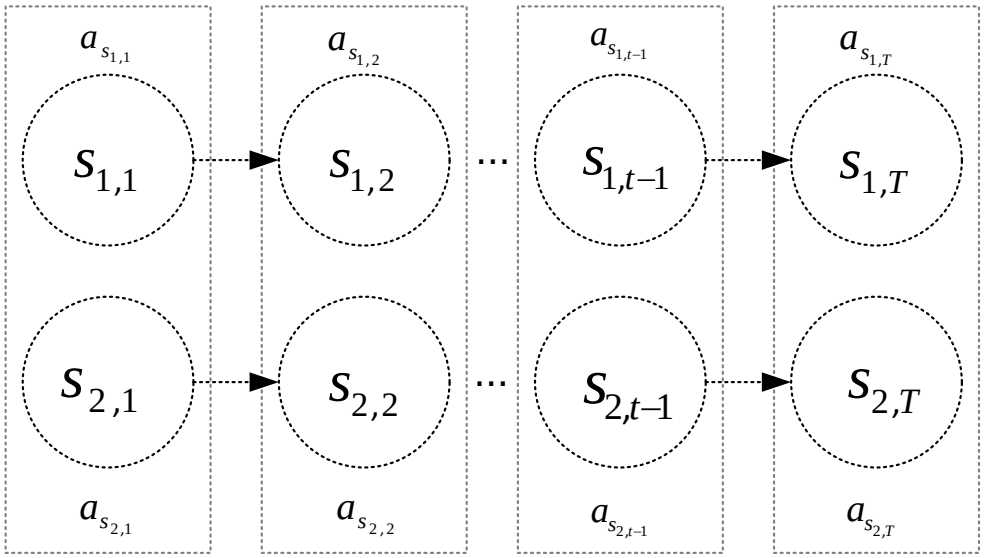


Figure 4.7: In structured mean field method, the Markov chains in a factorial HMM are decoupled with the variation parameters $a_{s_{k,t}} = \tilde{p}(y_t|s_{k,t})$ resided on the node to minimize the distance between the approximated distribution and the original distribution represented by the factorial HMM.

Denoting $\tilde{p}(y_t|s_{k,t}) = \sum_{\{s_{j\setminus j=k}\}} \prod_j p(s_j)p(y_t|\{s_{k,(t)}\})$ as the variational parameter, $Q(\mathbf{s_k})$ ensembles the conditional probability computed from a linear-chain HMM. This algorithm is referred to as structured mean field method in the literature [8], as $\tilde{p}(y_t|s_{k,t}) = \mathbb{E}_{p(\prod_{s_{j\setminus j=k}} p(s_j))} p(y_t|\{s_{k,(t)}\})$. The variational parameter $\tilde{p}(y_t|s_{k,t})$ can be considered as a mixture model, with $\prod_{j\setminus j=k} p(s_j)$ as the component weight and $p(y_t|\{s_{k,(t)}\})$ as a mixture component. Specifically when $p(y_t|\{s_{k,(t)}\})$ is modeled with multivariate Gaussian distribution, $\tilde{p}(y_t|s_{k,t})$ is a Gaussian mixture model which is widely used in automatic speech recognition. Forward-backward algorithm can be applied to infer individual chains efficiently with quadratic complexity for the frame-level marginal probabilities $Q(s_{k,t})$. Note that for speech application, $p(s_j)$ can be substituted with the prior probability of acoustic states. As there is no parameter dependency across different Markov chains, a single iteration is sufficient for $Q(s_{k,t})$, which is the major distinction from loopy belief propagation. The outline of this algorithm is given in Algorithm 4.1.

Algorithm 4.1 Structured mean field method for single-microphone speech separation with factorial HMM

function FHMM_SMF($\mathbf{y}$,$\{\mathbf{AM_k}\}$)

for $t = 1 : T$ **do**

Compute $p(y_t|\{s_{k,(t)}\}), \forall\{s_k\} = \{s_1 \dots s_K\}$.

end for

for $t = 1 : T$ **do**

for $k = 1 : K$ **do**

Compute $\tilde{p}(y_t|s_{k,t}) = \mathbb{E}_{p(\prod_{j\setminus j=k} p(s_j))} p(y_t|\{s_{k,(t)}\})$.

Perform forward-backward on $Q(\mathbf{s_k})$ for $\hat{p}(s_{k,t})$.

end for

end for

for $t = 1 : T$ **do**

$p(\{s_{k,(t)}\}|y_t) = \prod_k Q(s_{k,t}), \forall\{s_k\} = \{s_1 \dots s_K\}$

end for

return $p(\{s_{k,t}\}|y_t), \forall\{s_k\} = \{s_1 \dots s_K\}, t$

end function

While the complexity of temporal inference is reduced to quadratic in the structured mean field method, the complexity of acoustic inference with the MIXMAX model, i.e. the complexity of computing $p(y_t|\{s_{k,(t)}\})$, is still exponential to the number of sources. The acoustic inference would take up considerable run time during the separation process, but the overall complexity has reduced from $\mathcal{O}(S^{K+1})$ to $\mathcal{O}(S^K)$ with the approximated inference algorithms. Parallelizing the computation of $p(y_t|\{s_{k,(t)}\})$ helps to speed up the the exact acoustic inference. The structured mean field method further promotes parallelization of the inference algorithm of factorial HMM. As the Markov chains are decoupled, forward-backward algorithm can be performed independently for each chain. The process can be parallelized straightforwardly. If limited computational resource is available, a further approximation to acoustic inference is possible. A linear-time algorithm is developed to approximate the acoustic inference in [84]. However, speech separation performance is sacrificed.

4.7 Experiments

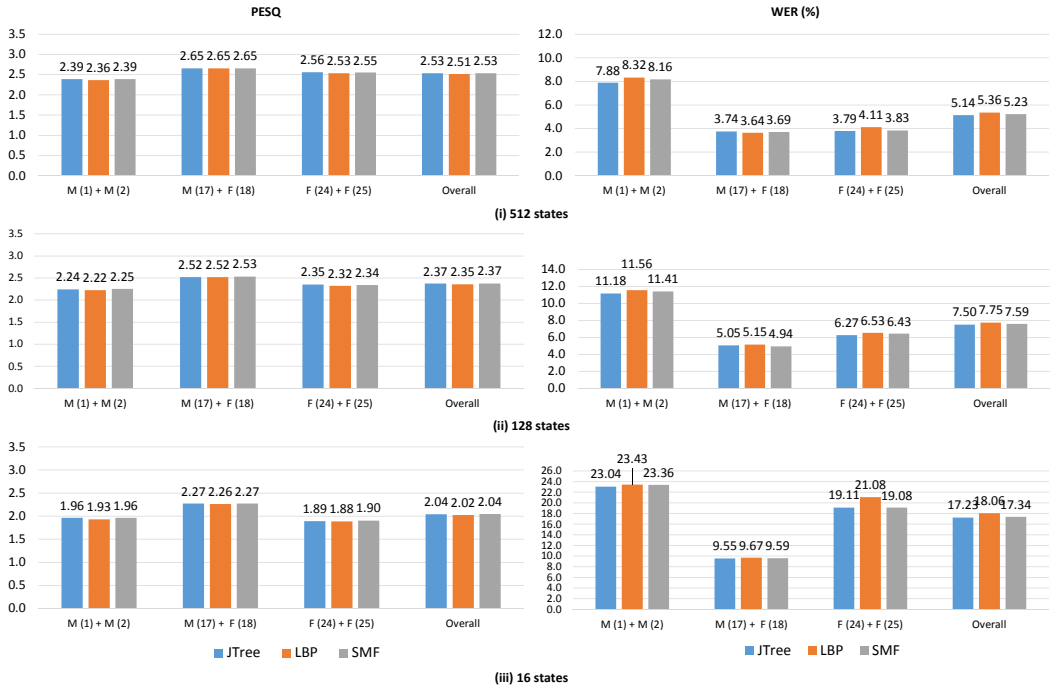
Speech separation experiments with factorial HMM are performed. The purpose of the experiments is to define a factorial HMM baseline for the later speech separation development. The signal-to-signal ratio of the mixtures is set to 0 dB. The log-spectra of the speech sources are estimated with MMSE. The waveforms are generated using the phase spectrum of the mixture by the overlap-add method. The experiments are performed with speaker-dependent acoustic models of 16, 128 and 512 states.

The experiments are performed with both the exact and approximated inference algorithms. Three inference algorithms: junction tree algorithm (JTREE) [97], loopy belief prorogation (LBP) [99] and structured mean field method (SMF) are investigated. We also compare the performance of using GMM modeling approach and the MIXMAX model for modeling the state-level interaction. For GMM modeling approach, two sets of training data, namely ENTIREDATA with all of the 200k training mixtures, and DATA_1% in which only 2k mixtures (about 1% of the entire set) are prepared for model training. For the MIXMAX model, we further include the speech separation results without speech dynamics. The results help to verify the importance of speech dynamics in speech separation.

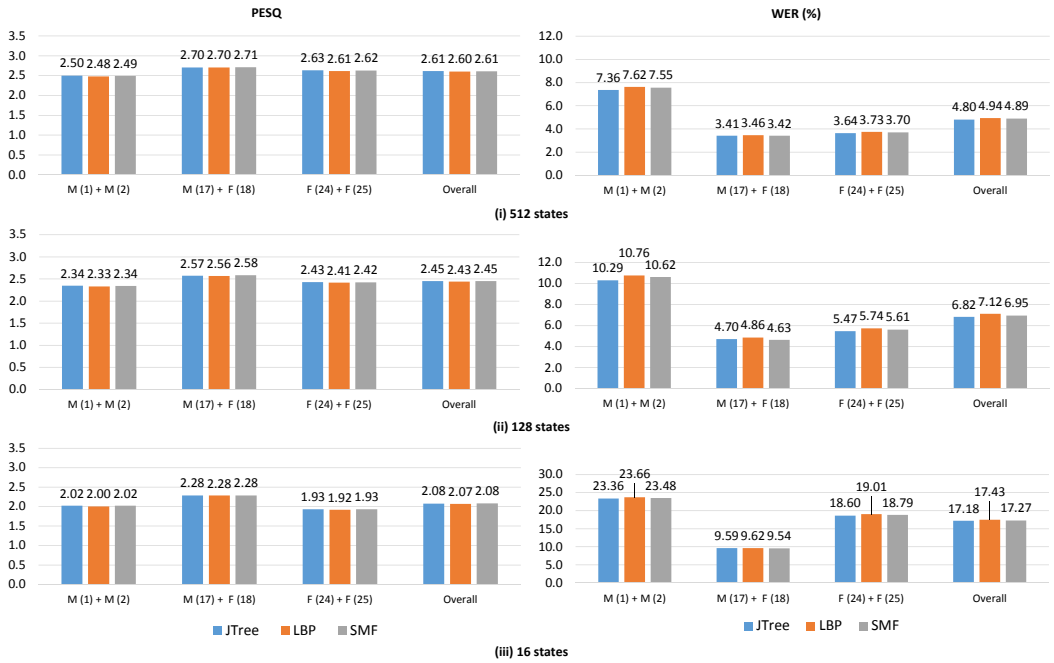
The key results in terms of PESQ and WER for the MIXMAX model and ENTIREDATA with different inference algorithms are shown in Figure 4.8. The results are averaged over 2500 trials with different speech mixtures of two speakers.

4.7.1 Comparison the inference methods

The separation performance of different inference algorithms are on the same trend across the MIXMAX model and the ENTIREDATA. The structured mean field (SMF) achieves the similar PESQ and WER as the junction tree algorithm (JTREE). SMF also achieves the similar performance as LBP. For ENTIREDATA, all three inference algorithms achieve similar PESQ. In terms of recognition accuracy, it is natural that junction tree algorithm achieves the lowest word error rate (WER) since it is an exact algorithm. However, the relative difference of WER between the junction tree algorithm and the structured mean field method is only 4% in the worst case, and on



(a) with the MIXMAX model



(b) with ENTIREDATA

Figure 4.8: PESQ and WER (%) of the three inference algorithms

Table 4.1: The complexity of the temporal and acoustic inference and the averaged runtime of the three algorithms with 512 acoustic states, where T, K, S are the number of frames, sources and states respectively.

Algorithm	Temporal	Acoustic	Runtime (s)
JTree	$\mathcal{O}(TKS^{K+1})$	$\mathcal{O}(TS^K)$	489.54
LBP	$\mathcal{O}(TKS^2)$	$\mathcal{O}(TS^K)$	137.42
SMF	$\mathcal{O}(TKS^2)$	$\mathcal{O}(TS^K)$	136.57

average it is less than 2%. The separation results of SMF is slightly better than LBP in terms of WER, although the relative improvement is small (overall less than 4% in WER in both settings). We conclude that the choice of inference algorithms is not a significant factor on the separation performance of factorial HMM.

The complexity of the algorithms and the average runtime for separating one speech mixture in the case of 512 acoustic states are shown in Table 4.1. The runtime is measured on a Linux machine, running on an Intel Core 2 Duo 3 GHz processor, with single CPU core allocated to the program. The runtime of SMF is only one-third of JTREE, and about the same as LBP. Considering the saving of the runtime and the small trade-off in accuracy, the results suggest that an exact inference algorithm is not necessary for a practical speech separation system. In the following contents, unless otherwise specified, SMF method is adopted for approximated inference of factorial HMM.

4.7.2 The choice of state-level interaction models

Figure 4.8 also reveals that when there are sufficient amount of training data, the empirical distribution from ENTIREDATA achieves slightly better overall performance than the MIXMAX model. The absolute improvement of PESQ is small ($<$0.1), but there is WER reduction (8.4% and 6.6% relatively with SMF) for 128 and 512 states respectively. For the case with 16 acoustic states, the relative WER reduction is less significant (3.5% maximum). This may indicate the performance saturation of factorial HMM. Nevertheless, the overall separation results of the MIXMAX model

Table 4.2: Speech separation results in terms of PESQ and WER (%) with the MIXMAX model, ENTIREDATA and DATA_1%

S		M (1) + M (2)		M (17) + F (18)		F (24) + F (25)		Overall	
		PESQ	WER	PESQ	WER	PESQ	WER	PESQ	WER
512	**MIXMAX**	2.39	8.16	2.65	3.69	2.55	3.83	2.53	5.23
	Data_1%	2.01	19.03	2.42	6.83	2.26	9.42	2.23	11.76
	EntireData	**2.49**	**7.55**	**2.71**	**3.42**	**2.62**	**3.70**	**2.61**	**4.89**
128	**MIXMAX**	2.25	11.41	2.53	4.94	2.34	6.43	2.37	7.59
	Data_1%	2.30	11.91	2.54	5.24	2.37	6.15	2.41	7.77
	EntireData	**2.34**	**10.62**	**2.58**	**4.63**	**2.42**	**5.61**	**2.45**	**6.95**
16	**MIXMAX**	1.96	**23.36**	2.27	9.59	1.90	19.08	2.04	17.34
	Data_1%	**2.02**	23.58	**2.29**	**9.47**	**1.93**	18.96	**2.08**	17.34
	EntireData	**2.02**	23.48	2.28	9.54	**1.93**	**18.79**	**2.08**	**17.27**

Table 4.3: The separation results of different state-level interaction models in terms of BSS_EVAL (SDR, SAR, SIR)

		M (1) + M (2)			M (17) + F (18)			F (24)+ F (25)			Overall		
		SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR
512	**MIXMAX**	6.17	9.63	**9.54**	9.75	12.13	14.26	8.93	11.79	12.66	8.28	11.19	**12.15**
	Data_1%	4.13	8.76	6.76	8.69	11.00	13.36	7.70	10.57	11.48	6.84	10.11	10.53
	EntireData	**6.36**	**10.39**	9.14	**9.96**	**12.40**	**14.34**	**9.21**	**12.17**	**12.79**	**8.51**	**11.65**	12.09
128	**MIXMAX**	5.62	9.09	**9.02**	9.03	11.32	13.92	8.03	10.75	12.02	7.56	10.39	**11.65**
	Data_1%	5.55	9.57	8.44	9.04	11.35	13.86	8.18	10.95	12.06	7.59	10.62	11.45
	EntireData	**5.79**	**9.83**	8.63	**9.26**	**11.59**	**14.03**	**8.34**	**11.15**	**12.18**	**7.80**	**10.86**	11.62
16	**MIXMAX**	**4.26**	8.27	**7.34**	7.78	10.09	**12.85**	5.73	8.69	**9.69**	5.92	9.01	**9.96**
	Data_1%	4.25	**8.83**	6.88	**7.92**	**10.32**	12.73	**5.85**	**8.92**	9.64	**6.01**	**9.35**	9.75
	EntireData	3.99	8.47	6.71	7.72	10.19	12.50	5.48	8.57	9.28	5.73	9.08	9.50

are still comparable to those of the GMM modeling approach. There are fewer requirements on training data for the MIXMAX model. It only requires source training data for the acoustic model, which is also the prerequisite of the GMM modeling approach. The slightly lower performance of the MIXMAX model can be considered as the trade-off for the convenience of model preparation.

When there are insufficient training data, the GMM modeling approach performs poorly. The separation results of DATA_1% are shown in Table 4.2. The case with 512 acoustic states is completely disastrous, with double of WER compared with the MIXMAX model. For the case with 128 acoustic states, the overall WER is slightly worsen (-2.3%). For the case with 16 acoustic states, the results are nearly the same as ENTIREDATA. This confirms that the GMM modeling approach requires

Table 4.4: PESQ and WER (%) of the reconstructed speech sources with or without speech dynamics during speech separation. Factorial HMM is denoted as FHMM. The setting without speech dynamics is denoted as PRIOR

		M (1) + M (2)		M (17) + F (18)		F (24) + F (25)		Overall	
		PESQ	**WER**	**PESQ**	**WER**	**PESQ**	**WER**	**PESQ**	**WER**
512	**Prior**	2.10	10.30	2.57	4.03	2.33	5.33	2.33	6.55
	FHMM	**2.39**	**8.16**	**2.65**	**3.69**	**2.55**	**3.83**	**2.53**	**5.23**
128	**Prior**	1.97	13.45	2.42	5.44	2.15	8.31	2.18	9.07
	FHMM	**2.25**	**11.41**	**2.53**	**4.94**	**2.34**	**6.43**	**2.37**	**7.59**
16	**Prior**	1.70	26.90	2.14	11.46	1.72	21.76	1.85	20.04
	FHMM	**1.96**	**23.36**	**2.27**	**9.59**	**1.90**	**19.08**	**2.04**	**17.34**

a significant amount of training data with the increased number of acoustic states.

By further analyzing the separation results with the BSS_EVAL metrics (Table 4.3), the higher SAR of ENTIREDATA reveals that the improvement is mainly due to the reduction of artifact in source reconstruction. Although the MIXMAX model performs better in suppressing the inferencing components with higher SIR, it is not enough to compensate the negative effects of more reconstruction artifacts, as indicated by the lower SDR.

4.7.3 Speech separation with or without speech dynamics

We also include the separation results without using speech dynamics in Table 4.4. The results clearly show that applying speech dynamics significantly and consistently improves speech separation performance in terms of PESQ and WER. Further investigation on the results of individual speaker pairs supports the claim that applying speech dynamics helps to resolve the ambiguity due to the close pitch ranges of the speakers. The *Male + Male* set has the closest pitch range between two speakers. The improvement in terms of PESQ and WER is the largest, followed by the *Female + Female* set. The improvement of *Male + Female* set, in which their pitch ranges are generally non-overlapped, is the smallest.

The results in terms of BSS_EVAL metrics (SDR, SAR, SIR) are shown in Table 4.5. The metrics reveal that applying speech dynamics helps to reduce the reconstruction artifacts and improve suppression of the frequency components of the inter-

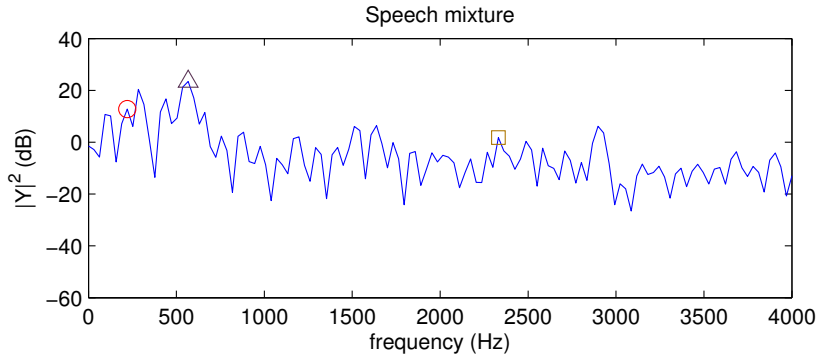
Table 4.5: BSS_EVAL metrics (SDR, SAR, SIR) of the reconstructed speech sources with (FHMM) or without (PRIOR) speech dynamics

		M (1) + M (2)			M (17) + F (18)			F (24)+ F (25)			Overall		
		SDR	**SAR**	**SIR**	**SDR**	**SAR**	**SIR**	**SDR**	**SAR**	**SIR**	**SDR**	**SAR**	**SIR**
512	**Prior**	5.31	8.70	8.80	9.50	11.98	13.87	8.14	11.04	11.86	7.65	10.57	11.51
	FHMM	**6.17**	**9.63**	**9.54**	**9.75**	**12.13**	**14.26**	**8.93**	**11.79**	**12.66**	**8.28**	**11.19**	**12.15**
128	**Prior**	4.86	8.25	8.38	8.73	11.12	13.47	7.27	10.04	11.30	6.95	9.80	11.05
	FHMM	**5.62**	**9.09**	**9.02**	**9.03**	**11.32**	**13.92**	**8.03**	**10.75**	**12.02**	**7.56**	**10.39**	**11.65**
16	**Prior**	3.51	7.43	6.74	7.37	9.77	12.35	4.87	8.29	8.42	5.25	8.49	9.17
	FHMM	**4.26**	**8.27**	**7.34**	**7.78**	**10.09**	**12.85**	**5.73**	**8.69**	**9.69**	**5.92**	**9.01**	**9.96**

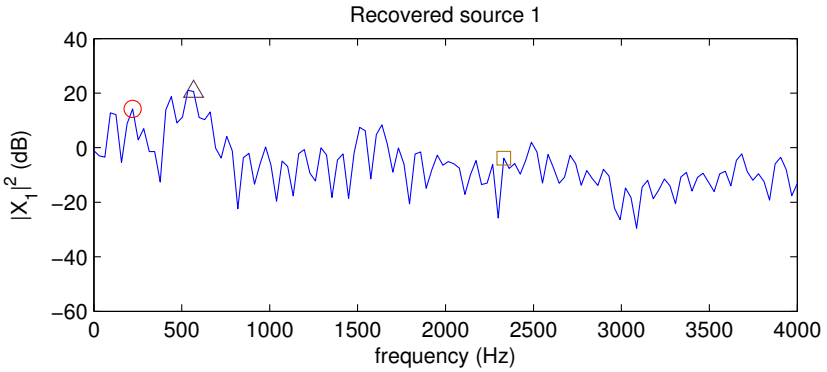
fering sources. The metrics further confirm the usefulness of speech dynamics on the speaker set with close pitch ranges. For *Male + Male* set with 512 acoustic states, SAR has improved by nearly 1 dB when speech dynamics is applied. In contrast, SAR of *Male + Female* set is only slightly improved with speech dynamics.

4.7.4 Samples of reconstructed speech frames

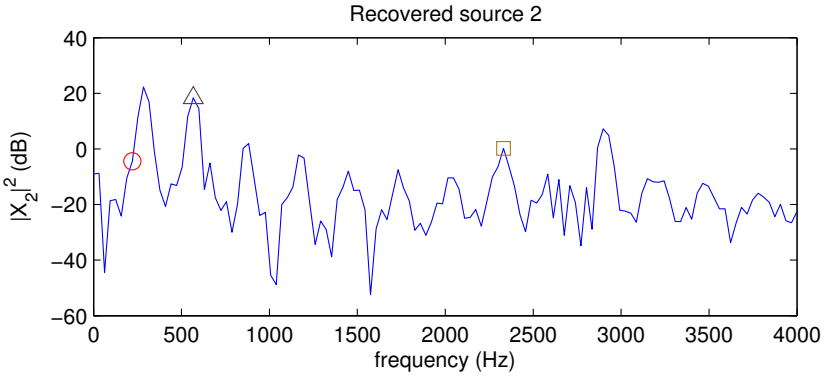
Figure 4.9 shows the speech sources recovered from the mixture in Figure 3.6c. The harmonicity of the sources are generally recovered. However, some distortions are still observed. The harmonic components at 2500-3000 Hz in the female source are lost. At some frequency components, the interfering sources are not attenuated sufficiently, such as the points marked with red circle for female speaker, and the points marked with brown square for male speaker.



(a) Speech mixture



(b) Recovered source 1 (Male)



(c) Recovered source 2 (Female)

Figure 4.9: Speech frames in spectral domain of the reconstructed (a) male sources (b) female sources by factorial HMM with the MIXMAX model

Chapter 5

Speech separation with conditional random fields

5.1 Direct modeling and conditional random fields

5.1.1 Direct modeling for speech separation

In the previous chapter, the posterior probability $p(\{\mathbf{s_k}\}|\mathbf{y})$ is obtained from the joint probability $p(\{\mathbf{s_k}\}, \mathbf{y})$ by a generative approach. Since the observation is always available, it is more straightforward to model $p(\{\mathbf{s_k}\}|\mathbf{y})$ directly from training data. Direct modeling is an inherently discriminative model. It estimates the posterior probability $p(\{\mathbf{s_k}\}|\mathbf{y})$ without involving the joint probability $p(\{\mathbf{s_k}\}, \mathbf{y})$. Compared with a generative model estimated with the maximum-likelihood criterion, a discriminative model is less sensitive to model mis-specification [106][107]. A well-matched generative model has good performance with a small amount of training data [108][109][110]. A discriminative model generally achieves better classification results for real-world data such as speech, where model mis-specification is common.

For single-microphone speech separation, several limitations lead to model mis-specification. They include the finite number of states in the acoustic models, and the use of approximated interaction model for $p(y_t|\{s_{k,(t)}\})$ in factorial HMM. When $p(y_t|\{s_{k,(t)}\})$ is modeled with training data, inaccurate distribution is often assumed to allow tractable maximum likelihood parameter estimation. Moreover, insufficient

training data can be a problem in both generative and discriminative models, as discussed in Section 4.7.2.

For simplicity, let $(\mathbf{s}, \mathbf{y})$ be the training data set, where $\{\mathbf{s_k}\}$ is denoted by $\mathbf{s}$. The problem is the estimation of $p(\mathbf{s}|\mathbf{y})$. The distribution should be consistent to the sufficient statistics of $(\mathbf{s}, \mathbf{y})$. There are a few principles accounting for the estimation. The principle of minimum cross-entropy states that, given the constraints (from sufficient statistics and other knowledge) and a prior distribution of the class labels $p(\mathbf{s})$, we should choose $p(\mathbf{s}|\mathbf{y})$ with the least cross-entropy $\sum_{\mathbf{s}} p(\mathbf{s}|\mathbf{y}) \log \frac{p(\mathbf{s}|\mathbf{y})}{p(\mathbf{s})}$ [111]. If $p(\mathbf{s})$ is a uniform distribution, the principle of minimum cross-entropy is equivalent to the principle of maximum entropy [112]. The estimation of $p(\mathbf{s}|\mathbf{y})$ is formulated as an entropy maximization problem,

$$\begin{aligned} &\underset{p}{\text{maximize}}\ H(p(\mathbf{s}|\mathbf{y})) \\ &\text{subject to } \mathbb{E}\{f_i(\mathbf{s}, \mathbf{y}, t)\} = \mu_i, \quad u_i \in \mu, \forall i \\ &\qquad \sum_{\mathbf{s}} p(\mathbf{s}|\mathbf{y}) = 1, \\ &\qquad p(\mathbf{s}|\mathbf{y}) \geq 0, \qquad\qquad \forall p(\mathbf{s}|\mathbf{y}) \in p \end{aligned} \tag{5.1}$$

where $H(p(\mathbf{s}|\mathbf{y})) = -\sum_{\mathbf{s}} p(\mathbf{s}|\mathbf{y}) \log p(\mathbf{s}|\mathbf{y})$ is the Shannon entropy, $f_i(\mathbf{s}, \mathbf{y}, t)$ is the i^{th} feature function or sufficient statistics associated with $\mathbf{y}$ and $\mathbf{s}$ at time instant t, and μ is a set of mean parameters. If the problem is strictly feasible, i.e., $p(\mathbf{s}|\mathbf{y}) > 0$, the solution would be a member of the exponential family [8]. The distribution $p(\mathbf{s}|\mathbf{y})$ is modeled as a log-linear model,

$$p(\mathbf{s}|\mathbf{y}) = \frac{\exp \sum_t \sum_i \lambda_i f_i(\mathbf{s}, \mathbf{y}, t)}{Z(\mathbf{y})} \tag{5.2}$$

which is known as the maximum entropy probability distribution [8]. If the exact canonical parameters λ_i are evaluated, this distribution is consistent with the sufficient statistics of $\mathbf{y}$ and $\mathbf{s}$. The exponential terms $\exp(\lambda_i f_i(\mathbf{s}, \mathbf{y}, t))$ are referred to as potential functions. The normalization term $Z(\mathbf{y}) = \sum_{\mathbf{s}} \exp \sum_t \sum_i \lambda_i f_i(\mathbf{s}, \mathbf{y}, t)$ is referred to as the partition function.

The probability distribution in Equation 5.2 can be represented by an undirected graphical model. Since the distribution is conditioned on the observations, this graphical model is a conditional random field (CRF) [15].

5.1.2 Dynamic conditional random fields

As discussed in Section 4.4, applying linear-chain CRF is basically infeasible for single-microphone speech separation. To improve the tractability, factorial structure is adopted. Figure 5.1 illustrates the CRF applied to a two-source case. It is not a linear-chain graphical structure. The graphical structure also repeats over the time axis. In [113], this type of CRF is referred to as dynamic conditional random field (DCRF)[1]. Since the conditional independence of $\{s_{k,(t)}\}$ on y_t is generally invalid, edges connecting the nodes of different sources are used to model this potential dependence. We further elaborate the feature functions into the state feature functions $f_\alpha(\cdot)$ which describe the sufficient statistics between the states of the individual sources and the observations, and the edge feature functions $f_\beta(\cdot)$ which describe the co-occurrence of the states, either within the same sources or across different sources. By defining the corresponding λ_α and λ_β, we rewrite (5.2) as

$$\begin{aligned} p(\{\mathbf{s_k}\}|\mathbf{y}) =& \frac{1}{Z(\mathbf{y})} \exp\left(\sum_k \sum_t \sum_\alpha \lambda_\alpha f_\alpha(s_{k,t}, y_t)\right) \\ & \times \exp\left(\sum_{(a,b)\in\mathcal{E}} \sum_\beta \lambda_\beta f_\beta(s_a, s_b)\right) \end{aligned} \tag{5.3}$$

which can define arbitrary graphical structures. The complexity of the modeling process is controlled by the corresponding graphical structure and the choice of the state and edge feature functions, depending on the amount of training data and the nature of the problem.

5.1.3 Relationship with factorial HMM

The DCRF described in Section 5.1.2 is a moral graph of factorial HMM. Given the same graphical structure after moralization, there exists a set of canonical parameters and feature functions of a DCRF corresponding to the parameters of a factorial HMM that are discriminative trained with maximum mutual information (MMI) criterion for the same posterior probability $p(\{\mathbf{s_k}\}|\mathbf{y})$ [17][16][114][115],

[1]We specifically refer DCRF to the type of CRF with the same graphical structure as the moral graph of factorial HMM.

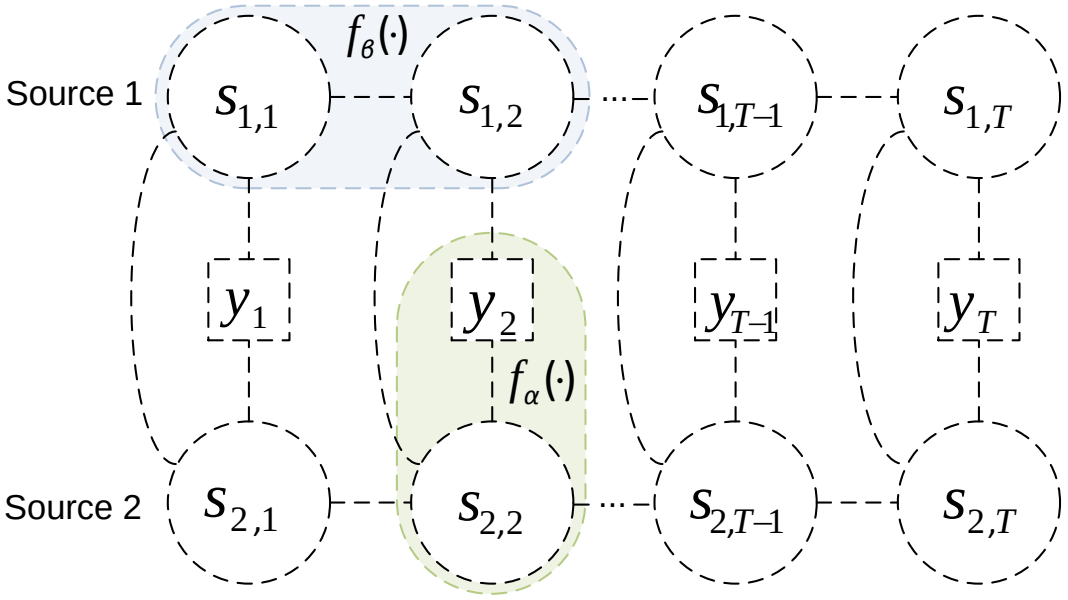


Figure 5.1: An example of DCRF for single-microphone speech separation with two sources. DCRF is defined according to an undirected graphical model. Examples of a state feature function $f_\alpha(\cdot)$ and an edge feature function $f_\beta(\cdot)$ are highlighted.

$$p(\{\mathbf{s_k}\}|\mathbf{y}) = \frac{\prod_{t=1}^{T} p(y_t|\{s_{k,(t)}\}) \times \left[\prod_{t=1}^{T}\prod_{k=1}^{K} p(s_{k,t}|s_{k,t-1})\right]}{\sum_{\{\mathbf{s_k}\}} \prod_{t=1}^{T} p(y_t|\{s_{k,(t)}\}) \times \left[\prod_{t=1}^{T}\prod_{k=1}^{K} p(s_{k,t}|s_{k,t-1})\right]}. \tag{5.4}$$

We may consider the parameter estimation for DCRF with the conditional maximum-likelihood criterion as a discriminative training technique for factorial HMM with the MMI criterion. However, there are some fundamental differences.

Parameter estimation with the MMI criterion is a constrained optimization problem. The terms $(y_t|\{s_{k,(t)}\})$ and $p(s_{k,t}|s_{k,t-1})$ in Equation 5.4 are constrained to be valid probability distributions (sum to one) at frame-level. Computing the denominator also requires the exploration of the whole search space. In automatic speech recognition research, the use of extended Baum-Welch algorithm and word lattices have been proposed to solve this problem [116][117][115]. The loopy graphical structure of factorial HMM further increases the difficulty of discriminative modeling with MMI criterion.

On the contrary, parameter estimation of DCRF is an unconstrained optimization problem. Each exponential term in Equation 5.3 is not necessary sum to one. Since

the problems are constrained differently, the parameters in factorial HMM with MMI criterion do not necessarily maximize the DCRF objective function with the corresponding feature functions. CRF requires the partition function to normalize for a valid probability distribution on sequence level. Conventional optimization techniques such as gradient descent can be applied. The gradient of the partition function can be computed effectively with forward-backward algorithm in a tree-structured graphical model, leading to a convex optimization problem for parameter estimation. In a generalized graphical model, there are some effective approximations for the gradient, although the convexity is generally lost.

Another interpretation of the discriminative capability of CRF is the possibility of integrating different types of observations or "evidence" to improve the classification performance. In HMM, the observations and the underlying labels are associated with the emission probability $p(o_t|s_t) = p(o_{t,1}, \cdots, o_{t,f}, \cdots, o_{t,F}|s_t)$, where F is the dimension of the observation vectors. When F is increased, the computation of $p(o_{t,1}, \cdots, o_{t,f}, \cdots, o_{t,F}|s_t)$ becomes challenging. Conditional independence of the observations on the class labels, i.e., $p(o_{t,1}, \cdots, o_{t,f}, \cdots, o_{t,F}|s_t) = \Pi_f^F p(o_{t,f}|s_t)$, is assumed, which is an example of model mis-specification. Although de-correlation methods such as principle component analysis (PCA) [118] may help to satisfy the conditional independent assumption, model mis-specification is still unavoidable.

In CRF, an observation is modeled as a linear combination of the feature functions. The feature functions do not necessarily correspond to the model parameters of HMM. The assumption of conditional independence is also not necessary. With appropriate feature functions, both continuous-valued and discrete-valued observations are combined in the log-linear terms. This leads to a potential advantage of flexible incorporation of different types of observations which are dependent to each other. Preliminary experiments on integrating different observations for single-microphone speech separation were performed in [80]. Although the improvement of the separation performance is marginal, this type of observation integration is shown to be feasible. Moreover, normalization of CRF at sequence level allows acoustic and temporal inference to be performed in a unified manner, with their contributions rescaled by the corresponding canonical parameters.

5.2 Statistical inference of conditional random fields

5.2.1 Parameter estimation of DCRF

The canonical parameters $\{\lambda_\alpha\}$ and $\{\lambda_\beta\}$ are estimated by minimizing the negative conditional log-likelihood of the correct state sequences. Let $\Lambda = [\ \lambda_{\alpha_1} \ \cdots \ \lambda_{\alpha_N} \ \lambda_{\beta_1} \ \cdots \ \lambda_{\beta_M}\ ]^T$ be a vector of canonical parameters, with M and N are the total number of the state and the edge feature functions respectively. The objective function for parameter estimation is written as,

$$\begin{aligned}\mathcal{L}(\Lambda) = -\sum_{r=1}^{R}\Bigg[\sum_{k}\sum_{t}\sum_{\alpha}\lambda_\alpha f_\alpha(s_{k,t}^{(r)}, y_{t,}^{(r)}) \\ + \sum_{(a,b)\in\mathcal{E}}\sum_{\beta}\lambda_\beta f_\beta(s_a^{(r)}, s_b^{(r)}) - \log Z(\mathbf{y^{(r)}})\Bigg] + c||\Lambda||_2^2\,,\end{aligned} \tag{5.5}$$

where r is the index of R training instances, c is the regularization factor and $||\Lambda||_2^2$ is the regularization term. The reference state sequences are obtained during HMM acoustic model training as described in Chapter 2. The minimization of $\mathcal{L}$ does not have a closed-form solution. The optimization is therefore performed by numerical optimization techniques such as gradient descent. The gradient descent method requires the gradient $\nabla_\Lambda \log Z(\mathbf{y^{(r)}})$,

$$\nabla_\Lambda \log Z(\mathbf{y^{(r)}}) = \begin{bmatrix} \mathbb{E}_{\lambda_{\alpha_1}}(f_{\alpha_1}(\cdot)) \\ \vdots \\ \mathbb{E}_{\lambda_{\alpha_N}}(f_{\alpha_N}(\cdot)) \\ \mathbb{E}_{\lambda_{\beta_1}}(f_{\beta_1}(\cdot)) \\ \vdots \\ \mathbb{E}_{\lambda_{\beta_M}}(f_{\beta_M}(\cdot)) \end{bmatrix}. \tag{5.6}$$

The gradients for λ_α and λ_β for each training sample are expressed as

$$\begin{aligned}\frac{\partial\mathcal{L}^{(r)}}{\partial\lambda_\alpha} &= -\sum_{t}\sum_{k} f_\alpha(s_{k,t}^{(r)}, y_{t,}^{(r)}) + \sum_{t}\sum_{k}\sum_{s_{k,t}}\mathcal{B}_{k,t}(s_{k,t}^{(r)}) f_\alpha(s_{k,t}^{(r)}, y_{t,}^{(r)}) + 2c\lambda_\alpha \\ \frac{\partial\mathcal{L}^{(r)}}{\partial\lambda_\beta} &= -\sum_{(a,b)\in\mathcal{E}} f_\beta(s_a^{(r)}, s_b^{(r)}) + \sum_{(a,b)\in\mathcal{E}}\sum_{\{s_a,s_b\}}\mathcal{B}_{ab}(s_a^{(r)}, s_b^{(r)}) f_\beta(s_a^{(r)}, s_b^{(r)}) + 2c\lambda_\beta\,,\end{aligned} \tag{5.7}$$

where $\mathcal{B}_{k,t}(s_{k,t}) = p(s_{k,t}|\mathbf{y})$ and $\mathcal{B}_{ab}(s_a, s_b) = p(s_a, s_b|\mathbf{y})$ are the the marginal probabilities. The marginal probabilities are subject to local marginalization constraints, i.e., $\sum_{s_a} \mathcal{B}_a(s_a) = 1$, $\sum_{s_a} \mathcal{B}_{ab}(s_a, s_b) = \mathcal{B}_b(s_b)$ and $\sum_{s_b} \mathcal{B}_{ab}(s_a, s_b) = \mathcal{B}_a(s_a)$. Since $\log Z(\mathbf{y}^{\mathbf{r}})$ and its gradient $\nabla_\Lambda \log Z(\mathbf{y}^{\mathbf{r}})$ are updated at each iteration, approximated inference is a trade off of accuracy and tractability in the training process. The negative log-partition function $-\log Z$ is approximated with loopy belief propagation (LBP) as [104],

$$\begin{aligned} & -\log Z \\ \approx & \sum_{(a,b)\in\mathcal{E}} \sum_{\{s_a,s_b\}} \mathcal{B}_{ab}(s_a, s_b)[\log \mathcal{B}_{ab}(s_a, s_b) - \log \eta(s_a, s_b)] \\ & - \sum_a (q_a - 1) \sum_{s_a} \mathcal{B}_a(s_a)[\log \mathcal{B}_a(s_a) - \log \Phi(s_a, y_a)] \,, \end{aligned} \tag{5.8}$$

where q_a is the number of neighbours of node a and $\eta(s_a, s_b) = \Phi(s_a, s_b)\Phi(s_a, y_a)\Phi(s_b, y_b)$. The edge potential function $\Phi(s_a, s_b)$ and the state potential function $\Phi(s_a, y_a)$ are computed from the related edge feature functions and state feature functions. Loopy belief propagation is an attempt to solve the Bethe variational problem [104],

$$\log Z \;=\; \sup_{\mu} \{\Lambda^T \mu - (-H_{Bethe}(\mathcal{B}))\} \,, \tag{5.9}$$

where $\mu = [\ \mu_{\alpha_1} \ \cdots \ \mu_{\alpha_N} \ \ \mu_{\beta_1} \ \cdots \ \mu_{\beta_M}\ ]^T$ is a vector of mean parameters of the corresponding feature functions. The mean parameters are defined as $\mu_\alpha = \sum_a \sum_{s_a} \mathcal{B}_a(s_a) f_\alpha(s_a, y_a)$ and $\mu_\beta = \sum_{(a,b)\in\mathcal{E}} \sum_{\{s_a,s_b\}} \mathcal{B}_{ab}(s_a, s_b) f_\beta(s_a, s_b)$. The Bethe entropy $H_{Bethe}(\mathcal{B})$ is defined as [104],

$$\begin{aligned} H_{Bethe}(\mathcal{B}) = & \sum_a (q_a - 1) \sum_{s_a} \mathcal{B}_a(s_a) \log \mathcal{B}_a(s_a) \\ & - \sum_{(a,b)\in\mathcal{E}} \sum_{\{s_a,s_b\}} \mathcal{B}_{ab}(s_a, s_b) \log \mathcal{B}_{ab}(s_a, s_b) \,. \end{aligned} \tag{5.10}$$

There is no guarantee on the convergence of loopy belief propagation solution. However, from the experience of factorial HMM for single-microphone speech separation, the convergence of loopy belief propagation is a minor problem on speech separation performance. If loopy belief propagation converges, the marginals $\mathcal{B} = \{\mathcal{B}_a, \mathcal{B}_{ab}\}$ are the zero gradient points of the Bethe variational problem and $-\log Z$ is re-

ferred to as the Bethe free energy [119]. Since $\log Z$ and negative Shannon entropy $-H(\mathcal{B})$ is a conjugate function pair, LBP is considered as a variational method [105][8]. However, the Shannon entropy is not the same as the Bethe entropy, i.e., $H(\mathcal{B}) \neq H_{Bethe}(\mathcal{B})$ except for a tree-structured graphical model. Even when loopy belief propagation converges, the approximated $\log Z$ is neither a upper-bound nor a lower-bound of the exact solution. As the convexity is lost, a suitable numerical optimization method is thus required.

5.2.2 Computing the posterior probabilities

The marginals $\mathcal{B}_{k,t}(s_{k,t})$ and pairwise marginals $\mathcal{B}_{ab}(s_a, s_b)$ are required in both parameter estimation and the computation of the frame-level posterior probability $p(\{s_{k,(t)}\}|\mathbf{y})$. Recall that the posterior probability $p(\{\mathbf{s_k}\}|\mathbf{y})$ is expressed with potential functions $\Phi(\cdot)$,

$$p(\{\mathbf{s_k}\}|\mathbf{y}) = \frac{1}{Z(\mathbf{y})} \prod_t \Phi(y_t, \{s_{k,(t)}\}) \prod_k \Phi(s_{k,t}, s_{k,t-1}) \,. \tag{5.11}$$

The transition potential functions $\Phi(s_{k,t}, s_{k,t-1})$ for temporal inference are computed from the edge feature functions within the same source. For acoustic inference, $\Phi(y_t, \{s_{k,(t)}\}) = \prod_k \Phi(y_t, s_{k,t}) \prod_{u,v \in K \setminus u=v} \Phi(s_u, s_v)$ are obtained from clique factorization, where $\Phi(s_u, s_v)$ are the cross-chain edge potential functions and $\Phi(y_t, s_{k,t})$ are the state potential functions. The clique size of each potential function is only two, leading to quadratic complexity with approximated inference algorithms such as loopy belief propagation. For a two-source case, the acoustic inference is expressed as $\Phi(y_t, s_{1,t}, s_{2,t}) = \Phi(y_t, s_{1,t})\Phi(y_t, s_{2,t})\Phi(s_{1,t}, s_{2,t})$.

Loopy belief propagation in CRF is similar to that of in factorial HMM. Figure 5.2 shows the propagation of messages in an undirected graph. Let $m_{ab}(s_b)$ be a message passing from node S_a to node S_b about a state variable s_b in node S_b. Belief propagation performs the following updates,

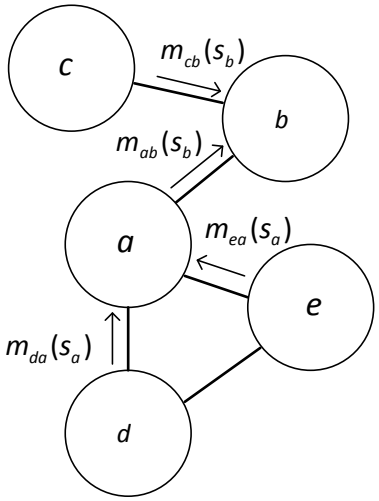


Figure 5.2: Loopy belief propagation in a general undirected graphical model. Note that every node is conditioned by observation y which is not shown in the figure

$$
\begin{aligned}
m_{ab}(s_b) \leftarrow & \kappa \sum_{s_a} \Phi(s_a, s_b)\Phi(s_a, y_a) \prod_{j \in N(a)\backslash b} m_{ja}(s_a) \\
\mathcal{B}_a(s_a) \leftarrow & \kappa' \Phi(s_a, y_a) \prod_{j \in N(a)} m_{ja}(s_a) \\
\mathcal{B}_{ab}(s_a, s_b) \leftarrow & \kappa'' \eta(s_a, s_b) \prod_{j \in N(a)\backslash b} m_{ja}(s_a) \prod_{k \in N(b)\backslash a} m_{kb}(s_b)
\end{aligned}
\tag{5.12}
$$

where $N(a)\backslash b$ denotes set of nodes neighbouring to S_a except S_b and $\kappa, \kappa', \kappa''$ are normalization constants.

Given the marginals $\mathcal{B}_{k,t}(s_{k,t})$ of each sources k and $\mathcal{B}_{(u,t)(v,t)}(s_{u,t}, s_{v,t})$ for $u, v \in K$, $u \neq v$ covering all the sources, $p(\{s_{k,(t)}\}|\mathbf{y})$ can be approximated by $\mathcal{B}(\{s_{k,(t)}\})$ as [8],

$$
\mathcal{B}(\{s_{k,(t)}\}) = \frac{\prod_{u,v \in K, u \neq v} \mathcal{B}_{(u,t)(v,t)}(s_{u,t}, s_{v,t})}{\prod_k \mathcal{B}_{k,t}(s_{k,t})^{|E|-1}}\,, \tag{5.13}
$$

where $|E|$ is the number of connections to other sources, which is equal to $|K| - 1$. For a two-source case, the posterior probability $p(s_{1,t}, s_{2,t}|\mathbf{y}) = \mathcal{B}_{(1,t)(2,t)}(s_{1,t}, s_{2,t})$ is the pairwise marginal across different sources at the same time instant. When there are more than two sources, since the marginals are only approximations due to the loops in the graphical structure, $p(\{s_{k,(t)}\}|\mathbf{y})$ can be simply approximated as

$$
p(\{s_{k,(t)}\}|\mathbf{y}) \approx \prod_k \mathcal{B}_{(k,t)}(s_{k,t})\,. \tag{5.14}
$$

Table 5.1: Speech separation results of DCRF with L-BFGS and ASGD by soft-mask filtering

(a) BSS_EVAL results

		M (1) + M (2)			M (17) + F (18)			F (24)+ F (25)			Overall		
		SDR	**SAR**	**SIR**	**SDR**	**SAR**	**SIR**	**SDR**	**SAR**	**SIR**	**SDR**	**SAR**	**SIR**
512	**L-BFGS**	6.19	9.83	**9.37**	9.58	11.96	14.22	8.81	11.57	12.74	8.19	11.12	12.11
	ASGD	**6.32**	**10.07**	9.35	**9.78**	**12.19**	**14.28**	**8.99**	**11.74**	**12.91**	**8.36**	**11.33**	**12.18**
128	**L-BFGS**	5.78	9.29	9.14	8.86	11.23	13.82	8.11	10.75	12.25	7.58	10.43	11.74
	ASGD	**5.87**	**9.49**	**9.10**	**9.05**	**11.42**	**13.98**	**8.27**	**10.96**	**12.35**	**7.73**	**10.62**	**11.81**

		M (1) + M (2)	M (17) + F (18)	F (24)+ F (25)	Overall
512	**L-BFGS**	2.41	2.67	2.58	2.56
	ASGD	**2.48**	**2.72**	**2.63**	**2.61**
128	**L-BFGS**	2.37	2.58	2.43	2.46
	ASGD	**2.41**	**2.62**	**2.46**	**2.50**

(b) PESQ results

5.2.3 Averaged stochastic gradient descent

In [113], the canonical parameters of DCRF are estimated by limited-memory Broyden–Fletcher–Goldfarb–Shanno method (L-BFGS), which is a quasi-Newton method [120]. We opt for the averaged stochastic gradient descent (ASGD) [121][122] as our optimization algorithm. During our development of the CRF formulations for single-microphone speech separation, we note that the parameters estimated by ASGD can achieve better separation performance in terms of objective quality measures, when compared with the parameters estimated by L-BFGS. We include the comparison of [81] in Table 5.1 for reference. Note that the source reconstruction is implemented with soft-mask filtering [92]. Our separation performance presented at the end of this Chapter has since been improved significantly with the estimators from the MIXMAX model introduced in Chapter 3.

The outline of the algorithm is given in Algorithm 5.1. ASGD has been successful in parameter estimation of CRFs [123][17]. The use of ASGD is justified by the following reasons. In speech processing, the training data contain a lot of redundancies. By updating parameters with each training sample, ASGD effectively

makes use of these redundancies to reduce the number of iterations. Due to the non-convexity with loopy belief propagation, we expect that ASGD helps to escape away from local extrema for better optimal points due to its stochastic nature. By choosing a suitable step-size and averaging the estimated parameters, the convergence rate of ASGD further improves [121]. We have applied the following step-size $\eta^{(m)}$ according to [124],

$$\eta^{(m)} = \frac{\eta^{(0)}}{(1 + \frac{cm\eta^{(0)}}{R})^{0.75}} \tag{5.15}$$

where η_0 is the initial step-size, m is the current iteration count, c is the regularization factor and R is the number of training samples. Let $a_{(\cdot)}^{(m)}$ be the averaged parameters. By initializing $a_{(\cdot)}^{(0)} = 0$, the averaged parameters is simply

$$a_{(\cdot)}^{(m)} = (1 - \frac{1}{m})a_{(\cdot)}^{(m-1)} + \frac{1}{m}\Lambda_{(\cdot)}^{(m)} . \tag{5.16}$$

5.3 The choice of feature functions

5.3.1 By sufficient statistics of observations

Conditional random field provides the flexibility in choosing suitable feature functions. The choice is critical to speech separation performance. When designing state feature functions $f_\alpha(\cdot)$, a straight-forward way is to associate the acoustic states with speech mixture observations. For example, a state feature function is designed for dimension f of observation y_t and state $s_{k,t}$ at frame index t,

$$f_{\alpha_{i,f}}^{(M1)}(s_{k,t}, y_{t,f}) = \begin{cases} y_{t,f} , & \text{if } s_{k,t} = i \\ 0 & \text{otherwise} \end{cases}$$

$$f_{\alpha_{i,f}}^{(M2)}(s_{k,t}, y_{t,f}) = \begin{cases} (y_{t,f})^2 , & \text{if } s_{k,t} = i \\ 0 & \text{otherwise} \end{cases}$$

where i denotes the specific state in the acoustic model, and $(M1)$ and $(M2)$ label the feature functions. The two feature functions correspond to the sufficient statistics for the first and second moments of the observations. For discrete-value observations,

Algorithm 5.1 Averaged stochastic gradient descent for CRF parameter estimation

Input: $\{\mathbf{y}\}, \{\mathbf{s_k}\}, \eta^{(0)}, c$

Output: $\Lambda, \mathbf{a}$ ▷ Stacked for element-wise operation

function ASGD($\{\mathbf{y}\}$,$\{\mathbf{s_k}\}$,$\eta^{(0)}$,c)

$m = 1$ ▷ Counter for parameter update

for $epoch$ = 1:$MaxItr$ **do** ▷ MaxItr: Maximum number of iteration

$rndOrder$=randPerm(1:R) ▷ Randomize the order of the training instances

for i = 1:R **do**

r=rndOrder(i)

$[f, \mathbf{g}]$=Loglikelihood_and_gradient($\mathbf{y}(r)$,$\mathbf{s_k}(r)$,Λ) ▷ $\mathbf{g} = \nabla_\Lambda \log Z(\mathbf{y^{(r)}})$

$\mathbf{g} = \mathbf{g} + c\nabla_\Lambda$norm($\Lambda$) ▷ norm($\Lambda$): the norm function

$sumLoss = sumLoss + f$ ▷ accumulate the conditional log-likelihood

η=adjustStepSize($\eta^{(0)}$,c,m, R)

$\Lambda = \Lambda - \eta\mathbf{g}$

$\mathbf{a} = (1 - \frac{1}{m})\mathbf{a} + \frac{1}{m}\Lambda$

$m = m + 1$

end for

$histSumLoss(epoch) = sumLoss$ ▷ Store the log-likelihood

if $epoch < period \,\&\&\, sumLoss < histSumLoss(epoch - period)$ **then**

break ▷ Reach a (local) mimimun

end if

$sumLoss = 0$

end for

return Λ,$\mathbf{a}$

end function

the sufficient statistics are represented with count-based feature functions,

$$f^{(COUNT)}_{\alpha_{i,f}}(s_{k,t}, y_{t,f}) = \begin{cases} 1\,, & \text{if } s_{k,t} = i \\ 0 & \text{otherwise} \end{cases}.$$

5.3.2 With non-linear transformations

Non-linear transformation such as a Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$ and a sigmoid function $G(u) = \int_{-\infty}^{u} K(v)dv$ can be incorporated with the feature functions. Given the multivariate Gaussian emission probability density for each acoustic state i with mean $\mu_{i,f}$ and standard derivation $\sigma_{i,f}$ at dimension f, the following state feature functions are defined,

$$f^{(KN)}_{\alpha_{i,f}}(s_{k,t}, y_{t,f}) = \begin{cases} \frac{1}{\sigma_{i,f}} K(\frac{y_{t,f}-\mu_{i,f}}{\sigma_{i,f}})\,, & \text{if } s_{k,t} = i \\ 0 & \text{otherwise} \end{cases}$$

$$f^{(SM)}_{\alpha_{i,f}}(s_{k,t}, y_{t,f}) = \begin{cases} G(\frac{y_{t,f}-\mu_{i,f}}{\sigma_{i,f}})\,, & \text{if } s_{k,t} = i \\ 0 & \text{otherwise} \end{cases}.$$

When y_t and $s_{k,t}$ are modeled in log-spectrum, the kernel $\frac{1}{\sigma_{i,f}} K(\frac{y_{t,f}-\mu_{i,f}}{\sigma_{i,f}})$ is equivalent to $p(x_{k,t,f} = y_{t,f}|s_{k,t} = i)$, and the sigmoid function $G(\frac{y_{t,f}-\mu_{i,f}}{\sigma_{i,f}})$ is equivalent to $p(x_{k,t,f} < y_{t,f}|s_{k,t} = i)$. They carry significant physical meanings according to the MIXMAX model [77][78]. The non-linear transformations map the speech mixture observations from real space $\mathbb{R}$ to probability space $[0, 1]$. We consider the application of these transformations as the revival of the idea of integrating initial separation results from factorial HMM in CRF formulations [80].

5.3.3 Defining edge features

For the edge feature functions, a count-based indicator is defined to collect the statistics of the state pair connected by an edge (a, b),

$$f_{\beta_{i,j}}(s_a, s_b) = \begin{cases} 1 & \text{, if } s_a = i \text{ and } s_b = j \\ 0 & \text{, otherwise} \end{cases} \tag{5.17}$$

where i, j denote the states of the corresponding acoustic models. The state pairs s_a and s_b can be within the same source with frames a and b adjacent to each other. This type of edge feature functions effectively models the transitions between the states along the time axis. The cross-chain edge feature functions correspond to the occurrence of a state pair of different sources but at the same time instant, as appears in Figure 5.1. They are part of the potential functions for acoustic inference in DCRF.

5.4 Experiments on DCRF

Speech separation performance with DCRF formulations is evaluated. Speech separation and automatic speech recognition experiments are carried out by following the same procedures in Chapter 4. The signal-to-signal ratio of mixtures is also set to 0 dB. The DCRF formulations are compared with the factorial HMM baseline with the GMM modeling approach. The GMM modeling approach achieves the best separation results in the previous experiments. The GMM modeling approach is a more appropriate baseline as both approaches require parameter estimation from training speech mixtures.

The results from factorial HMM ENTIREDATA and DATA_1% inferred with structured mean field (SMF) method are compared with those from DCRF formulations. Recall that ENTIREDATA is trained from 100 times more training data (200k training mixtures) than DATA_1%. For DCRF, the parameters are estimated from the same training set as DATA_1%, i.e., with 2k training mixtures. Statistical inference is performed with loopy belief propagation (LBP). Note that SMF performs slightly better than LBP in factorial HMM. The experiment setup is thus slightly favorable to factorial HMM.

Table 5.2 shows the configuration of different DCRF formulations in our experiments. DCRF formulations follow the same set of edge feature functions as defined in Section 5.3.3. To further demonstrate the benefits of integrating different observations, the 39-dimensional MFCC feature vectors from the speech mixtures are included as additional observations. MFCC observations and the log-spectra are in fact highly dependent. To apply non-linear transformations, the mean and the diagonal

Table 5.2: The settings of CRF formulations for single-microphone speech separation. State feature functions suffixed with "log" correspond to log-spectrum observations, "mfcc" correspond to MFCC observations of the speech mixtures

ID	State feature functions applied
DCRFMAG	M1-LOG, M2-LOG
DCRFTRANS	KN-LOG, SM-LOG
DCRFMFCC	KN-LOG, SM-LOG, KN-MFCC, SM-MFCC

covariance of the clean MFCC vectors from the sources are estimated for each state.

The number of parameters is an indicator of complexity. The empirical number of parameters of DCRF and factorial HMM for two-source case, and the corresponding theoretical maximum are given in Table 5.3. When applying the GMM modeling approach, the number of parameters in factorial HMM can be over 100 times of DCRF formulations. This implies that a large amount of training data may be required.

The BSS_EVAL metrics and PESQ are used to evaluate the objective speech quality of the reconstructed sources. Standard word error rate (WER) is adopted as the performance metric for the automatic speech recognition. Experimental results in terms of PESQ and WER are shown in Figure 5.3. The results in terms of BSS_EVAL metrics are shown in Table 5.4. The results are averaged from 2500 reconstructed sources for each speaker pair.

5.4.1 Comparison with factorial HMM

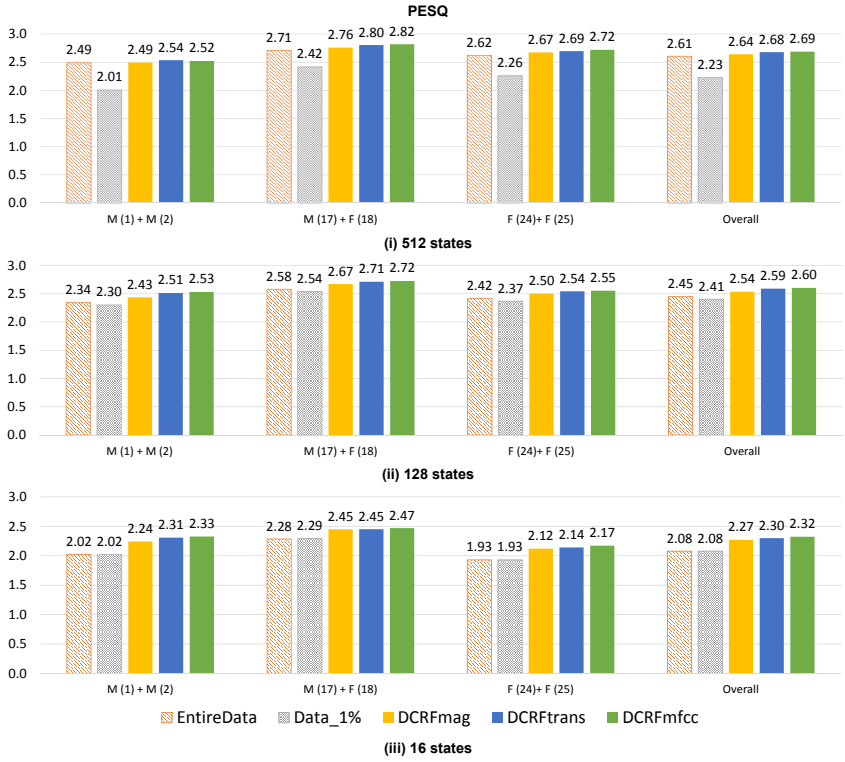
Almost all DCRF formulations achieve higher PESQ and significantly lower WER than those by factorial HMM ENTIREDATA. The exceptions are the cases with 512 states, and 128 states with DCRFMAG. This trend is consistent across different number of states and different speaker pairs. DCRF also demonstrates substantial improvement in terms of BSS_EVAL metrics in the successful cases. The improvement on SIR and SAR suggests that DCRF formulations perform better in suppressing the interfering sources and reducing the reconstruction artifacts. As the same conditional mean estimator $\mathbb{E}(x_{k,t}|\mathbf{y}, \{s_{k,t}\})$ is applied for the source reconstruction in both fac-

Table 5.3: The numbers of parameters of different DCRF and factorial HMM formulations. The number of sources is denoted as K, the number of acoustic states is denoted as S , the dimension of speech mixture observations is denoted as F, and $G = 2KSD$ is the number of parameters from the emission probabilities of the acoustic states. The statistics are averaged from the 3 evaluation sets

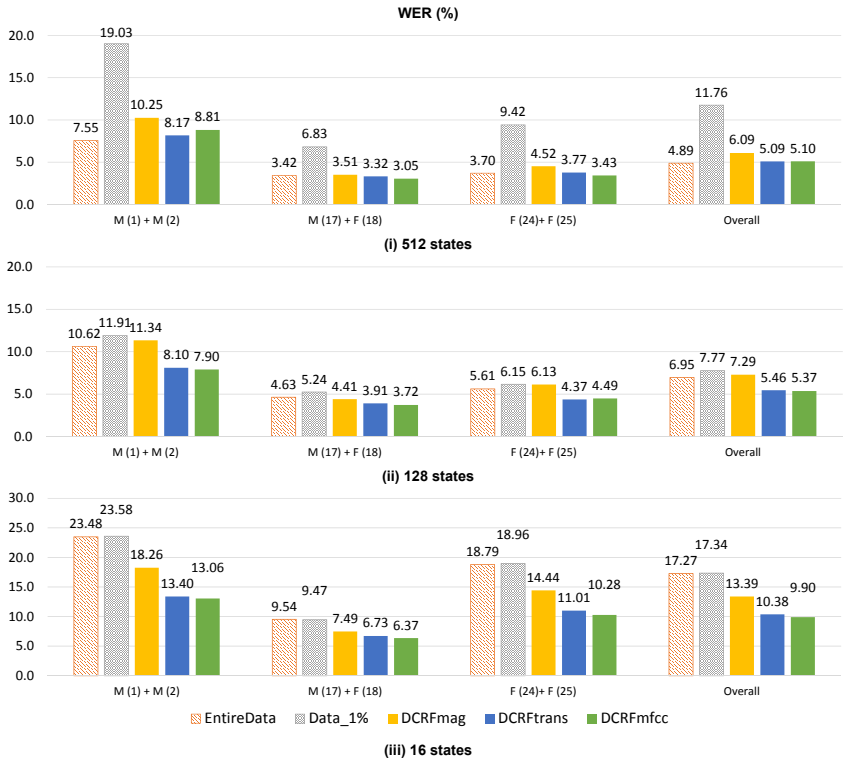
ID	Max. # of parameters	# parameters for 2 sources		
		$\mathbf{S = 16}$ $\mathbf{G = 16.4k}$	$\mathbf{S = 128}$ $\mathbf{G = 132k}$	$\mathbf{S = 512}$ $\mathbf{G = 526k}$
DCRFMAG	$KSF + KS^2 + \frac{1}{2}K(K-1)S^2$	15.6k	128k	498k
DCRFTRANS	$KSF + KS^2 + \frac{1}{2}K(K-1)S^2 + G$	15.6k+G	128k+G	498k+G
DCRFMFCC		17.9k+G	146k+G	559k+G
Factorial HMM	$2S^K D + KS^2$	132k	8.45M	135M

Table 5.4: Separation results of different DCRF formulations and factorial HMM in terms of BSS_EVAL with 16, 128 and 512 acoustic states

		M (1) + M (2)			M (17) + F (18)			F (24)+ F (25)			Overall		
		SDR	**SAR**	**SIR**	**SDR**	**SAR**	**SIR**	**SDR**	**SAR**	**SIR**	**SDR**	**SAR**	**SIR**
512	**EntireData**	6.36	**10.39**	9.14	9.96	12.40	14.34	9.21	12.17	12.79	8.51	11.65	12.09
	Data_1%	4.13	8.76	6.76	8.69	11.00	13.36	7.70	10.57	11.48	6.84	10.11	10.53
	DCRFmag	**6.52**	10.29	**9.54**	**10.31**	12.64	**14.76**	9.48	12.30	**13.21**	**8.77**	11.74	**12.50**
	DCRFtrans	6.49	10.27	9.51	10.18	12.59	14.62	9.41	12.29	13.07	8.69	11.72	12.40
	DCRFmfcc	6.48	10.22	**9.54**	10.23	**12.65**	14.64	**9.51**	**12.43**	13.11	8.74	**11.77**	12.43
128	**EntireData**	5.79	9.83	8.63	9.26	11.59	14.03	8.34	11.15	12.18	7.80	10.86	11.62
	Data_1%	5.55	9.57	8.44	9.04	11.35	13.86	8.18	10.95	12.06	7.59	10.62	11.45
	DCRFmag	6.21	9.83	9.42	**9.77**	11.87	**14.78**	8.79	11.33	**12.95**	8.26	11.01	**12.38**
	DCRFtrans	6.29	**10.24**	9.16	9.69	**12.09**	14.24	8.82	11.61	12.62	8.27	**11.31**	12.01
	DCRFmfcc	**6.43**	10.04	**9.59**	9.71	12.04	14.41	**8.87**	**11.62**	12.72	**8.34**	11.23	12.24
16	**EntireData**	3.99	8.47	6.71	7.72	10.19	12.50	5.48	8.57	9.28	5.73	9.08	9.50
	Data_1%	4.25	8.83	6.88	7.92	10.32	12.73	5.85	8.92	9.64	6.01	9.35	9.75
	DCRFmag	5.50	8.92	**9.00**	**8.60**	**10.91**	**13.62**	7.01	9.36	**11.66**	7.03	9.73	**11.42**
	DCRFtrans	5.55	**9.33**	8.67	8.47	10.79	13.49	6.91	9.69	10.94	6.98	9.94	11.03
	DCRFmfcc	**5.63**	9.26	8.86	8.52	10.82	13.60	**7.04**	**9.80**	11.10	**7.06**	**9.96**	11.19



(a)



(b)

Figure 5.3: Separation results of different DCRF formulations and factorial HMM in terms of (a) PESQ and (b)WER (%) with (i) 512, (ii) 128 and (iii) 16 acoustic states

torial HMM and DCRF formulations, the performance improvement is purely due to the improved inference of the posterior probabilities of the source acoustic states.

As a discriminative models, DCRF demonstrates its robustness against insufficient training data which is a kind of model mis-specification with 512 states. All the formulations maintain reasonable performance (slightly better PESQ than ENTIREDATA) even with much fewer training data than ENTIREDATA. With this amount of training data, factorial HMM (DATA_1%) is completely a failure with significant drop of objective quality metrics PESQ and BSS_EVAL.

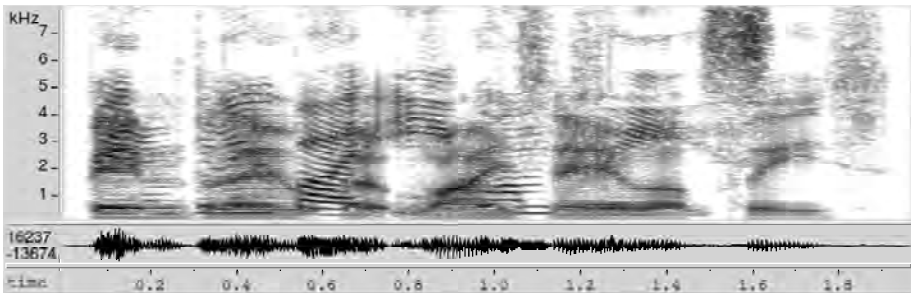
5.4.2 Comparison among DCRF

DCRF formulations with non-linear transformations (DCRFTRANS and DCRFMFCC) attain slightly better PESQ, and significantly lower WER than the formulation without non-linear transformations (DCRFMAG). DCRFTRANS and DCRFMFCC achieve considerable WER improvement with 128 states. With 512 states, the performance is slightly worse than factorial HMM. In contrast, the performance improvement with DCRFMAG reduces dramatically with increased number of states. The results of DCRFMFCC show the benefits of integrating multiple observations. It performs slightly better than DCRFTRANS in terms of both PESQ and WER.

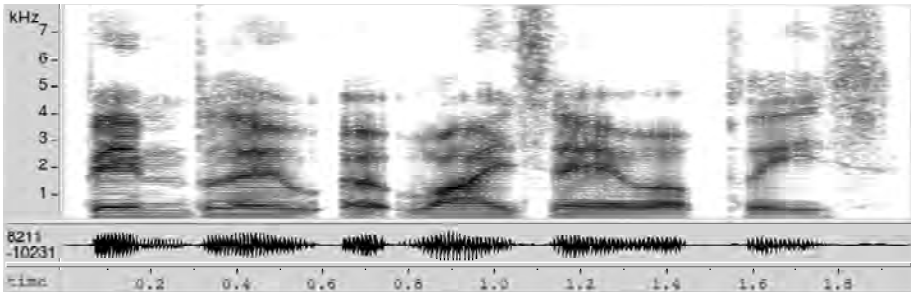
We suspect that over-fitting with insufficient training data begins to affect the performance of DCRF. A syndrome of over-fitting is observed in *Male* + *Male* set with DCRFTRANS and DCRFMFCC. Under these formulations, WER with 128 states is even lower than WER with 512 states. For DCRFMFCC, the performance degradation is as large as 11.5%. We expect that the problem will be disappeared when the amount of training data increases. In Chapter 6, we also investigate a large-margin parameter estimation method to minimize the effect of over-fitting.

5.4.3 Samples of the reconstructed sources

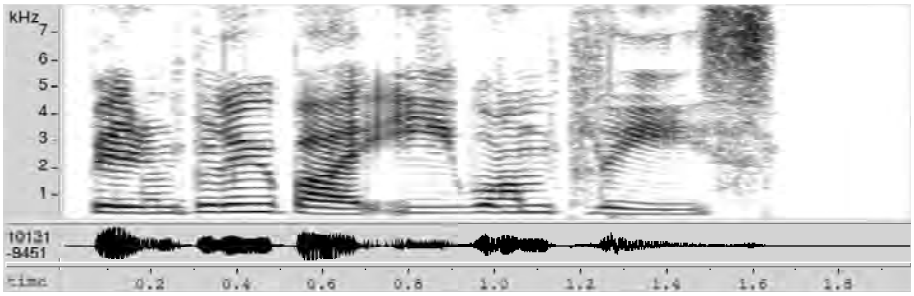
An example from *Male* + *Female* evaluation set is presented. The spectrograms of the speech mixture and the reference speech sources are shown in Figure 5.4. The spec-



(a) speech mixture



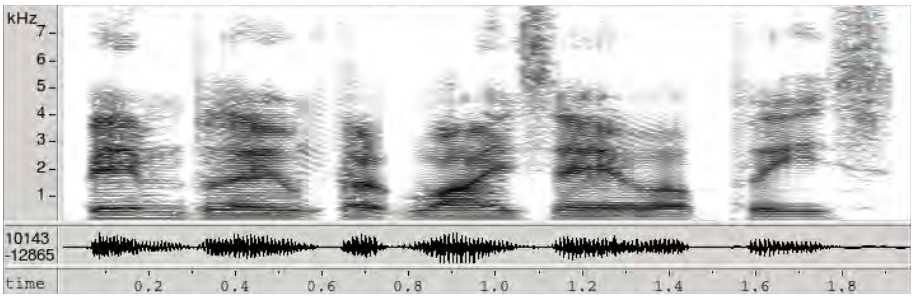
(b) Reference male



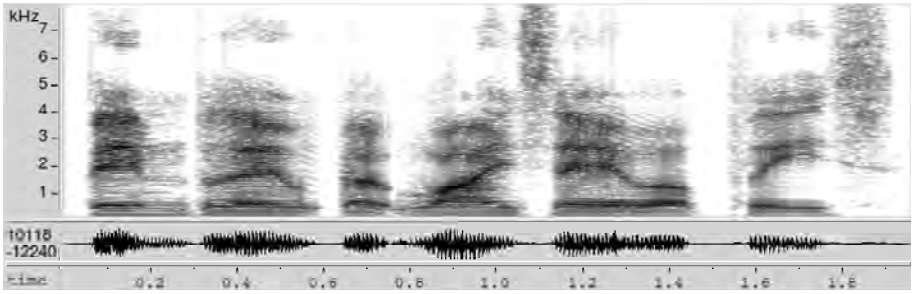
(c) Reference Female

Figure 5.4: Spectrograms and waveforms of a speech mixture and the reference signals from the *Male + Female* set

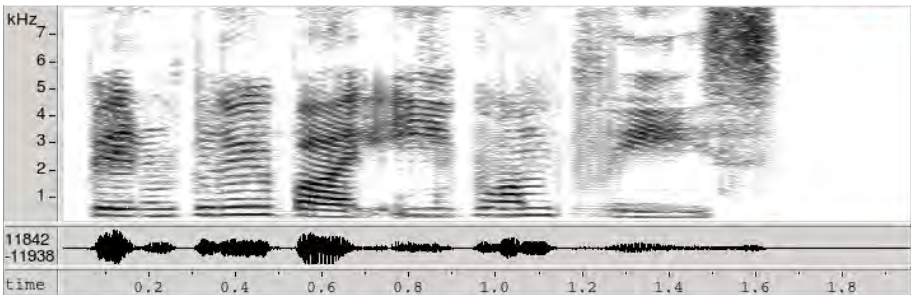
trogram of the reconstructed speech sources by factorial HMM and DCRFTRANS are shown in 5.5. Some residues of the harmonic structure of the competing source are still remained in the reconstructed source by factorial HMM (for example at time 0.6 s of the Male speaker) . These residues are much reduced in DCRFTRANS.



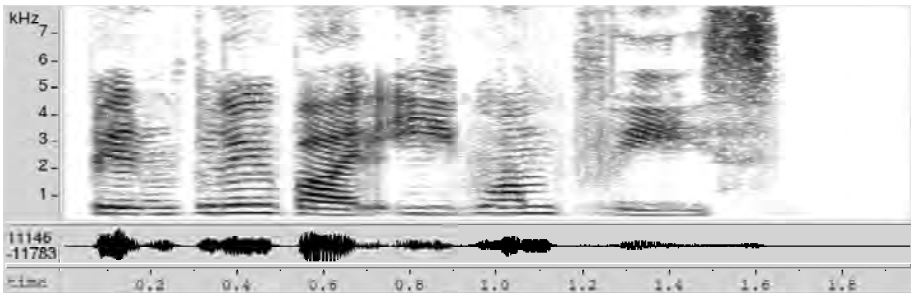
(a) Factorial HMM - male



(b) DCRFTRANS - male



(c) Factorial HMM - female



(d) DCRFTRANS - female

Figure 5.5: Spectrograms and waveforms of the reconstructed speech sources. Acoustic models with 512 acoustic states are applied

Chapter 6

Extensions of conditional random fields for speech separation

6.1 Large-margin training for CRF

Motivated by the success of supporting vector machine (SVM) [125], margin-based discriminative classifiers have gained much attention in various pattern recognition problems. Large-margin modeling aims at finding the decision boundaries which separate different classes with maximum distances. It has been applied to automatic speech recognition [126][127][128], hand-written digit recognition [129] and many other tasks. The success of large-margin classifiers is supported by strong theoretical guarantee on generalization [130][131]. Recently, large-margin technique has been extended for sequential classifications. Formulations such as max-margin Markov networks (M^3) [131] and large-margin hidden Markov models (HMM) [132] were proposed. Similar to large-margin HMM [132], a large-margin CRF formulation can be established with minimal modification from conditional maximum-likelihood criterion.

6.1.1 Objective function for large-margin CRF training

In speech separation with CRF, the classification output is represented by the acoustic state sequences $\{\mathbf{s_k}\}$, which maximizes $p(\{\mathbf{s_k}\}|\mathbf{y}) = \frac{1}{Z(\mathbf{y})} \exp\Big(\Phi(\{\mathbf{s_k}\}|\mathbf{y})\Big)$ or

$\log p(\{\mathbf{s_k}\}|\mathbf{y}) = \Phi(\{\mathbf{s_k}\}|\mathbf{y}) - \log Z(\mathbf{y})$. Denote $\{\mathbf{s_k}\}$ as $\mathbf{s}$ for simplicity. Let $\mathcal{H}(\mathbf{s}, \mathbf{s}')$ be a distance metric for the sequences $\mathbf{s}$ and $\mathbf{s}'$, e.g., the hamming distance [131]. Given that the correct sequence of r^{th} training sample is $\mathbf{s^{(r)}}$, large-margin criterion aims at finding the canonical parameters Λ such that the log-potential function $\Phi(\mathbf{s^{(r)}}|\mathbf{y})$[1] of the correct sequence is greater than $\Phi(\mathbf{s}|\mathbf{y})$ of any incorrect sequences by $\mathcal{H}(\mathbf{s}, \mathbf{s^{(r)}})$,

$$\begin{aligned} &\underset{\Lambda}{\text{minimize}} \sum_{\mathbf{s}\setminus\mathbf{s}=\mathbf{s^{(r)}},r} c||\Lambda||_2^2 \\ &\text{subject to } \Phi(\mathbf{s^{(r)}}|\mathbf{y^{(r)}}) - \Phi(\mathbf{s}|\mathbf{y^{(r)}}) \geq \mathcal{H}(\mathbf{s}, \mathbf{s^{(r)}}), \forall \mathbf{s} \neq \mathbf{s^{(r)}}, r\,. \end{aligned} \tag{6.1}$$

The parameters can be rescaled to produce arbitrary large margins [132]. The prior distribution of the parameters is defined by the regularization term $c||\Lambda||_2^2$ to limit the "size" of the parameters. Moreover, there are exponential number of constraints due to exponential number of possible sequences. As suggested in [132], the constraints can be rewritten into a single constraint for each training sample

$$\Phi(\mathbf{s^{(r)}}|\mathbf{y^{(r)}}) - \max_{\mathbf{s}\neq\mathbf{s^{(r)}}} \left\{\Phi(\mathbf{s}|\mathbf{y^{(r)}}) + \mathcal{H}(\mathbf{s}, \mathbf{s^{(r)}})\right\} \geq 0, \quad \forall r\,. \tag{6.2}$$

When we include the trivial case $\Phi(\mathbf{s}|\mathbf{y^{(r)}}) - \Phi(\mathbf{s^{(r)}}|\mathbf{y^{(r)}}) = 0$ for $\mathbf{s} = \mathbf{s^{(r)}}$, we get an equality constraint,

$$\Phi(\mathbf{s^{(r)}}|\mathbf{y^{(r)}}) - \max_{\mathbf{s}} \left\{\Phi(\mathbf{s}|\mathbf{y^{(r)}}) + \mathcal{H}(\mathbf{s}, \mathbf{s^{(r)}})\right\} = 0, \quad \forall r\,. \tag{6.3}$$

Since the $\max\{\cdot\}$ function is non-differentiable, soft-max approximation is applied. Moreover, since $\max\{\cdot\} \leq \text{logsumexp}\{\cdot\}$, slack variables $\zeta_r \geq 0$, $\forall r$ are introduced to the constraint set such that $\max\{\cdot\} = \text{logsumexp}\{\cdot\} - \zeta_r$. We should minimize the slack variables ζ_r such that logsumexp$\{\cdot\}$ is as close to $\max\{\cdot\}$ as possible. A small ζ_r also implies that $\Phi(\mathbf{s^{(r)}}|\mathbf{y^{(r)}})$ is much larger than $\Phi(\mathbf{s}|\mathbf{y^{(r)}}) + \mathcal{H}(\mathbf{s}, \mathbf{s^{(r)}})$, $\forall \mathbf{s} \neq \mathbf{s^{(r)}}$ when (6.3) is satisfied. When (6.3) is not satisfied, the slack variables allow the violation of the constraint as in soft-margin formulation of SVM [125]. The violation should be minimized in hope that $\Phi(\mathbf{s^{(r)}}|\mathbf{y^{(r)}})$ is still greater than $\Phi(\mathbf{s}|\mathbf{y^{(r)}})$, $\forall \mathbf{s} \neq \mathbf{s^{(r)}}$. The optimization problem is now formulated as

[1] Here we re-define $\Phi(\cdot)$ as a log-potential function for simplicity.

$$\begin{aligned} &\underset{\zeta,\Lambda}{\text{minimize}} \ \sum_r \zeta_r + c||\Lambda||_2^2 \\ &\text{subject to} \ -\Phi(\mathbf{s^{(r)}}|\mathbf{y^{(r)}}) + \log\sum_{\mathbf{s}} e^{\Phi(\mathbf{s}|\mathbf{y^{(r)}})+\mathcal{H}(\mathbf{s},\mathbf{s^{(r)}})} = \zeta_r, \ \forall r \\ &\qquad\qquad \zeta_r \geq 0, \qquad\qquad \forall r\,. \end{aligned} \tag{6.4}$$

The slack variable ζ_r is the conditional negative log-likelihood of the correct sequence $\mathbf{s^{(r)}}$ of r^{th} training data. Hence, (6.4) is transformed into an unconstrained minimization problem with objective function

$$\mathcal{L}(\Lambda) = \sum_{r=1}^{R} \left(-\Phi(\mathbf{s^{(r)}}|\mathbf{y^{(r)}}) + \log\sum_{\mathbf{s}} e^{\Phi(\mathbf{s}|\mathbf{y^{(r)}})+\mathcal{H}(\mathbf{s},\mathbf{s^{(r)}})} \right) + c||\Lambda||_2^2\,. \tag{6.5}$$

An identical objective function was derived in [133], but our derivation emphasizes the bounding property of soft-maximum. The probability distribution $p(\mathbf{s}|\mathbf{y^{(r)}})$ is derived as

$$p(\mathbf{s}|\mathbf{y^{(r)}}) = \frac{\exp\left[\Phi(\mathbf{s}|\mathbf{y^{(r)}}) + \mathcal{H}(\mathbf{s},\mathbf{s^{(r)}})\right]}{\sum_{\mathbf{s}}\exp\left[\Phi(\mathbf{s}|\mathbf{y^{(r)}}) + \mathcal{H}(\mathbf{s},\mathbf{s^{(r)}})\right]}\,. \tag{6.6}$$

Similar to CRF training and MMI discriminative training for HMM, large-margin CRF training resembles boosted MMI discriminative training [134][135]. For speech separation, the large-margin objective function is expressed as

$$\begin{aligned} \widetilde{\mathcal{L}}(\Lambda) = \sum_{r=1}^{R} \Bigg[&-\left(\sum_k \sum_t \sum_\alpha \lambda_\alpha f_\alpha(s_{k,t}^{(r)}, y_{t,}^{(r)}) + \sum_{(a,b)\in\mathcal{E}} \sum_\beta \lambda_\beta f_\beta(s_a^{(r)}, s_b^{(r)}) \right) \\ &+ \log\widetilde{Z}(\mathbf{y^{(r)}}) \Bigg] + c||\Lambda||_2^2\,, \end{aligned} \tag{6.7}$$

where $\widetilde{Z}(\mathbf{y^{(r)}}) = \sum_{\mathbf{s}}\exp\left[\Phi(\mathbf{s}|\mathbf{y^{(r)}}) + \mathcal{H}(\mathbf{s},\mathbf{s^{(r)}})\right]$ is the modified partition function and

$$\mathcal{H}(s_{k,t}, s_{k,t}^r) = \begin{cases} 1, & s_{k,t} \neq s_{k,t}^r \\ 0, & s_{k,t} = s_{k,t}^r \end{cases},$$

which is the error count of recognized states. The gradients of $\widetilde{\mathcal{L}}$ are exactly the same as the ones in Equation 5.7 and hence the procedures for parameter updates. The only modification is to include $\mathcal{H}(s_{k,t}, s_{k,t}^{(r)})$ in the computation of the potential functions. No modification is required for computing the posterior probability $p(\{s_{k,(t)}\}|\mathbf{y})$ during the separation process.

6.1.2 Separation performance

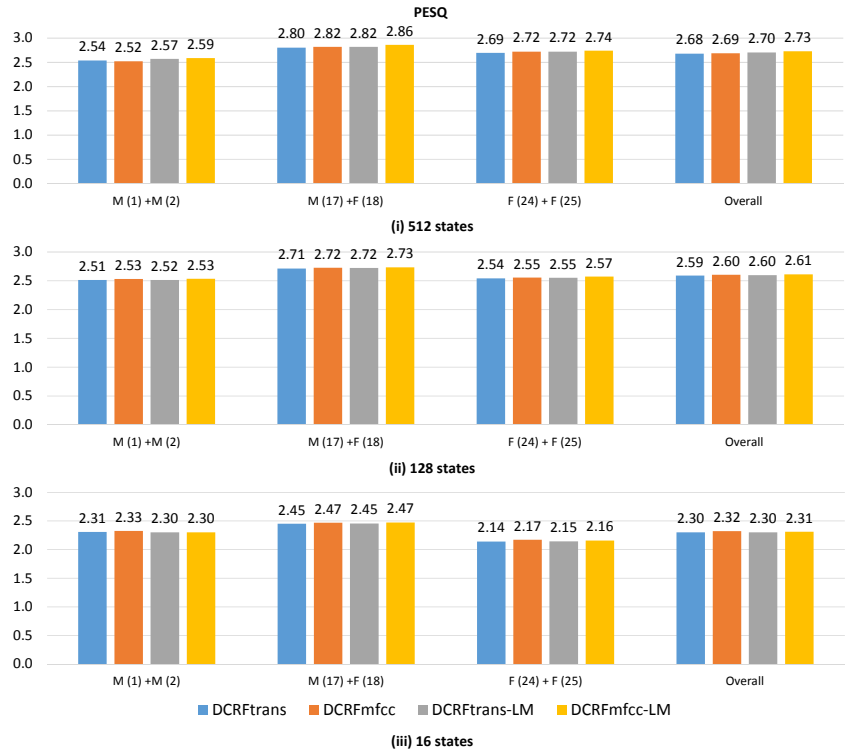
The experiments with DCRFTRANS and DCRFMFCC are repeated, except that the parameters are obtained from the large-margin criterion. The new results suffixed with "-LM" are listed in Figure 6.1.

For the cases with 16 and 128 acoustic states, the large-margin method slightly improves WER (up to 8.24% relative WER reduction) for the speech recognition task on the reconstructed sources. The improvement on PESQ is insignificant. For 512 states, the improvement on speech recognition accuracy becomes more noticeable. A relative WER reduction up to 16.57% is observed. There is also some improvement on PESQ. For *Male + Male* set, the WER with 512 states is now lower than that with 128 states under DCRFTRANS-LM and DCRFMFCC-LM. This indicates that the over-fitting problem is alleviated by better generalization abilities of large-margin training. Moreover, DCRF with large-margin criterion finally achieves the lowest word recognition error.

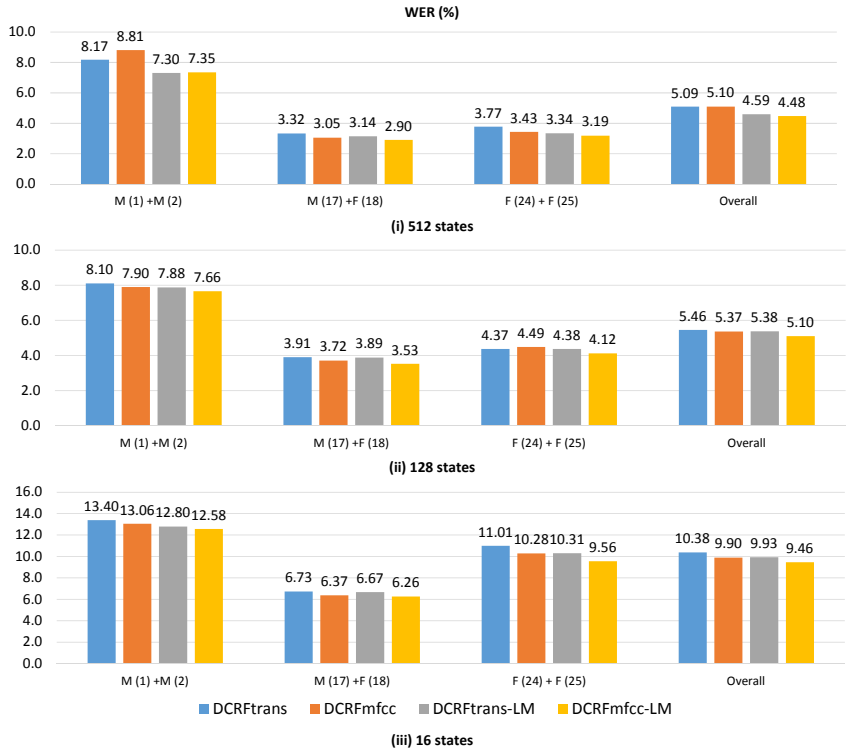
6.1.3 Convergence analysis

The effects of different initial step-sizes $\eta^{(0)}$ and regularization factors c are investigated. Two initial step-sizes $\eta^{(0)} = 0.00781 \approx \frac{1}{128}$ and $\eta^{(0)} = \frac{0.00781}{4}$ are evaluated on the *Male + Female* set with 128 states. For both initial step-sizes $\eta^{(0)}$, ASGD terminates before reaching the maximum number of iterations (1000 iterations), suggesting that the algorithm converges successfully. The separation results with canonical parameters Λ obtained from different number of iterations are shown in Figure 6.2. The results show that the choice of initial step-size does not seriously affect the speech separation performance. Similar PESQ and WER ($< 3\%$ difference) are achieved with different $\eta^{(0)}$. However, the number of iterations varies significantly with different $\eta^{(0)}$. For $\eta^{(0)} = \frac{0.00781}{4}$, 550 iterations are required for convergence, but only 90 iterations are required for $\eta^{(0)} = 0.00781$.

The results further confirm that ASGD requires much fewer iterations to reach reasonable separation performance. The results with the parameters obtained suring the mid-way of the gradient descent (50 and 250 iterations for $\eta^{(0)} = 0.00781$ and



(a)



(b)

Figure 6.1: Separation results in terms of (a) PESQ and (b) WER (%) of large-margin CRF formulations

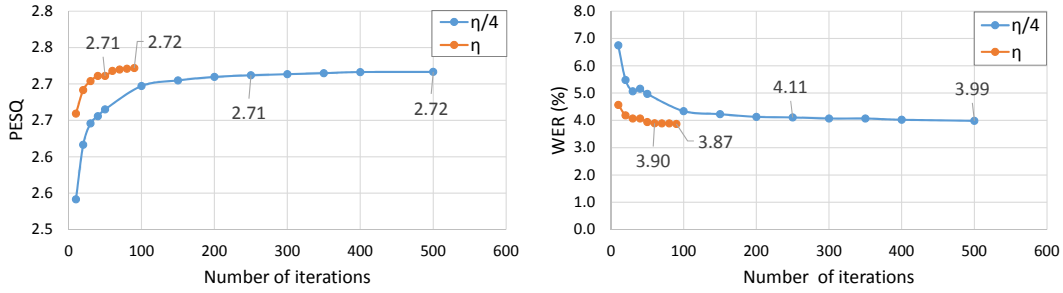


Figure 6.2: Separation results of different step-sizes $\eta = 0.00781$ and $\frac{\eta}{4}$ on *Male* + *Female* set with 128 states and regularization factor $c = 0.2$ according to DCRFTRANS-LM formulation

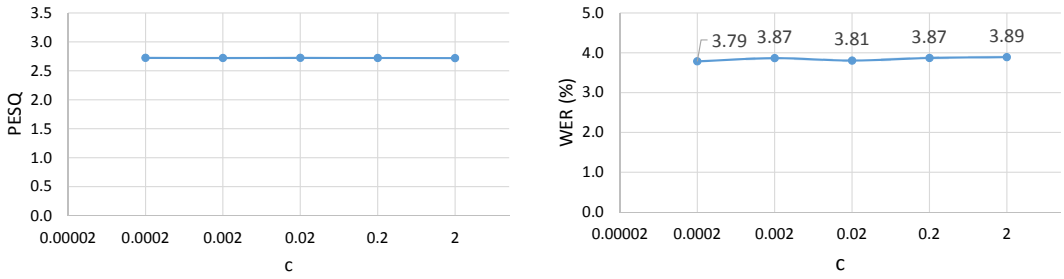


Figure 6.3: Separation results of different regularization factors $c = \{2, 0.2, 0.02, 0.002, 0.0002\}$ on *Male* + *Female* set with 128 states according to DCRFTRANS-LM formulation

$\eta^{(0)} = \frac{0.00781}{4}$ respectively) are already comparable to the results with parameters obtained after the termination of ASGD. The difference in terms of WER is less than 3%. The difference in terms of PESQ is insignificant.

We also evaluate the separation performance with different regularization factors $c = \{2, 0.2, 0.02, 0.002, 0.0002\}$. As shown in Figure 6.3, the speech separation performance generally is not affected by the choice of the regularization factors when large-margin parameter estimation is applied.

6.2 Simplified CRF formulations

The computational problem in DCRF is due to the loopy structure made up by the edges across different sources. A relaxation is possible by removing these cross-

chain edges and retaining the edges within the same source. This formulation is referred to as JOINTCRF and was proposed in [80]. The underlying assumption is that the source states are conditionally independent given the speech mixture. With this assumption, $p(\{\mathbf{s_k}\}|\mathbf{y})$ is defined as

$$\begin{aligned} p(\{\mathbf{s_k}\}|\mathbf{y}) &= \prod_k p(\mathbf{s_k}|\mathbf{y}) \\ &= \prod_k \frac{\exp \sum_t \sum_i \lambda_i f_i(\mathbf{s_k}, \mathbf{y}, t)}{Z_k(\mathbf{y})} \,. \end{aligned} \tag{6.8}$$

By removing the cross-chain edges, the graphical structure can be considered as a forest of two linear-chain CRFs conditioned on the same observations, as shown in Figure 6.4. Observation integration can be performed as if conventional CRF. Parameter estimation of JOINTCRF is effective given a modest amount of training data, since the partition function $Z_k(\mathbf{y})$ of each linear-chain CRF can be computed exactly with the forward-backward algorithm [95]. Forward-backward algorithm is also applied for exact computation of the posterior probabilities of the source states during speech separation. Since the partition function and the marginal probabilities can be computed exactly with respect to the graphical model structure, statistical inference of JOINTCRF is a convex optimization problem.

JOINTCRF also reduces the complexity of statistical inference. The overall complexity is $O(TKS^2)$, which increases linearly with the number of sources. In contrast, DCRF formulation with the loopy belief propagation requires quadratic complexity with the number of sources. Moreover, without the cross-chain edge feature functions, JOINTCRF requires fewer canonical parameters. This may lessen the required amount of training data.

6.2.1 A discussion on the "correct" model

JOINTCRF is a typical example of mis-specified model. The conditional independence assumption of the source states given the speech mixture are sometimes violated. A counterexample is that given a speech mixture of two sources, if we know that one of the sources is silent, conditional independence assumption is broken as we immediate know that the another source is the same as the speech mixture.

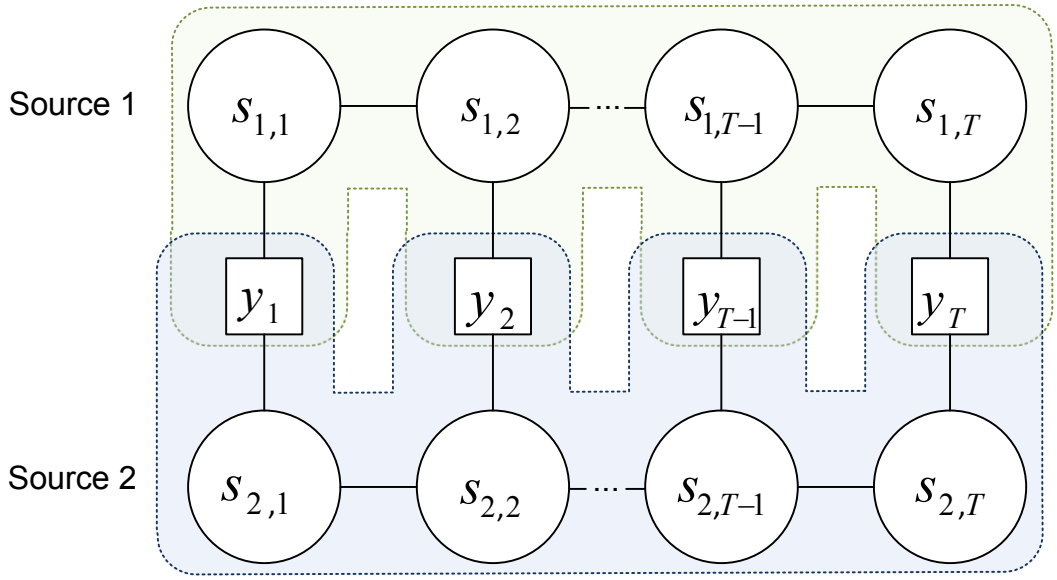


Figure 6.4: A two-source single-microphone separation problem modeled in JOINTCRF formulation. It is a forest of 2 linear-chain CRFs sharing the same observations of speech mixture.

Both factorial HMM and DCRF model the potential state dependency and achieve quite good speech separation performance. However, we cannot claim that factorial HMM and DCRF are the canonical models for speech separation. We cannot guarantee that all the properties of single-microphone speech separation are included in these models. As we generally do not know the underlying true model, discriminative modeling is justified in statistical model-based methods to handle the potential model mis-specification.

The choice of JOINTCRF and DCRF should be better described as a trade-off in practical situation. DCRF is a more accurate model but also more complex due to the additional constraints for potential state dependency. When the training condition is well-matched with speech separation condition, e.g., with the same signal-to-signal ratio, DCRF formulations are expected to achieve better separation results. However, the performance of DCRF may be limited by insufficient training data due to more parameters and the locally optimal solution from approximated inference. Although JOINTCRF is a simpler model, the estimated parameters are (near) globally optimal due to the convexity of the parameter estimation problem. Observation integration is

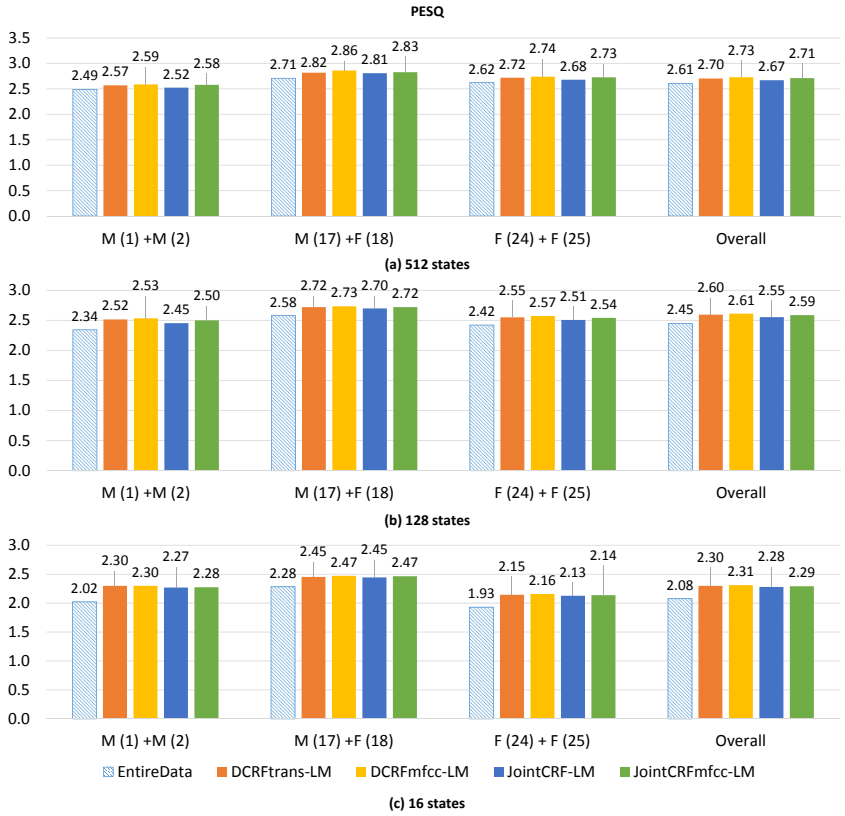
always feasible for JOINTCRF, which also helps to compensate the potential performance loss.

When the training condition are mis-matched with speech separation condition, such as under different signal-to-signal ratios. The input observations are perturbed with respect to the parameters obtained in training stage. In this case, the non-convex approximated inference algorithms of DCRF can be problematic. There may be several locally optimal solutions and the algorithm may converge to a solution far from the true one. Algorithms such as loopy belief propagation may even fail to converge with the perturbed input, resulting a poor solution. In contrast, the statistical inference of JOINTCRF is a convex problem, which always converges to a unique solution. JOINTCRF is hence considered as a more stable model [136]. Compared with DCRF, JOINTCRF may achieve better performance when mismatch becomes more serious.

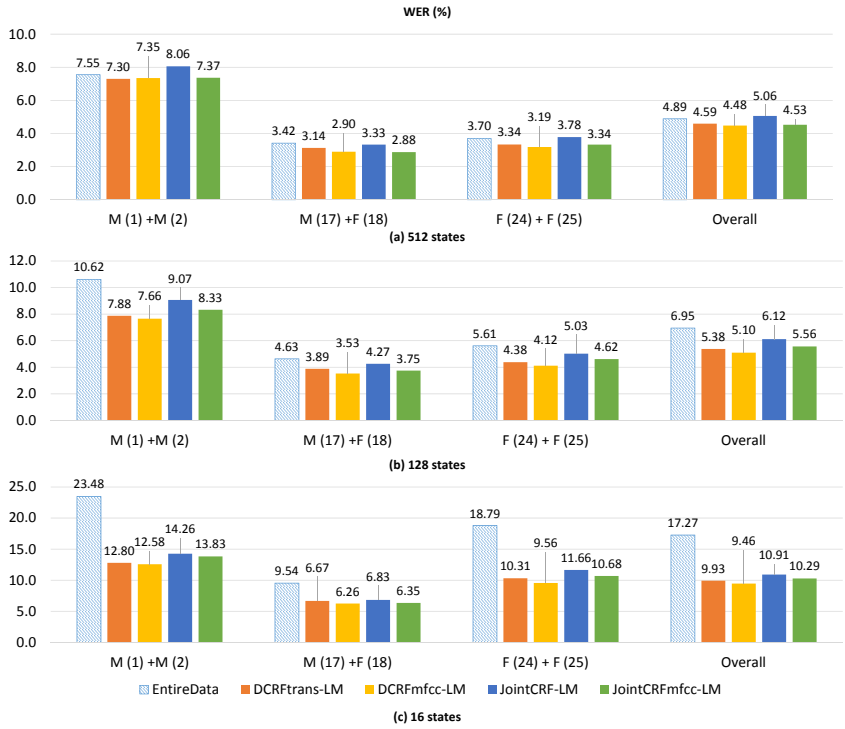
6.2.2 Experimental results

In the experiments, we compare the speech separation results from large-margin formulation of JOINTCRF denoted as JOINTCRF-LM. We also include the settings integrated with MFCC speech mixture observations denoted as JOINTCRFMFCC-LM. We compare the separation results with factorial HMM baseline ENTIREDATA, and large-margin DCRF formulations DCRFTRANS-LM and DCRFMFCC-LM. Following the settings of DCRF formulations, the speech mixture observations consist of 257-dimension log-spectrum for JOINTCRF-LM and DCRFTRANS-LM, and additional 39-dimensional MFCC feature vectors for JOINTCRFMFCC-LM and DCRFMFCC-LM. Non-linear transformations are applied on all the CRF formulations. The empirical number of parameters of DCRFTRANS-LM and JOINTCRF-LM formulations are listed in Table 6.1. The numbers of parameters of JOINTCRF-LM with 16 and 128 acoustic states are fewer than those of DCRFTRANS-LM formulations by 1% and 5% respectively. When the number of acoustic states grows to 512, the number of parameters of JOINTCRF-LM reduces more significantly by about 15%.

The speech separation results are listed in Figure 6.5. While JOINTCRF-LM



(a)



(b)

Figure 6.5: Separation results of JOINTCRF in terms of (a) PESQ and (b) WER (%)

Table 6.1: The numbers of parameters between DCRF and JOINTCRF formulations

	$\mathbf{S = 16}$	$\mathbf{S = 128}$	$\mathbf{S = 512}$
JOINTCRF-LM	15.4k	122k	427k
DCRFTRANS-LM	15.6k	128k	498k

achieves similar PESQ (< 0.1 difference) as DCRFTRANS-LM, the performance of JOINTCRF-LM in terms of WER is generally poorer (about 10% WER difference). Nevertheless, JOINTCRF-LM still performs better than factorial HMM in terms of PESQ and WER with 16 and 128 states. For 512 states, the performance of JOINTCRF-LM is still reasonable (about 5% WER). Note that factorial HMM ENTIREDATA is trained with 100 times more training data. With the same amount of training data, JOINTCRF-LM performs much better than factorial HMM HMMDATA_1% as shown Chapter 4. The results of JOINTCRF-LM demonstrate the robustness of discriminative models for withstanding model uncertainties.

The integration of MFCC speech mixture observations is more effective in JOINTCRF than DCRF. For 128 and 512 states, about 10% relative WER reduction is observed after the observation integration. For DCRF, less than 5% relative WER reduction are observed. The performance gap between JOINTCRFMFCC-LM and DCRFMFCC-LM becomes smaller with the increased number of acoustic states. JOINTCRFMFCC-LM finally achieves similar WER as DCRFMFCC-LM with 512 states. The results show that observation integration helps to compensate the potential performance loss in JOINTCRF.

6.3 Different signal-to-signal ratios

We further evaluate the speech separation algorithms under different signal-to-signal ratios. We define the ratio between the power of current source and the reference power level in the acoustic model as the gain. For speech separation problem, the gain of the sources can be estimated by algorithms such as [10] and [137]. In this experiment, we assume that the gain factor is known as a *priori*.

When the sources are modeled in log-power spectral domain, the model parame-

ters can be easily adjusted given the gain of the sources. Let $|X|^2$ be the linear power spectrum of the training data, $|\hat{X}|^2$ be the linear power spectrum of the observed data, $a^2 \in \mathbb{R}^+$ be the gain in linear power spectral domain, i.e., $|\hat{X}|^2 = a^2|X|^2$. In log-power spectral domain, we have

$$\log |\hat{X}|^2 = \log |X|^2 + \log a^2 \,. \tag{6.9}$$

If the gain is fixed, the mean parameter of the observed source is a shifted version of the original one. The covariance of the observed source is unchanged. For a state with multivariate Gaussian distributed emission probability $\mathcal{N}(x; \mu, \Sigma)$, the distribution becomes $\mathcal{N}(x; \mu + \log a^2, \Sigma)$ after the adjustment. This relationship is also applicable in log-magnitude domain since $\log |\hat{X}| = \log |X| + \log a$.

For factorial HMM, the MIXMAX model makes use of this relationship to perform speech separation under different signal-to-signal ratios. GMM modeling approach is not suitable since it requires a new model for each signal-to-signal ratio. For CRF formulations, the adjusted acoustic model parameters can be applied with state feature functions using non-linear transformation.

6.3.1 Experiments

The speech mixtures are created from two sources in 5 different signal-to-signal ratios, i.e., -6 dB, -3 dB, 0 dB, 3 dB and 6 dB. The results at 0 dB have been reported in the previous experiments. The mixing procedures are described as following. For signal-to-signal ratios greater than 0 dB, the target source is kept at the same power level as in the acoustic model, but the power of the interfering source is scaled down. Similarly, for signal-to-signal ratios smaller than 0 dB, the power of the target source is scaled down, the power of the interfering source remains unchanged.

The experiment is performed with acoustic models of 128 states. At this setting, both factorial HMM and CRF formulations have reasonable performance. CRF formulations with large-margin parameter estimation DCRFTRANS-LM and JOINTCRF-LM are compared with factorial HMM with the MIXMAX model. The separation performance of the target sources at different SSR in terms of PESQ and WER are shown in Figure 6.6 and 6.7 respectively.

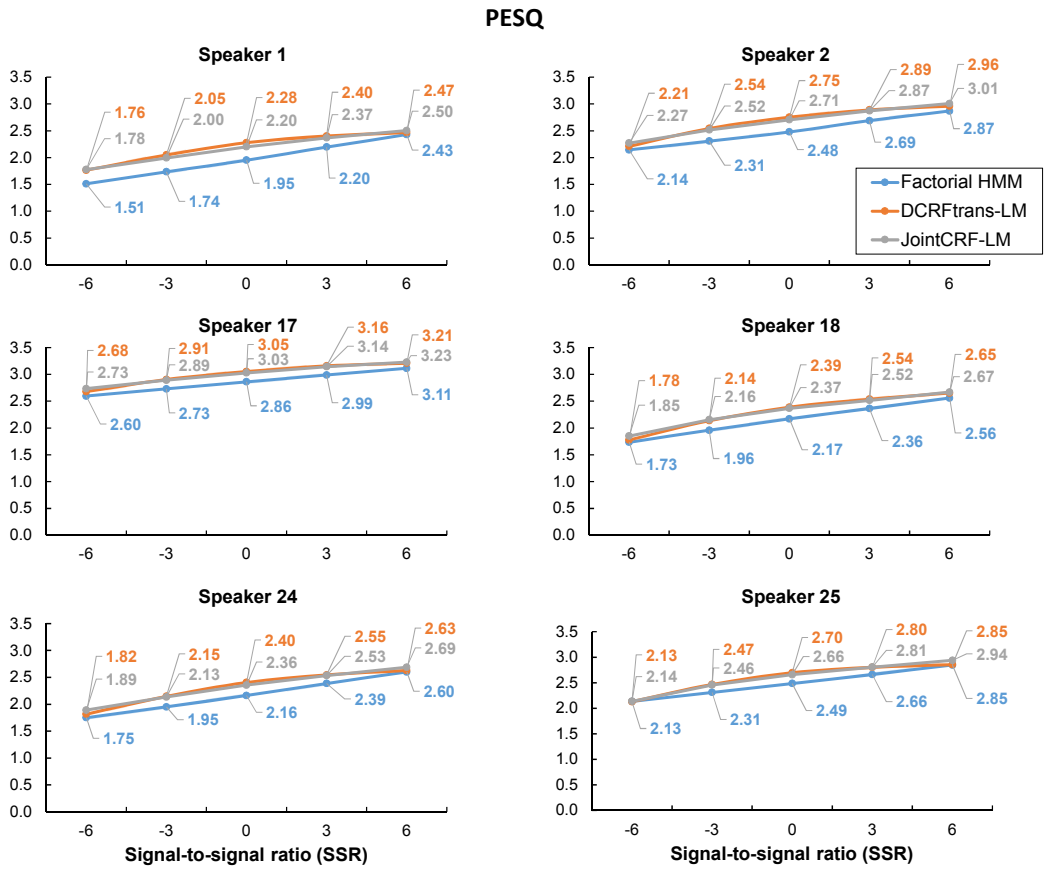


Figure 6.6: The separation results in terms of PESQ of the individual speakers in different signal-to-signal ratios

DCRFTRANS-LM and JOINTCRF-LM consistently achieve higher PESQ than factorial HMM at different SSR from -6 to 6 dB. The greatest improvement is at 0 dB SSR, which is the training condition of the current CRF parameters estimation. PESQ of DCRFTRANS-LM and JOINTCRF-LM are similar (< 0.1 PESQ difference) throughout the 5 different SSR. For factorial HMM, PESQ improves consistently and linearly with the increased SSR of the target sources. At -6 or 6 dB SSR, PESQ of factorial HMM is just slightly lower than that of CRF formulations.

When the sources are competing with similar signal power, DCRFTRANS-LM achieves lower WER than factorial HMM, and JOINTCRF-LM lies in the middle. For factorial HMM, WER decreases linearly with the increase of SSR. At -3 and 0 dB SSR, CRF formulations achieve significantly lower WER than factorial HMM. At 3 dB SSR, CRF formulations only achieve slightly better WER. When the aver-

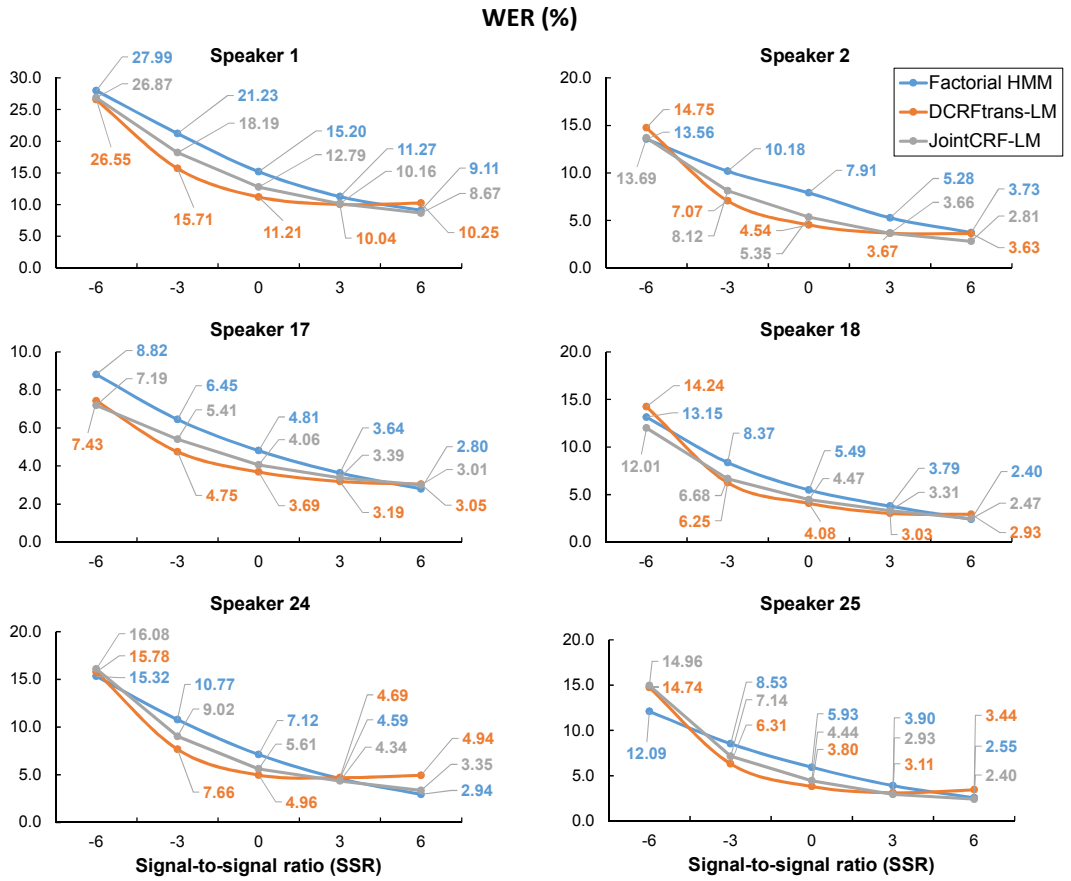


Figure 6.7: WER (%) of the separated speeches of the individual speakers in different signal-to-signal ratios

aged power of the sources are significantly different, the assumption that the speech mixture is dominated by single source is satisfied. The MIXMAX model becomes more accurate. The performance improvement of factorial HMM finally overwhelms the benefits from the discriminative ability of CRF formulations. Factorial HMM now achieves lower WER than DCRFTRANS-LM at -6 dB or 6 dB SSR.

At -6 dB or 6 dB SSR, JOINTCRF-LM generally performs better than DCRFTRANS-LM with slightly lower WER. As discussed previously, a possible reason is that the convex statistical inference procedure of JOINTCRF-LM begins to gain advantage when the mismatch of the training stage (SSR=0 dB) and the speech separation stage (SSR≠0 dB) becomes more serious.

Chapter 7

Conclusion and future directions

7.1 How far is here to the oracle?

Towards the conclusion of this work, there are several important questions. How good are the proposed CRF formulations? How much potential room is there for improvement? In order to answer these questions, an oracle experiment is carried out to determine the performance upper-bound. In this experiment, it is assumed that the source state sequences are known. They are obtained by decoding the clean reference sources with speaker-dependent acoustic models. With the oracle state sequences, the sources are recovered from speech mixtures. We compare the oracle results with the best separation results obtained in previous experiments. The best separation results are achieved mostly by CRF formulations. Figure 7.1 shows the results in terms of PESQ and WER, at 0 dB signal-to-signal ratio.

When the underlying state sequences are correct, the perceptual speech quality and speech recognition accuracy of the reconstructed sources can be greatly improved. Nevertheless, distortion still exists, as the mixture phase spectra are used for the source reconstruction. It is hard to achieve the extreme low WER in the clean sources, but a WER of 1.57% with 512 states is reasonable for many practical applications. It is observed that separation performance in terms of PESQ are similar among the three evaluation sets with 512 acoustic states. This suggests that with accurate and sufficient number of source states, the perceptual quality of the reconstructed sources tends to be independent of speakers.

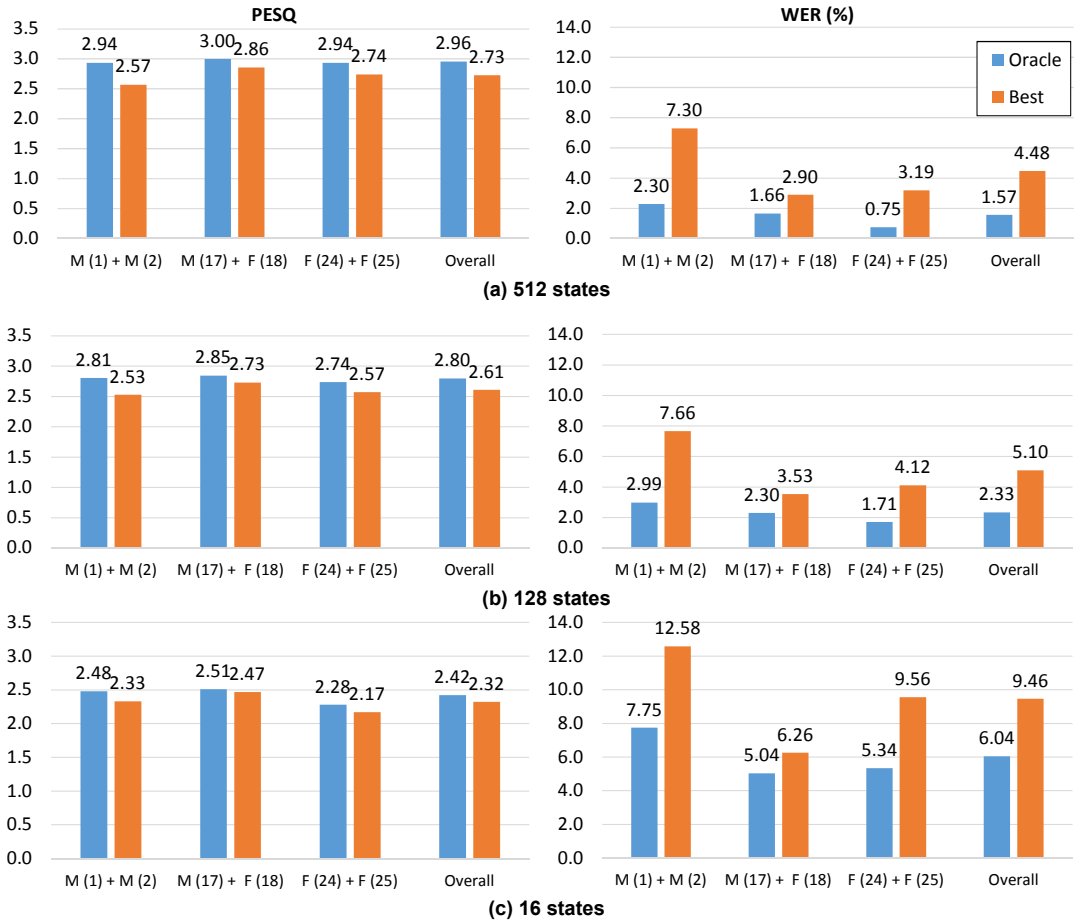


Figure 7.1: PESQ and WER (%) of the reconstructed speech sources of the oracle and from the best experimental setup

In terms of PESQ, the speech separation performance attained by the proposed methods are quite close to the oracle. The results are about 0.2 points lower than the oracle in most of the cases. For the *Male* + *Female* set with 16 acoustic states, PESQ of the best experimental results (which is by DCRFMFCC-LM) are only slightly lower than the oracle. For speech recognition task, there is still a significant performance gap between the experimental results and the oracle results. For the case with the smallest performance gap in PESQ (*Male* + *Female* with 16 states), the WER is still 24.4% higher than that of the oracle. For other cases, WER can be double or even triple of the oracle results. The baseline speech recognition system is sensitive to the distortion due to mis-classification of the source states.

7.2 Conclusion

Two types of graphical models, namely factorial hidden Markov models (HMM) and conditional random fields (CRF) are investigated for statistical model-based single-microphone speech separation. Graphical models are utilized to compute the posterior probabilities of source states given the speech mixture. The state posterior probabilities are required for MMSE source estimation.

For factorial HMM, we aim at building a comprehensive baseline system for CRF formulations. To achieve the goal, the performance of factorial HMM on single-microphone speech separation has been evaluated in detail. We have compared an analytical model from log-spectrum approximation (the MIXMAX model) and an empirical model obtained from training speech mixtures (GMM modeling approach) for modeling the state-level interaction of the sources. Experimental results show that the MIXMAX model can achieve reasonable performance on speech separation and recognition tasks in terms of PESQ and WER respectively. The GMM modeling approach gives better results when there are sufficient amount of training data. Factorial HMM with the GMM modeling approach is henceforth considered as our baseline. The results also confirm the importance of speech dynamics in single-microphone speech separation. Incorporating speech dynamics information can improve speech separation performance significantly. Applying approximated inference algorithms on factorial HMM can significantly speed up the computation at the cost of mild performance degradation. We thus conclude that the choice of inference algorithm is not critical for speech separation performance.

A generalized version of CRF, referred to as dynamic conditional random field (DCRF), has been investigated for single-microphone speech separation. A specialized graphical structure is chosen such that DCRF formulations resemble discriminative modeling of factorial HMM. DCRF and discriminative training of factorial HMM are different in the choice of sufficient statistics and the constraints for parameter estimation. DCRF formulations achieve better separation performance than factorial HMM baseline with much fewer training data. Lower speech recognition word error rate on reconstructed sources is also achieved by DCRF formulations. MFCC observations which are highly correlated with the log-spectral observations are ap-

plied in DCRF. Improvement on the separation results is observed. This demonstrates the advantage of CRF formulations in integrating multiple and correlated observations. Applying non-linear transformations on feature functions further improves the separation performance, and allows DCRF to handle speech mixtures with signal-to-signal ratios different from the training condition. Averaged stochastic gradient descent (ASGD) and loopy belief propagation are applied for approximated parameter estimation. Large-margin criterion on CRF parameter estimation further improves the separation performance by improving model generalization. In one of the evaluation set, large-margin formulation successfully relieves the over-fitting problem due to insufficient amount of training data.

JOINTCRF is proposed as a simplified CRF formulation. The forest structure of JOINTCRF promotes exact statistical inference. JOINTCRF assumes that the source states given the speech mixture are conditional independent. The assumption is sometimes violated, causing a major model mis-specification on the graphical structure. The inherent discriminative ability of CRF still leads to the comparable performance of JOINTCRF to factorial HMM. The performance of JOINTCRF is also comparable to DCRF after integrating MFCC speech mixture observations. The exact statistical inference is a convex optimization problem which leads to a unique, globally optimal solution for both parameter estimation and speech separation.

Single-microphone speech separation algorithms are evaluated under different signal-to-signal ratios. With parameters obtained at 0 dB signal-to-signal ratio, CRF formulations (DCRF and JOINTCRF) achieve better speech separation performance when the sources are competing with similar signal power. When the power difference between the target and the interference sources further increases, the assumption of the MIXMAX model is better satisfied, leads to a more accurate factorial HMM formulation. The performance of factorial HMM begins to catch up and finally outperforms CRF formulations.

7.3 Contributions

The main contributions of this work are summarized as follows.

1. Conditional random field (CRF) formulations for single-microphone speech separation are developed.

    - The application of dynamic conditional random fields (DCRF) has demonstrated the idea of discriminative modeling and integration of multiple observations for speech separation. Large-margin parameter estimation further improves the model generalization and separation performance.
    - Different feature functions for CRF formulations are evaluated. Feature functions with non-linear transformations inspired from the *mixture-maximization* (MIXMAX) model are proposed. In addition to improving separation performance, non-linear transformations allow CRF formulations to perform speech separation on mixtures with signal-to-signal ratios different from the training condition.
    - A simplified CRF formulation JOINTCRF is proposed. JOINTCRF demonstrates the power of discriminative modeling although the formulation is less accurate. The simplified graphical structure of JOINTCRF allows exact and efficient statistical inference, which is a convex problem. A unique, globally optimal solution is achievable for parameter estimation and speech separation. The performance of JOINTCRF is comparable to factorial HMM. After integrating MFCC speech mixture observations, the performance of JOINTCRF is also comparable to DCRF.

2. A factorial hidden Markov model (HMM) baseline is developed for single-microphone speech separation.

    - Exact and approximated inference algorithms are evaluated with different speaker combinations. The results show that the choice of inference algorithms is not critical to factorial HMM for speech separation. The

experimental results also show that modeling speech dynamics can improve speech separation performance significantly. The speech dynamics can be modeled as the state transition probabilities in the source acoustic models.

- An empirical distribution of source interaction is modeled in Gaussian mixture model (GMM). The empirical distribution is compared with the distribution derived from the MIXMAX model. The experimental results confirm that the MIXMAX model can achieve reasonable separation performance. The empirical distribution from the GMM modeling approach achieves better separation performance only given sufficient amount of training speech mixtures. The experimental results also reveal that the performance gap between factorial HMM and CRF formulations reduces with the increased number of source acoustic states.
- We also evaluate the separation performance of factorial HMM with the MIXMAX model on speech mixtures of different signal-to-signal ratios. We found that the separation performance improves linearly with increased signal-to-signal ratio. When the power difference of the sources further increases, the MIXMAX model becomes more accurate and factorial HMM can perform better than CRF formulations.

7.4 Future directions

The current experiments are based on speech mixtures with two competing speakers. A more realistic scenario is that the speech source is corrupted by background noise. Both factorial HMM and CRF formulations are capable to perform speech enhancement under other noisy conditions such as babble noise or speech-shape noise, but in this thesis we have not evaluated the corresponding performance. A further evaluation on these noisy conditions is thus preferred.

In terms of human perceptual speech quality, the oracle experiments in Section 7.1 reveal that reasonable performance is achieved by CRF formulations. In terms of machine intelligibility which is presented by word error rate, there is still a significant

performance gap between the best separation results and the oracle results. Moreover, an important future work is to improve CRF parameters adaptation to different signal-to-signal ratios.

The CRF formulations can be incorporated with several recent advances of machine learning techniques. The current CRF training set consists of only about 2000 training speech mixtures for each speaker pair, which is only about 1% of the available training data. Semi-supervised learning [138][110] may help to improve the performance by making use of the training data unlabeled with acoustic state sequences. Semi-supervised learning also fits to the true scenario since most of the real world speech mixtures are unlabeled.

The integration of more effective observations is always a topic of machine learning. There is no exception in single-microphone speech separation. With the gigantic improvement of the computational power, the use of deep learning is a recent trend of many pattern recognition problems [139]. A method to integrate the output of multilayer perceptron into CRF was proposed in [140]. Deep learning architecture for speech separation is already being investigated [79].

There are also rooms to further improve the perceptual quality of the reconstructed sources. Currently, the waveform of the source is reconstructed with the phase spectrum of the speech mixture. Distortion from the interfering sources are brought back to the reconstructed sources. There are researches on estimating the source phase spectrum [141][142]. An alternative approach is to perform post-processing on the reconstructed sources, such as by comb filtering [53] and periodicity enhancement [7]. Periodicity enhancement requires a robust pitch tracking algorithm. Our recent study shows that higher pitch tracking accuracy is achievable on the sources reconstructed from CRF formulations than those reconstructed by factorial HMM [6].

Appendix A

Derivation of the exact interaction model

A.1 Derivation from characteristic functions

The exact interaction model can be derived from the complex Fourier domain. After discrete Fourier transform, the instantaneous additive mixing model is $Y = \sum_k X_k$ for each frequency component, where X_k and Y are complex with a and b are the real part and the imaginary part respectively. Let $\Phi_{(\cdot)}(\widetilde{a}, \widetilde{b})$ be the characteristic functions of $p_{(\cdot)}(a, b)$. They form a double Fourier transform pair,

$$\Phi_{(\cdot)}(\widetilde{a}, \widetilde{b}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{(\cdot)}(a, b)e^{-j\widetilde{a}a}e^{-j\widetilde{b}b}\mathrm{d}a\mathrm{d}b \tag{A.1}$$

$$p_{(\cdot)}(a, b) = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Phi_{(\cdot)}(\widetilde{a}, \widetilde{b})e^{j\widetilde{a}a}e^{j\widetilde{b}b}\mathrm{d}\widetilde{a}\mathrm{d}\widetilde{b}\,. \tag{A.2}$$

Let $p_{(.)}(r)$ be the probability density function of the magnitude r. We follow the proof in [88]. The variables are firstly transformed into a polar coordinate system. By expressing

$$a + jb = re^{j\theta}$$

$$\widetilde{a} + j\widetilde{b} = qe^{j\phi}$$

and

$$a = r\cos\theta$$

$$b = r\sin\theta$$

$$r = \sqrt{a^2 + b^2}$$

$$\widetilde{a} = q\cos\phi$$

$$\widetilde{b} = q\sin\phi$$

$$q = \sqrt{\widetilde{a}^2 + \widetilde{b}^2}$$

then we obtain

$$p_{(\cdot)}(a,b) = p_{(\cdot)}(r,\theta) = \frac{1}{2\pi r} p_{(\cdot)}(r)$$

by assuming the phases are uniformly distributed. By substituting the above equations into Equation A.1 and A.2, the characteristic function $\Phi_{(\cdot)}(q)$ and the probability density function $p_{(\cdot)}(r)$ are a transform pair of zeroth-order Hankel transform,

$$\Phi_{(\cdot)}(q) = \int_0^\infty p_{(\cdot)}(r) J_0(qr)\mathrm{d}r \tag{A.3}$$

$$p_{(\cdot)}(r) = r\int_0^\infty \Phi_{(\cdot)}(q) J_0(qr) q\mathrm{d}q, \tag{A.4}$$

where $J_0(u) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-j(-u\sin(\tau))}\mathrm{d}\tau$ is the zeroth-order Bessel function of the first kind. If the random variables $\{X_k\}$ are independent, by the fundamental of random process [143], the characteristic function of Y is

$$\Phi_Y(q) = \prod_k \Phi_{X_k}(q)\,. \tag{A.5}$$

The corresponding probability density function is thus,

$$p_Y(r) = r\int_0^\infty J_0(qr)\left[\prod_k \Phi_{X_k}(q)\right] q dq\,. \tag{A.6}$$

Given that the magnitude of source X_k is $|X_k|$, $p_{X_k}(r)$ of source X_k is trivial,

$$p_{X_k}(r) = \delta(r - |X_k|) \tag{A.7}$$

where $\delta(\cdot)$ is a Dirac delta function. The corresponding characteristic function is,

$$\Phi_{X_k}(q) = J_0(|X_k|q)\,. \tag{A.8}$$

The exact interaction in magnitude domain $p\Big(|Y|\Big|\{|X_k|\}\Big) = p(\Big(r = |Y|\Big|\{|X_k|\}\Big)$ are derived from Equation A.5 and A.6,

$$p(\Big(|Y|\Big|\{|X_k|\}\Big) = |Y| \int_0^\infty J_0(|Y|q)\Bigg[\prod_k J_0(|X_k|q)\Bigg] q dq\,. \tag{A.9}$$

With the help of an integral table [144, Eq. 6.578.9], the two-source case is derived as,

$$\begin{aligned}
&p\Big(|Y|\Big||X_1|,|X_2|\Big)\\
=&|Y| \int_0^\infty J_0(|Y|q)\Bigg[J_0(|X_1|q)J_0(|X_2|q)\Bigg] q dq\\
=&\begin{cases}
\frac{2|Y|}{\pi\sqrt{4|X_1|^2|X_2|^2-(|X_1|^2+|X_2|^2-|Y|^2)^2}} & \text{for } \Big||X_1|-|X_2|\Big| < |Y| < |X_1|+|X_2| \\
0 & \text{otherwise}\,.
\end{cases}
\end{aligned} \tag{A.10}$$

By substituting $x_k = \log(|X_k|^2)$ and $y = \log(|Y|^2)$, the interaction model can be expressed in log-power domain by change of variables. As we obtain

$$\begin{aligned}
|X_k|^2 &= \exp(x_k)\\
|Y| &= \exp(\frac{y}{2})\\
g_i^{-1}(y) &= (-1)^i \exp(\frac{y}{2}) \text{ for } i = 1,2\\
\frac{\mathrm{d}g_i^{-1}(y)}{\mathrm{d}y} &= (-1)^i \frac{1}{2} \exp(\frac{y}{2}) \text{ for } i = 1,2,
\end{aligned}$$

the derived interaction model in log-power spectral domain is,

$$\begin{aligned}
&p\Big(y\Big|x_1,x_2\Big)\\
=&\sum_i \left|\frac{\mathrm{d}g_i^{-1}(y)}{\mathrm{d}y}\right| p\Big(e^y\Big|e^{x_1},e^{x_2}\Big)\\
=&\begin{cases}
\frac{e^{y-\frac{x_1+x_2}{2}}}{\pi\sqrt{1-\frac{1}{4}(e^{y-\frac{x_1+x_2}{2}}-e^{\frac{x_1-x_2}{2}}-e^{-\frac{x_1-x_2}{2}})^2}} & \text{for } \Big|e^{\frac{x_1}{2}}-e^{\frac{x_2}{2}}\Big| < e^{\frac{y}{2}} < e^{\frac{x_1}{2}}+e^{\frac{x_2}{2}} \\
0 & \text{otherwise}\,.
\end{cases}
\end{aligned} \tag{A.11}$$

A.2 Derivation from the first principle

This derivation is specifically for the two-source case. Recall Equation 3.7 for a two-source case,

$$|Y|^2 = |X_1|^2 + |X_2|^2 + 2|X_1||X_2|\cos(\theta) \tag{A.12}$$

where $\theta = \theta_1 - \theta_2$ is the phase difference. Given $|X_1|, |X_2|$ and θ, $|Y|$ can be determined analytically, hence $p(|Y|\Big||X_1|, |X_2|, \theta)$ is represented by the Dirac delta function $\delta(\cdot)$,

$$p(|Y|\Big||X_1|, |X_2|, \theta) = \delta\Big(|Y| - \sqrt{|X_1|^2 + |X_2|^2 + 2|X_1||X_2|\cos(\theta)}\Big)\,, \tag{A.13}$$

and the interaction model $p(|Y|\Big||X_1|, |X_2|)$ is obtained by marginalizing θ, i.e.,

$$\begin{aligned} p\Big(|Y|\Big||X_1|, |X_2|\Big) &= \int_{-\infty}^{\infty} p(\theta)\delta(|Y| - \sqrt{|X_1|^2 + |X_2|^2 + 2|X_1||X_2|\cos(\theta)})\mathrm{d}\theta \\ &= \sum_i \frac{p(\theta_i)}{|g'(\theta_i)|} \end{aligned} \tag{A.14}$$

where θ_i are the roots of $g(\theta) = |Y| - \sqrt{|X_1|^2 + |X_2|^2 + 2|X_1||X_2|\cos(\theta)}$. By observing that the phase of the individual sources are uniformly distributed at $[-\pi, \pi]$, the distribution of the phase difference is

$$p(\theta) = \begin{cases} \frac{1}{2\pi}(1 - |\frac{\theta}{2\pi}|) & \text{for } |\theta| \leq 2\pi \\ 0 & \text{otherwise} \end{cases} \tag{A.15}$$

and for $g(\theta_i) = 0$, $g'(\theta_i) = \frac{\sqrt{4|X_1|^2|X_2|^2 - (|X_1|^2 + |X_2|^2 - |Y|^2)^2}}{2|Y|}$. Since

$$\theta_i = \begin{cases} -\cos^{-1}\left|\frac{|Y|^2 - (|X_1|^2 + |X_2|^2)}{2|X_1||X_2|}\right| + (i-1)\pi & \text{for } i \text{ is odd} \\ \cos^{-1}\left|\frac{|Y|^2 - (|X_1|^2 + |X_2|^2)}{2|X_1||X_2|}\right| - (i-2)\pi & \text{for } i \text{ is even,} \end{cases}$$

and there are four roots at $[-2\pi, 2\pi]$, $\sum_i p(\theta_i) = \frac{1}{\pi}$. Hence

$$p\Big(|Y|\Big||X_1|, |X_2|\Big) = \frac{2|Y|}{\pi\sqrt{4|X_1|^2|X_2|^2 - (|X_1|^2 + |X_2|^2 - |Y|^2)^2}} \quad . \tag{A.16}$$

Similarly for log-power spectral domain, $g(\theta) = y - \log(e^{x_1} + e^{x_2} + 2e^{\frac{x_1+x_2}{2}}\cos(\theta))$ and $g'(\theta_i) = \frac{\sqrt{1 - \frac{1}{4}(e^{y - \frac{x_1+x_2}{2}} - e^{\frac{x_1-x_2}{2}} - e^{-\frac{x_1-x_2}{2}})^2}}{e^{y - \frac{x_1+x_2}{2}}}$, the interaction model is the same as Equation A.11.

Appendix B

Proof of forward-backward updates

B.1 Forward-backward algorithm on factorial HMM

A junction tree can be constructed from an undirected graph or the moral graph of a directed graph if and only if the graph is triangulated [145]. Figure B.1 and Figure B.2 shows the moral graph and the chordal graph after triangulation of factorial HMM respectively.

Let X, Y be the clique nodes (denoted in circles) in the junction tree and M be a separator (denoted in rectangles) as illustrated in Figure B.4. The separator M contains a set of random variables s_M which are the intersection of the random variable sets s_X and s_Y of X and Y, i.e. $s_M = s_X \cap s_Y$. Let $\Phi_X(s_X)$ be the potential function of the clique node X, $m(s_M)$ be the potential function of the separator M. An asterisk "$*$" is added to the potential function at each update. Notably if we apply the update as in forward-backward algorithm, $m^*(\cdot)$ is a forward message and $m^{**}(\cdot)$ is a backward message.

The junction tree of a factorial HMM with two sources is shown in Figure B.3. By setting the initial condition of $m(\cdot) = 1$, the forward-backward algorithm of factorial

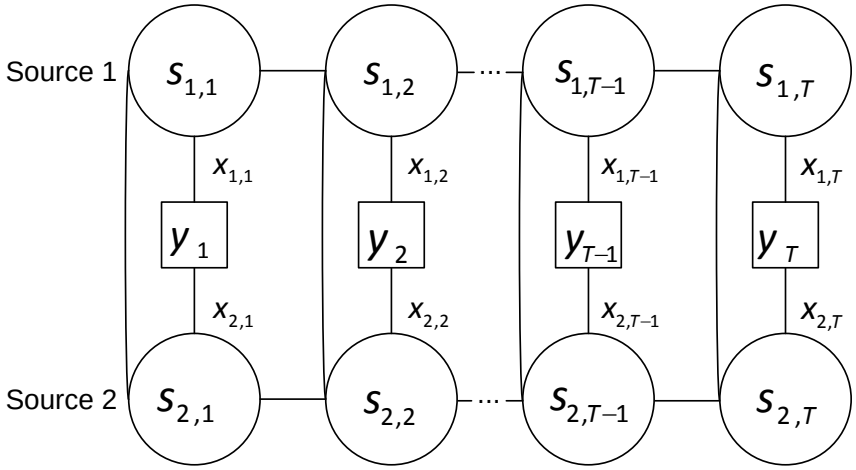


Figure B.1: The factorial HMM in Figure 4.5 after moralization

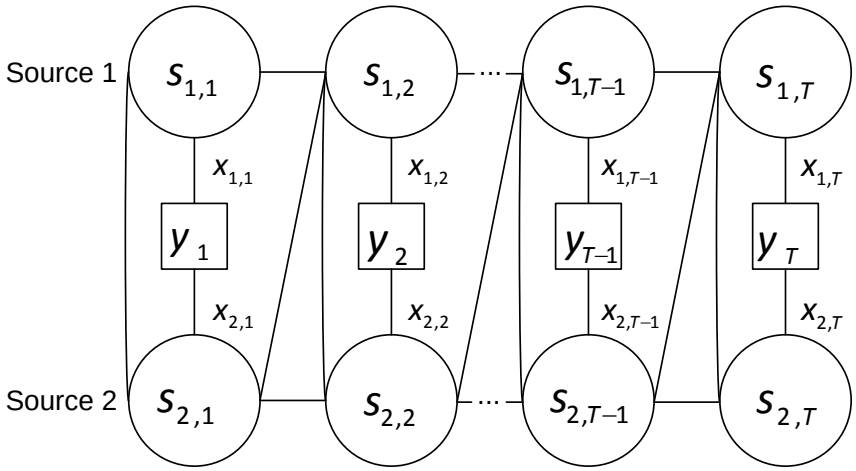


Figure B.2: The factorial HMM in Figure 4.5 after moralization and triangulation

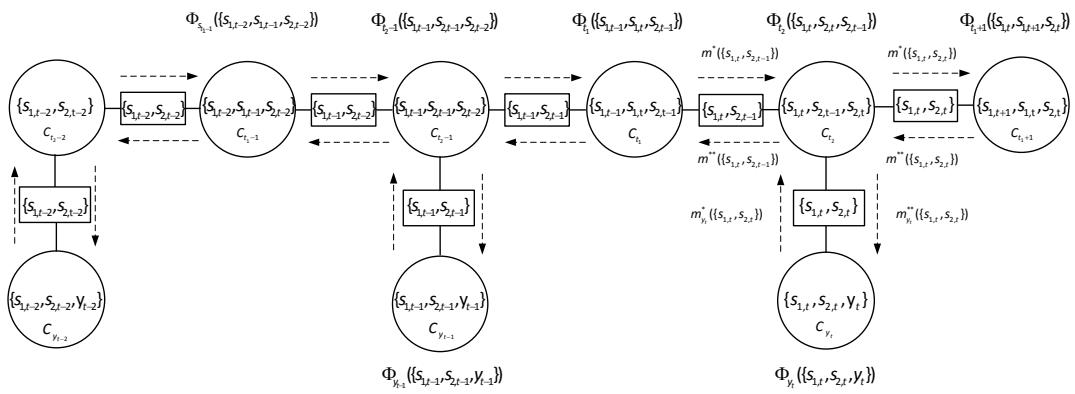


Figure B.3: The junction tree of factorial HMM in Figure 4.5.

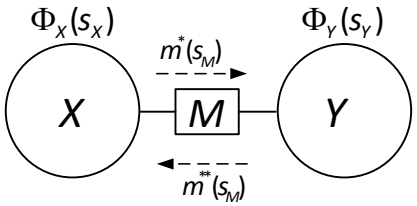


Figure B.4: A message is passing from clique X to clique Y in a junction tree

HMM can be derived from the junction tree update. By setting

$$\begin{aligned}
\Phi_t(s_{1,1}, s_{2,1}) &= p(s_1)p(s_2) \\
\Phi_{y_t}(y_{1,t}, s_{1,t}, s_{2,t}) &= p(y_{1,t}|s_{1,t}, s_{2,t}) \\
\Phi_t(s_{1,t}, s_{2,t}, s_{2,t-1}) &= p(s_{2,t}|s_{2,t-1}) \\
\Phi_t(s_{1,t-1}, s_{1,t}, s_{2,t-1}) &= p(s_{1,t}|s_{1,t-1}) \\
m^*_{y_t}(s_{1,t}, s_{2,t}) &= \Phi_{y_t}(y_{1,t}, s_{1,t}, s_{2,t})
\end{aligned}$$

and applying dynamic programming on the junction tree, the update rules for the forward messages $\alpha^\star_t$ and α_t are derived as,

$$\begin{aligned}
\alpha_t &= m^*(s_{1,t}, s_{2,t}) \\
&= \sum_{s_{2,t-1}} m^*(s_{1,t}, s_{2,t-1})\Phi_{y_t}(y_{1,t}, s_{1,t}, s_{2,t})\Phi_t(s_{1,t}, s_{2,t}, s_{2,t-1}) \\
&= p(y_{1,t}|s_{1,t}, s_{2,t}) \sum_{s_{2,t-1}} \underbrace{m^*(s_{1,t}, s_{2,t-1})}_{\alpha^\star_t} p(s_{2,t}|s_{2,t-1}) \\
&= p(y_{1,t}|s_{1,t}, s_{2,t}) \sum_{s_{2,t-1}} p(s_{2,t}|s_{2,t-1})\alpha^\star_t \blacksquare, \qquad \text{(B.1)} \\
\alpha^\star_t &= m^*(s_{1,t}, s_{2,t-1}) \\
&= \sum_{s_{1,t-1}} \underbrace{m^*(s_{1,t-1}, s_{2,t-1})}_{\alpha_{t-1}} \Phi_t(s_{1,t-1}, s_{1,t}, s_{2,t-1}) \\
&= \sum_{s_{1,t-1}} \alpha_{t-1} p(s_{1,t}|s_{1,t-1}) \blacksquare. \qquad \text{(B.2)}
\end{aligned}$$

The update rules for the backward messages $\beta^\star_t(s_t)$ and $\beta_t(s_t)$ are derived as,

$$
\begin{aligned}
\beta_t &= m^{**}(s_{1,t}, s_{2,t}) \\
&= \sum_{s_{1,t+1}} \underbrace{m^{**}(s_{1,t+1}, s_{2,t})}_{\beta_t^\star} \Phi_t(s_{1,t}, s_{1,t+1}, s_{2,t}) \\
&= \sum_{s_{1,t+1}} \beta_t^\star p(s_{1,t+1}|s_{1,t}) \blacksquare, \qquad \text{(B.3)} \\
\beta_t^\star &= \sum_{s_{2,t+1}} \Phi_t(s_{1,t+1}, s_{2,t+1}, s_{2,t}) \Phi_{y_t}(y_{1,t+1}, s_{1,t+1}, s_{2,t+1}) \underbrace{m^{**}(s_{1,t+1}, s_{2,t+1})}_{\beta_{t+1}} \\
&= \sum_{s_{2,t+1}} p(s_{2,t+1}|s_{2,t}) p(y_{t+1}|s_{1,t+1}, s_{2,t+1}) \beta_{t+1} \blacksquare. \qquad \text{(B.4)}
\end{aligned}
$$

The joint probability hence is derived as

$$
\begin{aligned}
p(s_{1,t}, s_{2,t}, \mathbf{y}) &= \Phi_{y_t}(y_{1,t}, s_{1,t}, s_{2,t}) \sum_{s_{1,t-1}} m^*(s_{1,t}, s_{2,t-1}) m^{**}(s_{1,t}, s_{2,t}) \Phi_t(s_{1,t}, s_{2,t}, s_{2,t-1}) \\
&= \beta_t \underbrace{p(y_{1,t}|s_{1,t}, s_{2,t}) \sum_{s_{2,t-1}} p(s_{2,t}|s_{2,t-1}) \alpha_t^\star}_{\alpha_t} \\
&= \alpha_t \beta_t \blacksquare. \qquad \text{(B.5)}
\end{aligned}
$$

Bibliography

[1] C. H. Hansen, "Fundamentals of acoustics," in *Occupational exposure to noise: evaluation, prevention and control*, B. Goelzer, C. H. Hansen, and G. A. Sehrndt, Eds. Dortmund, Germany: World Health Organization, 2001.

[2] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 12, pp. 1846–1856, 1989.

[3] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *Signal Processing, IEEE Transactions on*, vol. 40, no. 4, pp. 725–735, 1992.

[4] K. Han and D. L. Wang, "Towards generalizing classification based speech separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 1, pp. 168–177, 2013.

[5] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech & Language*, vol. 24, no. 1, pp. 1–15, 2010.

[6] F. Huang, Y. T. Yeung, and T. Lee, "Evaluation of pitch estimation algorithms on separated speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 6807–6811.

[7] F. Huang, "Speech periodicity enhancement based on transform-domain signal decomposition and robust pitch estimation," Ph.D. dissertation, The Chinese University of Hong Kong, 2012.

[8] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[9] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems*. MIT Press, 2001, pp. 793–799.

[10] 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 & Language*, vol. 24, no. 1, pp. 45–66, 2010.

[11] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Single-channel multitalker speech recognition," *Signal Processing Magazine, IEEE*, vol. 27, no. 6, pp. 66–80, 2010.

[12] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Communication*, vol. 49, no. 5, pp. 402–417, May 2007.

[13] P. W. Dawson, S. J. Mauger, and A. A. Hersbach, "Clinical evaluation of signal-to-noise ratio-based noise reduction in Nucleus® cochlear implant recipients," *Ear and hearing*, vol. 32, no. 3, pp. 382–90, 2011.

[14] A. A. Hersbach, K. Arora, S. J. Mauger, and P. W. Dawson, "Combining directional microphone and single-channel noise reduction algorithms: A clinical evaluation in difficult listening conditions with cochlear implant users," *Ear and hearing*, vol. 33, no. 4, pp. e13–23, 2012.

[15] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 282–289.

[16] E. Fosler-Lussier, Y. He, P. Jyothi, and R. Prabhavalkar, "Conditional random fields in speech, audio, and language processing," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1054–1075, May 2013.

[17] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification," in *9th European Conference on Speech Communication and Technology, Interspeech 2005*, 2005, pp. 1117–1120.

[18] J. Morris and E. Fosler-Lussier, "Conditional random fields for integrating local discriminative classifiers," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 3, pp. 617–628, 2008.

[19] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, ser. NAACL '03, vol. 1. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 134–141.

[20] R. Prabhavalkar, Z. Jin, and E. Fosler-Lussier, "Monaural segregation of voiced speech using discriminative random fields," in *Tenth Annual Conference of the International Speech Communication Association, Interspeech 2009*, vol. 1, 2009, pp. 856–859.

[21] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.

[22] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1998–2000, 1999.

[23] J. Beh and H. Ko, "A novel spectral subtraction scheme for robust speech recognition: Spectral subtraction using spectral harmonics of speech," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP). 2003 IEEE International Conference on*, vol. 1, 2003, pp. 648–651 vol.1.

[24] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Communication*, vol. 50, no. 6, pp. 453–466, 2008.

[25] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462–1469, 2006.

[26] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 4, 2002, pp. 4164–4167.

[27] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[28] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Communication*, vol. 10, no. 1, pp. 45–57, 1991.

[29] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[30] G.-J. Jang and T.-W. Lee, "A maximum likelihood approach to single-channel source separation," *J. Mach. Learn. Res.*, vol. 4, pp. 1365–1392, 2003.

[31] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[32] J. F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[33] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, no. 6, pp. 1129–1159, Jan. 1995.

[34] D. Smith, J. Lukasiak, and I. S. Burnett, "An analysis of the limitations of blind signal separation application with speech," *Signal Processing*, vol. 86, no. 2, pp. 353–359, 2006.

[35] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[36] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Ninth International Conference on Spoken Language Processing*, 2006, pp. 2614–2617.

[37] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1066 –1074, March 2007.

[38] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: From coding to source separation," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 995–1005, 2010.

[39] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3397–3415, 1993.

[40] P. J. Durka, D. Ircha, and K. J. Blinowska, "Stochastic time-frequency dictionaries for matching pursuit," *Signal Processing, IEEE Transactions on*, vol. 49, no. 3, pp. 507–510, 2001.

[41] S. P. Boyd and L. Vandenberghe, *Convex optimization*. New York: Cambridge, 2004.

[42] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, Oct. 1994.

[43] G. J. Brown and D. L. Wang, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Hoboken, N.J.: Wiley-Interscience, 2006.

[44] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*. Springer US, 2005, pp. 181–197.

[45] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, p. 4007, 2006.

[46] M. Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, p. 1562, 2006.

[47] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction." *The Journal of the Acoustical Society of America*, vol. 123, no. 3, pp. 1673–82, Mar. 2008.

[48] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[49] K. Han and D. Wang, "An SVM based classification approach to speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 4632–4635.

[50] F. R. Bach and M. I. Jordan, "Learning spectral clustering, with application to speech separation," *J. Mach. Learn. Res.*, vol. 7, pp. 1963–2001, Dec. 2006.

[51] W. Hartmann and E. Fosler-Lussier, "ASR-driven top-down binary mask estimation using spectral priors," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4685–4688.

[52] M. Weintraub, "A computational model for separating two simultaneous talkers," *Acoustics, Speech, and Signal Processing, 1986. (ICASSP '86). IEEE International Conference on (Volume:11)*, pp. 81–84, 1986.

[53] S. W. Lee, "Model-based speech separation and enhancement with single-microphone input," Ph.D. dissertation, The Chinese University of Hong Kong, 2008.

[54] Y. Gu and W. Van Bokhoven, "Co-channel speech separation using frequency bin non-linear adaptive filtering," in *Acoustics, Speech, and Signal Processing,*

1991. Proceedings. (ICASSP-91). IEEE International Conference on, 1991, pp. 949–952 vol.2.

[55] M. Wu, D. L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 3, pp. 229–241, 2003.

[56] F. Bach and M. I. Jordan, "Discriminative training of hidden Markov models for multiple pitch tracking," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 5, 2005, pp. 489–492 Vol. 5.

[57] M. Wohlmayr and F. Pernkopf, "Multipitch tracking using a factorial hidden Markov model," in *Ninth Annual Conference of the International Speech Communication Association, Interspeech 2008*, Brisbane, 2008, pp. 147–150.

[58] M. Stark, M. Wohlmayr, and F. Pernkopf, "Source-filter-based single-channel speech separation using pitch information," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 2, pp. 242–255, 2011.

[59] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 47–56, Jan. 2011.

[60] International Telecommunication Union, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, 2001, ITU-T Recommendation P.862.

[61] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech, and Signal Processing, 2001. (ICASSP '01). IEEE International Conference on*, vol. 2, 2001, pp. 749–752.

[62] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[63] P. Mowlaee, R. Saeidi, M. G. Christensen, and R. Martin, "Subjective and objective quality assessment of single-channel speech separation algorithms," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 69–72.

[64] P. Loizou, *Speech enhancement: Theory and Practice*, ser. Signal processing and communications. CRC Press, 2007.

[65] J. Ma and P. C. Loizou, "SNR Loss: A new objective measure for predicting speech intelligibility of noise-suppressed speech." *Speech communication*, vol. 53, no. 3, pp. 340–354, Mar. 2011.

[66] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions." *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–405, May 2009.

[67] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[68] International Telecommunication Union, *Methods for subjective determination of transmission quality*, 1996, ITU-T Recommendation P.800.

[69] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge University, 1995.

[70] N. Merhav and Y. Ephraim, "Maximum likelihood hidden Markov modeling using a dominant sequence," *Signal Processing, IEEE Transactions on*, vol. 39, no. 9, pp. 2111–2115, 1991.

[71] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series, with Engineering Applications*. Cambridge, MA: Technology Press of the Massachusetts Institute of Technology, 1949.

[72] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME - Journal of Basic Engineering*, no. 82 (Series D), pp. 35–45, 1960.

[73] Y. Ephraim, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[74] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, no. 2, pp. 245–273, 1997.

[75] B. Frey, L. Deng, A. Acero, and T. Kristjansson, "ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition," in *Proc. Eurospeech*, vol. 2, 2001, pp. 901–904.

[76] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, 1990, pp. 845–848 vol.2.

[77] A. Nadas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 10, pp. 1495–1503, 1989.

[78] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 6, pp. 341–351, 2002.

[79] S. Rennie, P. Fousek, and P. Dognin, "Factorial hidden restricted Boltzmann machines for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4297–4300.

[80] Y. T. Yeung, T. Lee, and C.-C. Leung, "Integrating multiple observations for model-based single-microphone speech separation with conditional random fields," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 257–260.

[81] ——, "Using dynamic conditional random field on single-microphone speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 146–150.

[82] J. Carmona, J. Barker, A. Gomez, and N. Ma, "Speech spectral envelope enhancement by hmm-based analysis/resynthesis," *Signal Processing Letters, IEEE*, vol. 20, no. 6, pp. 563–566, 2013.

[83] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, Jun. 2001.

[84] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Variational loopy belief propagation for multi-talker speech recognition," in *Tenth Annual Conference of the International Speech Communication Association, Interspeech 2009*, 2009, pp. 1331–1334.

[85] P. Moreno, B. Raj, and R. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. Conference Proceedings. (ICASSP-96). IEEE International Conference on*, vol. 2, 1996, pp. 733–736 vol. 2.

[86] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," pp. 352–359, 1996.

[87] M. H. Radfar, A. H. Banihashemi, R. M. Dansereau, and A. Sayadiyan, "Nonlinear minimum mean square error estimator for mixture-maximisation approximation," *Electronics Letters*, vol. 42, no. 12, p. 724, 2006.

[88] G. D. Durgin, T. S. Rappaport, and D. A. De Wolf, "New analytical models and probability density functions for fading in wireless communications," *Communications, IEEE Transactions on*, vol. 50, no. 6, pp. 1005–1015, 2002.

[89] J. R. Hershey, P. A. Olsen, and S. J. Rennie, "Signal interaction and the devil function," in *Eleventh Annual Conference of the International Speech Communication Association, Interspeech 2010*, 2010, pp. 334–337.

[90] L. Deng, J. Droppo, and A. Acero, "Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 133–143, Mar. 2004.

[91] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *Signal Processing, IEEE Transactions on*, vol. 52, no. 7, pp. 1830–1847, 2004.

[92] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2299–2310, 2007.

[93] C.-M. Mak, T. Lee, and S. W. Lee, "Spectral trajectory estimation using non-negative matrix factorization for model-based monaural speech separation," in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, Nov. 29 2010-Dec. 3 2010 2010, pp. 23 –28.

[94] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 625–638, May 2009.

[95] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[96] G. Ye, D. Chen, and B. Mak, "Transition probabilities are more important than we once thought," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 4809 –4812.

[97] F. V. Jensen, *An Introduction to Bayesian Networks*. London: UCL Press, 1996.

[98] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[99] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Single-channel speech separation and recognition using loopy belief propagation," in *Acoustics, Speech and Signal Processing (ICASSP), 2009 IEEE International Conference on*, 2009, pp. 3845–3848.

[100] S. L. Lauritzen and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 50, no. 2, pp. 157–224, Jan. 1988.

[101] R. Diestel, *Graph Theory, 4th Edition*, ser. Graduate texts in mathematics. Springer, 2012, vol. 173.

[102] F. V. Jensen and F. Jensen, "Optimal junction trees," in *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, 1994, pp. 360–366.

[103] K. P. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief propagation for approximate inference: An empirical study," in *Proceedings of Uncertainty in AI*, vol. 9, 1999, pp. 467–475.

[104] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Generalized belief propagation," in *Advances in Neural Information Processing Systems*. MIT Press, 2001, pp. 689–695.

[105] M. I. Jordan, Z. Ghahramani, and T. S. Jaakkola, "An introduction to variational methods for graphical models," *Machine learning*, vol. 233, pp. 183–233, 1999.

[106] a. Nadas, D. Nahamoo, and M. Picheny, "On a model-robust training method for speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 9, pp. 1432–1436, 1988.

[107] P. Liang and M. I. Jordan, "An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators," *Proceedings of the 25th international conference on Machine learning - ICML '08*, pp. 584–591, 2008.

[108] A. Nadas, "A decision theorectic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood," pp. 814–817, 1983.

[109] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," *Advances in neural information processing systems 14*, pp. 841–848, 2002.

[110] J. Lasserre, C. Bishop, and T. Minka, "Principled hybrids of generative and discriminative models," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, 2006, pp. 87–94.

[111] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Transactions on Information Theory*, vol. 26, no. 1, pp. 26–37, Jan. 1980.

[112] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 939–952, 1982.

[113] C. Sutton, A. Mccallum, and K. Rohanimanesh, "Dynamic conditional random fields : Factorized probabilistic models for labeling and segmenting sequence data," *Journal of Machine Learning Research*, vol. 8, pp. 693–723, 2007.

[114] G. Heigold, H. Ney, P. Lehnen, T. Gass, and R. Schluter, "Equivalence of generative and log-linear models," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1138–1148, 2011.

[115] L. Bahl, P. Brown, P. De Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Acoustics, Speech, and Signal Processing, 1986. (ICASSP '86). IEEE International Conference on*, vol. 11, 1986, pp. 49–52.

[116] V. Valtchev, J. J. Odell, P. C. Woodland, and S. J. Young, "MMIE training of large vocabulary recognition systems," *Speech Communication*, vol. 22, no. 4, pp. 303–314, 1997.

[117] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 25–47, 2002.

[118] I. Jolliffe, "Principal component analysis," in *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd, 2005.

[119] H. A. Bethe, "Statistical theory of superlattices," *Proceedings of the Royal Society of London. Series A,*, vol. 150, no. 871, pp. 552–575, 1935.

[120] J. Nocedal, "Updating quasi-Newton matrices with limited storage," *Mathematics of Computation*, vol. 35, no. 151, pp. 773–782, 1980.

[121] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM Journal on Control and Optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[122] L. Bottou, "Online algorithms and stochastic approximations," in *Online Learning and Neural Networks*, D. Saad, Ed. Cambridge, UK: Cambridge University Press, 1998.

[123] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy, "Accelerated training of conditional random fields with stochastic gradient methods," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 969–976.

[124] W. Xu, "Towards optimal one pass large scale learning with averaged stochastic gradient descent," *ArXiv e-prints*, Jul. 2011.

[125] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[126] F. Sha and L. Saul, "Large margin Gaussian mixture modeling for phonetic classification and recognition," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, 2006, pp. I–I.

[127] ——, "Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, 2007, pp. IV–313–IV–316.

[128] L. Xiao and L. Deng, "A geometric perspective of large-margin training of Gaussian Models," *Signal Processing Magazine, IEEE*, vol. 27, no. 6, pp. 118–123, 2010.

[129] K. Q. Weinberger, F. Sha, and L. K. Saul, "Convex optimizations for distance metric learning and pattern classification," *Signal Processing Magazine, IEEE*, vol. 27, no. 3, pp. 146–158, 2010.

[130] Y. Freund, "Large margin classification using the perceptron algorithm," *Machine Learning*, vol. 296, pp. 277–296, 1999.

[131] B. Taskar, C. Guestrin, and D. Koller, "Max-margin Markov networks," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004.

[132] F. Sha and L. K. Saul, "Large margin hidden Markov models for automatic speech recognition," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 1249–1256.

[133] M. Kim, "Large margin cost-sensitive learning of conditional random fields," *Pattern Recognition*, vol. 43, no. 10, pp. 3683–3692, Oct. 2010.

[134] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 4057–4060.

[135] G. Heigold, T. Deselaers, R. Schlüter, and H. Ney, "Modified MMI/MPE: A direct evaluation of the margin in speech recognition," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: ACM, 2008, pp. 384–391.

[136] M. J. Wainwright, "Estimating the "wrong" graphical model: Benefits in the computation-limited setting," *Journal of Machine Learning Research*, vol. 7, pp. 1829–1859, Dec. 2006.

[137] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 191–199, 2006.

[138] F. Jiao, S. Wang, C.-H. Lee, R. Greiner, and D. Schuurmans, "Semi-supervised conditional random fields for improved sequence segmentation and labeling," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ser. ACL-44. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 209–216.

[139] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, Jan. 2009.

[140] R. Prabhavalkar and E. Fosler-Lussier, "Backpropagation training for multilayer conditional random field based phone recognition," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 5534–5537.

[141] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 421–424, May 2010.

[142] N. Sturmel and L. Daudet, "Iterative phase reconstruction of Wiener filtered signals," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 101–104.

[143] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 3rd ed. New York: McGraw-Hill Companies, Feb. 1991.

[144] I. S. Gradshteyn, A. Jeffrey, and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 5th ed. Boston: Academic Press, 1994.

[145] S. L. Lauritzen, *Graphical Models*, ser. Oxford Science Publications. Oxford University Press, 1996.