SINGLE-CHANNEL SPEECH-MUSIC SEPARATION FOR ROBUST ASR WITH MIXTURE OF NMF MODELS

by

Cemil Demir

B.S., Electrical and Electronics Engineering, Bilkent University, 2004

M.S., Electrical and Electronics Engineering, Boğaziçi University, 2007

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Electrical and Electronics Engineering
Boğaziçi University
2014

to my wife

ACKNOWLEDGEMENTS

I am very grateful to my thesis supervisors Murat Saraçlar and Ali Taylan Cemgil for their invaluable guidance. Without their help, this thesis could not be completed.

I would like to thank Hakan Erdoğan and Levent Arslan for being members of the thesis progress committee and the defense jury. Their precious feedback contributed a lot to the work in this dissertation. I would like to thank Tuomas Virtanen for traveling from Tampere to participate in my jury.

I also would like to thank the members of TÜBİTAK-BİLGEM Speech and Language Processing Laboratory for their support to finish my thesis. Special thanks to M. Uğur Doğan for his endless encouragement to complete the thesis.

Last but not least I would like to mention my parents, for their unconditional support. Finally, my deepest thanks go to my wife, Filiz and my lovely daughter Cemile Neda for their love and their enduring support. Without their patience and encouragement, this thesis would never have been written.

ABSTRACT

SINGLE-CHANNEL SPEECH-MUSIC SEPARATION FOR ROBUST ASR WITH MIXTURE OF NMF MODELS

In this dissertation, we analyze the single-channel speech-music separation problem for automatic speech recognition (ASR). The motivation of the study is to increase the performance of the ASR systems by decreasing the effect of background music. We describe a single-channel speech-music separation method based on a mixture of non-negative matrix factorization (NMF) model. Given a catalog of background music material, we propose a generative model for the superposed speech and music spectrograms. The background music signal is assumed to be generated by a jingle in the catalog and it is modeled by a scaled conditional mixture model representing the jingle. The speech signal is modeled by an NMF model that is estimated in a semi-supervised manner from the mixed signal. The approach is tested with Poisson and complex Gaussian observation models that correspond respectively to Kullback-Leibler (KL) and Itakura-Saito (IS) divergence measures. Our experiments show that the proposed mixture model outperforms a standard NMF method both in speech-music separation and automatic speech recognition (ASR) tasks. Moreover, we extend the mixture of NMF based single-channel speech-music separation method such that it incorporates prior speech information to enhance the separation performance of the method. Finally, we propose to use sub-word NMF-based speech models for the separation of speech and music signals. By applying such a strategy, it is demonstrated that the recognition accuracy can be improved as compared to using a general speech model.

ÖZET

GÜRBÜZ KONUŞMA TANIMA İÇİN NOMA KARIŞIM MODELLERİYLE TEK-KANALDA KONUŞMA-MÜZİK AYRIŞTIRMA

Bu çalışmada otomatik konuşma tanıma (OKT) için tek kanalda konuşma-müzik ayrıştırma problemini inceledik. Çalışmanın motivasyonu, tanıma hatalarını arttıran arka-plan müziğinin etkisini azaltarak konuşma tanıma başarımını arttırmaktır. Bu çalışmada tek kanalda konuşma-müzik ayrıştırma metodu olarak Negatif Olmayan Matris Ayrıştırma (NOMA) karışımı modeli tabanlı bir yöntem tanımlanmıştır. Arka-plan müziklerini içeren bir katalog verildiği ve müziğin katalogdaki bir cıngıl tarafından üretildiği varsayımı altında karma konuşma ve müzik spektogramları için bir üretici model önerilmiştir. Önerilen yöntemde konuşma sinyali karma sinyalden yarı güdümlü biçimde kestirilen bir NOMA modeli ile temsil edilmektedir. Bu yöntem sırası ile Kullback-Leibler (KL) ve Itakura-Saito (IS) ıraksay ölçütlerine karşılık düşen Poisson ve karmaşık Gauss gözlem modelleri ile test edilmiştir. Deneylerimize göre önerilen karışım modeli hem konuşma-müzik ayrıştırma hem de konuşma tanıma testlerinde standart NOMA modellerinden daha iyi sonuçlar vermektedir. Daha sonra, önerilen NOMA karışım tabanlı yöntemin ayrıştırma başarımını iyileştirmek için önerilen olasılıksal model ve yöntem konuşma sinyali hakkındaki önsel bilgiyi kullanacak şekilde geliştirilmiştir. Son olarak, konuşma-müzik ayrıştırma için NOMA tabanlı kelime altı konuşma modellerinin kullanılması önerilmiştir. Bu stratejinin genel bir konuşma modeline kıyasla daha iyi bir konuşma tanıma başarımı sağladığı gösterilmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

b	Template vector index
f	Frequency bin index
h	Frequency filtering parameter
m	Latent music source
$\hat{m}$	Separated music signal
q	Posterior distribution of the latent sources
r	Active jingle index parameter
$\hat{s}$	Separated speech signal
s	Latent speech source
r_t	Active jingle index for time frame t
t	Time frame index
v	Gain parameter
$\mathbf{C}$	Magnitude or power Spectrum of Jingle
$\mathbf{D}$	Template or Dictionary Matrix of a Non-negative Matrix Factorization Model
$\mathbf{E}$	Excitation Matrix of a Non-negative Matrix Factorization Model
F	Total number of frequency bins
$\mathcal{M}$	Multinomial Distribution
$\mathcal{N}_c$	Complex Gaussian Distribution
$\mathbf{M}$	Magnitude or complex Spectrum of Music Signal
$\mathcal{PO}$	Poisson Distribution
Q	Expectation of the joint log-likelihood
$\mathbf{P}$	Power spectrum of a signal
$\mathbf{S}$	Magnitude or complex spectrum of speech signal
T	Total number of frames
$\mathbf{X}$	Magnitude or complex spectrum of mixed signal
λ	Intensity parameter of Poisson Distribution

μ	Mean parameter of complex Gaussian Distribution
Σ	Variance parameter of complex Gaussian Distribution
Θ	Parameter set

LIST OF ACRONYMS/ABBREVIATIONS

AR	Auto Regressive
ASR	Automatic Speech Recognition
BSS	Blind Source Separation
BW	Baum-Welch
CF	Cost Function
CGMM	Complex Gaussian Mixture Model
CNMF	Convolutional Non-negative Matrix Factorization
EM	Expectation Maximization
FHMM	Factorial Hidden Markov Model
GMC	Gamma Markov Chain
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IGMC	Inverse-Gamma Markov Chain
ICA	Independent Component Analysis
IS	Itakura-Saito
KL	Kullback-Leibler
MAP	Maximum A Posterior Probability
MCMC	Markov Chain Monte Carlo
MFCC	Mel Frequency Cepstral Coefficients
ML	Maximum Likelihood
MLLR	Maximum Likelihood Linear Regression
MMM	Multinomial Mixture Model
MMSE	Minimum Mean Square Error
MMSE-SPU	Minimum Mean Square Error Speech Presence Uncertainty
MPP	Maximum Posterior Probability
MSR	Music to Speech Ratio
NMF	Non-negative Matrix Factorization
PCA	Principal Component Analysis
PCE	Piece-wise Constant Estimation

PLP	Perceptual Linear Prediction
PMM	Poisson Mixture Model
PNCC	Power Normalized Cepstral Coefficients
PSD	Power Spectral Density
SAR	Speech to Artifact Ratio
SCNMF	Sparse Convolutional Non-negative Matrix Factorization
SDR	Speech to Distortion Ratio
SIR	Signal to Interference Ratio
SMR	Speech to Music Ratio
SNMF	Sparse Non-negative Matrix Factorization
SNR	Signal to Noise Ratio
STFT	Short Time Fourier Transform
VQ	Vector Quantization
WAcc	Word Accuracy Rate
WER	Word Error Rate

1. INTRODUCTION

Recently automatic speech recognition (ASR) applications have become popular in broadcast news transcription systems. One major problem is the serious drop in the performance with the presence of background music that is often present in radio and television broadcasts [1, 2]. The effect of background music in ASR application can be seen from the Table 1.1. In Table 1.1, the left column represents different language modeling (LM) methods which are often used in morphologically rich language such as Turkish. ASR performances are shown using Word Error Rate (WER) in Table 1.1. For all LM methods, the presence of the background music is decreasing the recognition performance of the system. In order to obtain more robust ASR systems, the recognition accuracy in the speech segments with the background music has to be increased.

Table 1.1. Turkish ASR results (in WER) for different acoustic conditions [1].

LM Method	Clean Speech	Background Music
Word	27.7	45.9
Morphs	19.9	38.3
Stem-Ending	19.4	38.2

1.1. Statement of the Problem

As stated, removing the background music is important for developing robust ASR systems. The first step in the removal of the background music is the detection of such segments in the incoming audio signal. Then, a separation system can obtain separated speech and music signals from the mixed signal with or without using the pure speech and music segments in the audio. Therefore, a real-world ASR system should contain a front-end processing unit capable of segmenting and separating music and speech from the incoming audio. Such a system is shown in Figure 1.1. Since there is only one observation from the mixed signal, the separation problem is classified as

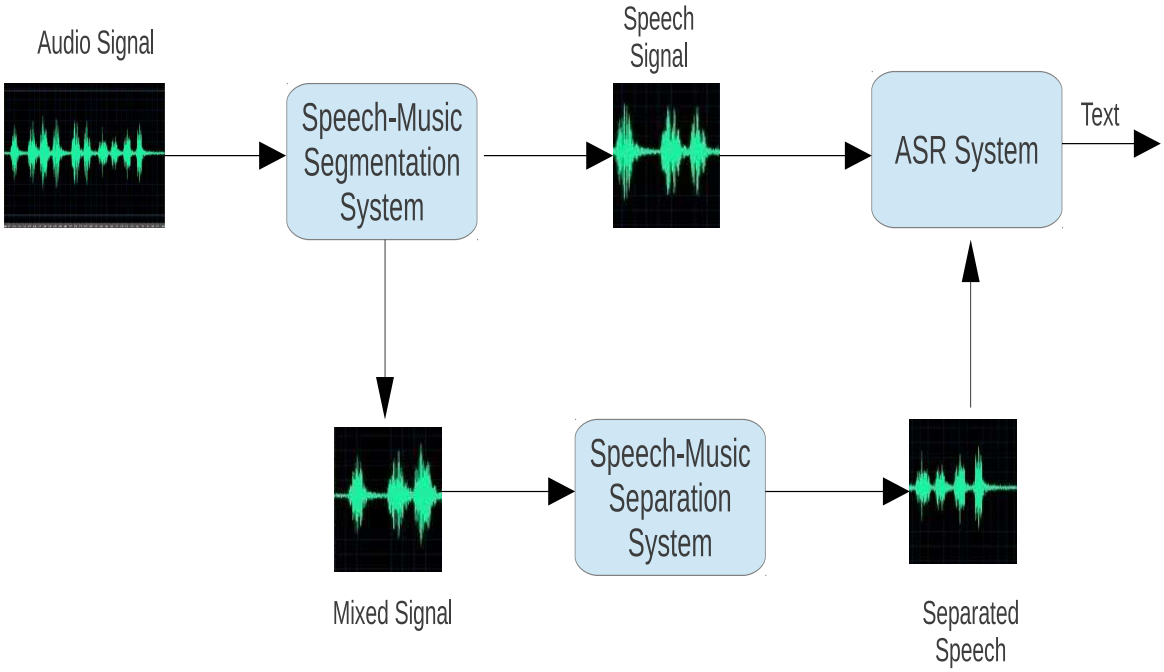

Figure 1.1. Speech-music segmentation and separation front-end for ASR.

single-channel source separation problem. This is an ill-conditioned problem [3] which has infinitely many solutions. Therefore, some extra conditions or information on the source signals have to be imposed to provide the optimal separated source signals.

In fact, this is not a Blind Source Separation (BSS) problem in the sense that we have some prior information about the source signals. At least, it is known that the mixed signal contains both speech and music signals. Moreover, it should be emphasized that the speech and music signals can be modeled using some training data. Since the separation system is used as a front-end for an ASR system, the training data to learn parameters of the acoustic model can be used for the speech signal modeling. Moreover, the main motivation of the background music removal is to increase the ASR performance on broadcast news data. Therefore, the music signals can be limited to those which are widely used to generate the background music in the broadcast news.

The aim of this study is to develop such a music-speech separation technique that

can be used as a front-end for an ASR system. In [2], it was shown experimentally that background music does not affect the ASR performance as seriously as white noise at the same SNR values. However, standard noise reduction techniques are not applicable to background music. Therefore, we approach the problem as a single-channel source separation task. The main difference between the speech-music separation and the noise reduction problem is the goal: the speech-music separation for ASR tasks aims to increase the recognition accuracy whereas the noise reduction techniques aim to increase the speech-music ratio (SMR) value in the separated signal.

1.2. Main Contribution of the Thesis

The contribution of this study is to develop a probabilistic approach to single-channel speech-music separation problem and to analyzing the performance improvement not only with source separation measures but also with ASR performance measures. Our approach proposes a representation to the source signals using different probabilistic models, a non-negative matrix factorization (NMF) model for the speech signal and a mixture model for the music signal. Representing the source signal with different probabilistic models for the speech-music separation task is proposed in this dissertation for the first time.

The motivation behind our approach is that, especially in broadcast news, most of the time, the background music is composed of the same piece of music, called a 'jingle'. Jingles are often produced by repeating a fixed set of music material such as drum loops or sound samples. Therefore, we assume that we can learn a catalog of these jingles and hope to improve separation performance. In this study, the identity of the jingle is assumed to be known as a prior for each mixed signal.

We introduced [4] a probabilistic model-based approach to separate speech and music signals. Unlike other probabilistic approaches, we do not model the speech in great detail, but instead focus on modeling the music. In our model, the catalog corresponds to a conditional mixture model. We assume, for each mixed signal, the jingle which generates the background music can be detected using the music segment of the

audio. Therefore, each frame of the spectrogram of the background music is assumed to be generated by scaling a single mixture component, i.e., a jingle frame. The scaling parameter consists of a filtering parameter for each frequency bin and a gain parameter for each time frame. The speech spectrogram is generated from an NMF model. The observed spectrogram is the sum of the speech and music spectrograms. Separation is achieved by joint estimation of the unknown parameters and latent variables of this hierarchical model, a mixture of NMF models. From a probabilistic model point of view, the main contribution of the thesis is to combine NMF models for the speech and a mixture model for the music signals, respectively, and hence obtain a mixture of NMF model for the representation of the mixed signal. Moreover, the inference method for the mixture of NMF models is developed in our thesis for the first time.

The probabilistic interpretation of the NMF models [5, 6] is used for developing the separation algorithm. However, in [5,6], both of the source signals are represented using NMF models. In our thesis, the source signals are represented using different probabilistic models (an NMF model for the speech and a mixture model for music signals, respectively) and an combined overall model is obtained for the mixed signal. Since a probabilistic approach to the source separation is developed, we can easily combine the NMF model for the speech signal and the mixture model for the music signal. Moreover, due to using the probabilistic approach, the proposed models can be easily extended such that it contains prior information about the sources. The separation algorithm includes finding the active jingle frame index for each time frame with its scaling parameter, filtering and gain parameters. Moreover, the algorithm estimates the parameters of the NMF model to reconstruct the spectrogram of the speech sources.

In this probabilistic framework, Poisson or complex Gaussian observation models are used for representing the magnitude or power spectrograms of the sources, respectively. While using Poisson observation model, we are minimizing the Kullback-Leibler (KL) divergence between the magnitude spectrogram of the signal and its estimated value, whereas using complex Gaussian observation model, we are minimizing Itakura-Saito (IS) divergence between the power spectrogram of the signal and its estimated

value. As a result, the music model corresponds to a Poisson mixture model (PMM) or a complex Gaussian mixture model (CGMM) and the overall probabilistic model consists of the combination of an NMF model for the speech signal and a mixture model for the music signal.

The main contributions of the thesis can be summarized as follows:

- *Mixture of NMF Model*: In this thesis, we propose using an NMF model for the speech signal with a mixture model for the music signal due to the repetitive structure of the background music in broadcast news. The usage of different probabilistic model for the source signals in the speech-music separation framework is proposed in this study [4]. Moreover, an inference method (Expectation-Maximization (EM) Algorithm) for the proposed probabilistic model is also developed. As a baseline strategy, NMF modeling of both sources (speech and music signals) is used and the separation performances of the methods are compared. The advantage of using a mixture model for the music signal with the proposed method is experimentally shown as compared to a conventional NMF method [7].
- *Comparison of Divergence Measures for Speech-Music Separation:* KL and IS divergences, which correspond to Poisson and complex Gaussian observation models, respectively, are used with mixture and NMF models. The separation performances with the divergence measures are compared in the experiments. It is shown that IS divergence has better separation performance [8].
- *Gain Estimation Problem in Poisson Mixture Model:* In the experimental study, it was pointed that the gain values which correspond to the volume changes in the background music are not estimated accurately in Poisson model. The reasons for poor estimation performance are analyzed in this study [9]. Moreover, 3 different gain estimation strategies are proposed which are:
  (i) Maximum A Posteriori Estimation
  (ii) Piece-wise Constant Estimation
  (iii) Gamma Markov Chain (GMC) Estimation

  Although the gain estimation performance of complex Gaussian model is better as compared to the Poisson case, it is shown that using GMC for gain values

improved the separation performance of the complex Gaussian method [10] even more.

- *Markovian Extension to Mixture Model:* In the proposed mixture model for the music signal, the temporal dependency between the jingle frames is ignored. In order to benefit from the continuity information between the jingle frames, Markovian extension to the proposed mixture model is also developed in this thesis [10]. The advantage of the incorporating temporal dependency between the jingle frames is also proved experimentally. Moreover, it is shown that Markovian extension results more performance improvement in Poisson model as compared to complex Gaussian model due to the fact that the baseline separation performance of IS model with mixture model is better compared to the KL model.
- *Separation Experiment with Real Data Recordings:* In this thesis work, the separation performance of the proposed methods are not only tested with synthetically mixed signals but also with real speech data with background music obtained from the broadcast news recordings. The difference from the synthetic case is that the reference results cannot be calculated due to the fact that the unmixed signals are not provided. However, it is experimentally shown that the proposed mixture of NMF based method improves the ASR performance as compared to the mixed case [10]. Moreover, ASR performance improvement with the proposed method is higher than the conventional NMF methods.
- *Speech Modeling For Speech-Music Separation:* In previous studies, speech templates are trained using the NMF models and the separation is performed using these fixed templates. However, in this study, we propose to train prior speech models using the training data. Then we perform the separation using the prior models combined with the mixture model for the music signal. Moreover, the effect of training data types is analyzed in this work [11, 12].
- *Sub-word Modeling For Speech-Music Separation:* In a previous work [13], phone-based modeling of the speech signal was proposed. However, in [13], the effect of the method to speech recognition is not analyzed. In other words, there was no performance improvement with objective criteria such as WAcc. In this study, we analyze the effect of phone-based models in the ASR task. Moreover, we analyze the oracle separation performance with the sub-word models such as phones or

states. The usage of N-best lists is also tested in this study.

1.3. Organization of the Thesis

The dissertation is organized as follows: In Chapter 2, we give an overview of speech recognition and single-channel source separation problems. The model-based source separation is emphasized in this chapter. Moreover, a review of the previous work that the dissertation is based on is given. In Chapter 3, we review the NMF model as a data modeling approach and then give the probabilistic interpretation of NMF model for both KL and IS divergences. Moreover, speech-music separation method with NMF models is described in Chapter 3. Furthermore, speech-music separation experiments are carried out in this chapter using different training sets for speech and music signals. In Chapter 4, a mixture model for the music signal is proposed and corresponding probabilistic model is described. An inference method for the overall mixture of NMF model with both divergence measures, KL and IS, is developed in this chapter. Gain estimation problem for KL divergence case is investigated and three different solutions are proposed in Chapter 4. Moreover, the mixture model for the music signal is improved by incorporating temporal dependency information between the jingle frames using Markovian structure. The proposed methods are not only tested with the synthetically mixed signals in this chapter but also tested with a real data set which are taken from the broadcast news speeches. In Chapter 5, we focus on modeling the speech signal by using some prior training data. For both of divergence measures, KL and IS, the mixture of NMF model with a prior model on the speech template is described and a variational inference method for this model is developed in Chapter 5. Moreover, for the speech signal, different type of training data set is tested in Chapter 5. In Chapter 6, we improve the speech model by using sub-word units for the representation of the speech signal in the separation process. Instead of a general speech model, phones or states are used in this chapter as a modeling unit.

2. BACKGROUND

The main focus of this dissertation is to investigate single-channel source separation problem for improving performance of ASR systems. This chapter gives a summary of the background about the related tasks in this thesis. The main components in this thesis are ASR and source separation systems. In the next sections, brief description of the main components will be given.

2.1. Foundations of Automatic Speech Recognition

In recent years the statistical approach to speech recognition has prevailed over other approaches. Given a sequence of acoustic observations $s_1^T = s_1, ..., s_T$, the aim of speech recognition is to find the best possible word sequence $w_1^N = w_1, ..., w_N$ as

$$[w_1^N]^* = \arg\max_{w_1^N} p(w_1^N|s_1^T) \tag{2.1}$$

With use of Bayes Rule, the equation can be written as follows:

$$[w_1^N]^* = \arg\max_{w_1^N} \frac{p(s_1^T|w_1^N)p(w_1^N)}{p(s_1^T)} \tag{2.2}$$

It is assumed that the probability of the observation sequence is the same as for all word combinations w_1^N.Therefore, Equation 2.2 can be written as:

$$[w_1^N]^* = \arg\max_{w_1^N} p(s_1^T|w_1^N)p(w_1^N) \tag{2.3}$$

The first component in Equation 2.3, $p(s_1^T|w_1^N)$, represents the acoustic likelihood of the observation sequence given the word sequence.

The acoustic likelihood can be computed using an acoustic model which is trained with the transcribed speech data. The acoustic models are based on the concatenation of hidden Markov models (HMM) for each phone in the word. Instead of using training

only one model for each phone, for each different context, a context dependent model is trained using the transcribed speech data. Most often, Gaussian mixture models (GMM) are used for representing the acoustic variability in each context dependent model.

The language model score, $p(w_1^N)$, is regarded as prior probability given to the word sequence, w_1^N. The prior probability of the word sequence is computed using a training text corpus. Due to the sparsity problem in computing the probability of a long sentence, an approximation to this value can be calculated using the n-gram approach which can be defined as using the following equation:

$$p(w_1^N) \approx \prod_{i=1}^{N} p(w_i | w_{i-n+1}^{i-1}) \tag{2.4}$$

In other words, in calculating the word given a history is limited to the previous $n-1$ words in a n-gram language model to make a more robust estimation for the language model score.

Actually, ASR is a search problem in which all possible word sequences have to be tested for finding the best possible word sequence corresponding to the incoming speech signal. While all possible word sequences are contained in the language model, acoustic model is used for calculating the likelihood of each possible word sequence. This overall process is shown in Figure 2.1.

2.1.1. Robustness in Speech Recognition

The recognition accuracy of an ASR system is dependent on the amount of mismatch between the content of the target speech and trained models (acoustic and language models). In this thesis, the focus is on the mismatch between the speech signal and the acoustic model. The main reason for the acoustic mismatch is the noisy target signal. There are two approaches to compensate the mismatch between the acoustic model and the noisy speech signal. The first one is to adapt the acoustic

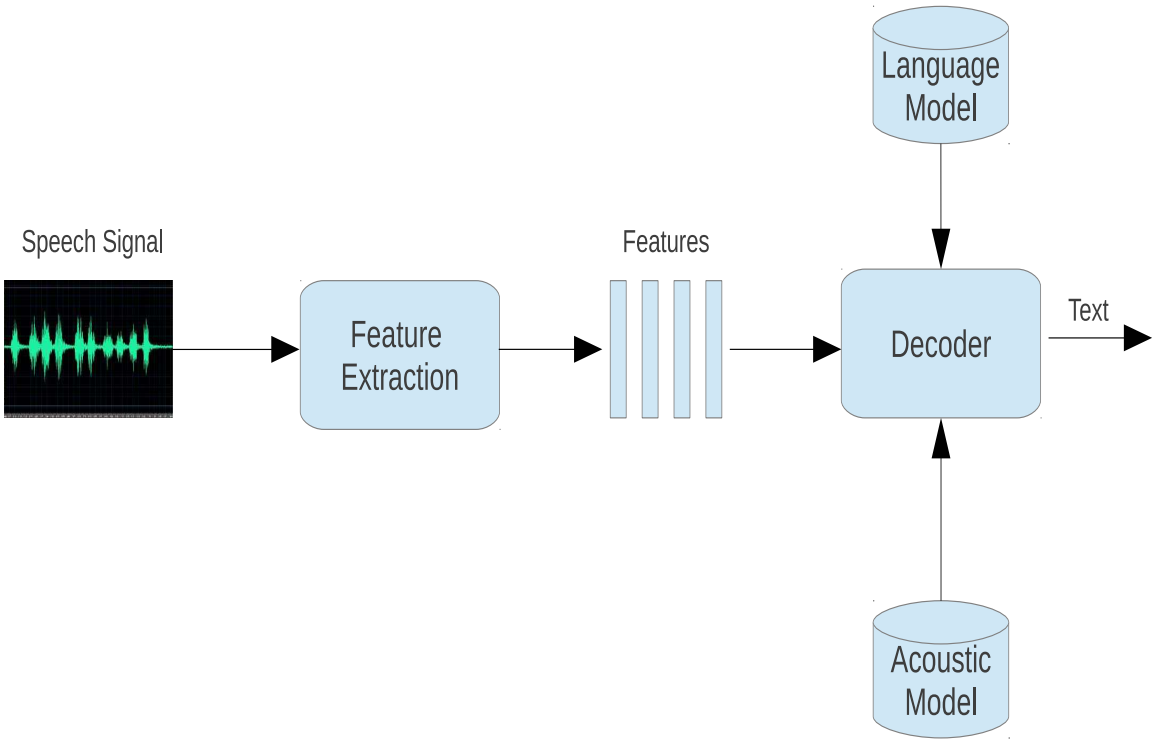

Figure 2.1. ASR system components.

model to the noisy environment. Since it is impossible to adapt the acoustic model to each noisy condition, the model adaptation is not much preferred for robust speech recognition.

The second approach is to suppress noise in the target speech signal, and hence decrease the mismatch between the acoustic model and the speech signal. Although background music has a more structured characteristic than background noise, it also increases the mismatch between acoustic model and test signal. Therefore, it is significant to decrease the mismatch effect of background music in ASR systems. With a more structured characteristic, it is more appropriate to use pre-trained model for the music signal in the source separation framework.

In order to increase the robustness of ASR systems against the noise signal, it is common to use robust features in ASR system such as perceptual linear prediction (PLP) and power normalized cepstral coefficients (PNCC) features. However, since this type of features are designed by considering noise characteristics, they are not effective against the background music.

2.1.2. ASR Performance Measure

Evaluation of the speech recognition performance of an algorithm is based on the distance between the reference and hypothesized word sequences. There are three type of errors between the reference and hypothesis:

- Insertion (I): Inserting a non-existent word to the reference.
- Deletion (D): Deleting a word from the reference.
- Substitution (S): Substituting a word in the reference.

If there are N words in a reference, word error rate (WER) is defined as:

$$WER = \frac{I + D + S}{N} * 100 \tag{2.5}$$

In this study, word accuracy (WAcc), which is defined as $100 - WER$, is used to measure the speech recognition performance of the systems.

2.2. Foundations of Source Separation

The single-channel source separation for background music removal problem can be defined as the estimation of the original speech signal given only an observed mixture of speech and music signals. In mathematical formulation, we may write the mixing process as

$$x = s + m \tag{2.6}$$

where x,s and m represent mixed, speech and music signals respectively. Single-channel source separation is an under-determined problem and its solution requires additional information about the sources. In source-modeling approach, pre-trained source models correspond to the additional information about the sources. We want to find a method that uses mixed signal and source models to estimate the speech signal. This type of the method with pre-trained models is called as 'Model-Based' approaches to the source separation problem. A typical scheme of a model-based technique is shown Figure 2.2.

When we have more than one mixture of the sources, we can use different statistical properties of the sources in the separation of the mixed signals from each other. However, in the case of the single-channel mixture, we have to use prior information about the sources in the separation. In order to use pre-trained source models in the separation method, there are five issues that must be considered as shown in Figure 2.2:

- Training Sets
- Feature Extraction
- Modeling Techniques
- Separation Method

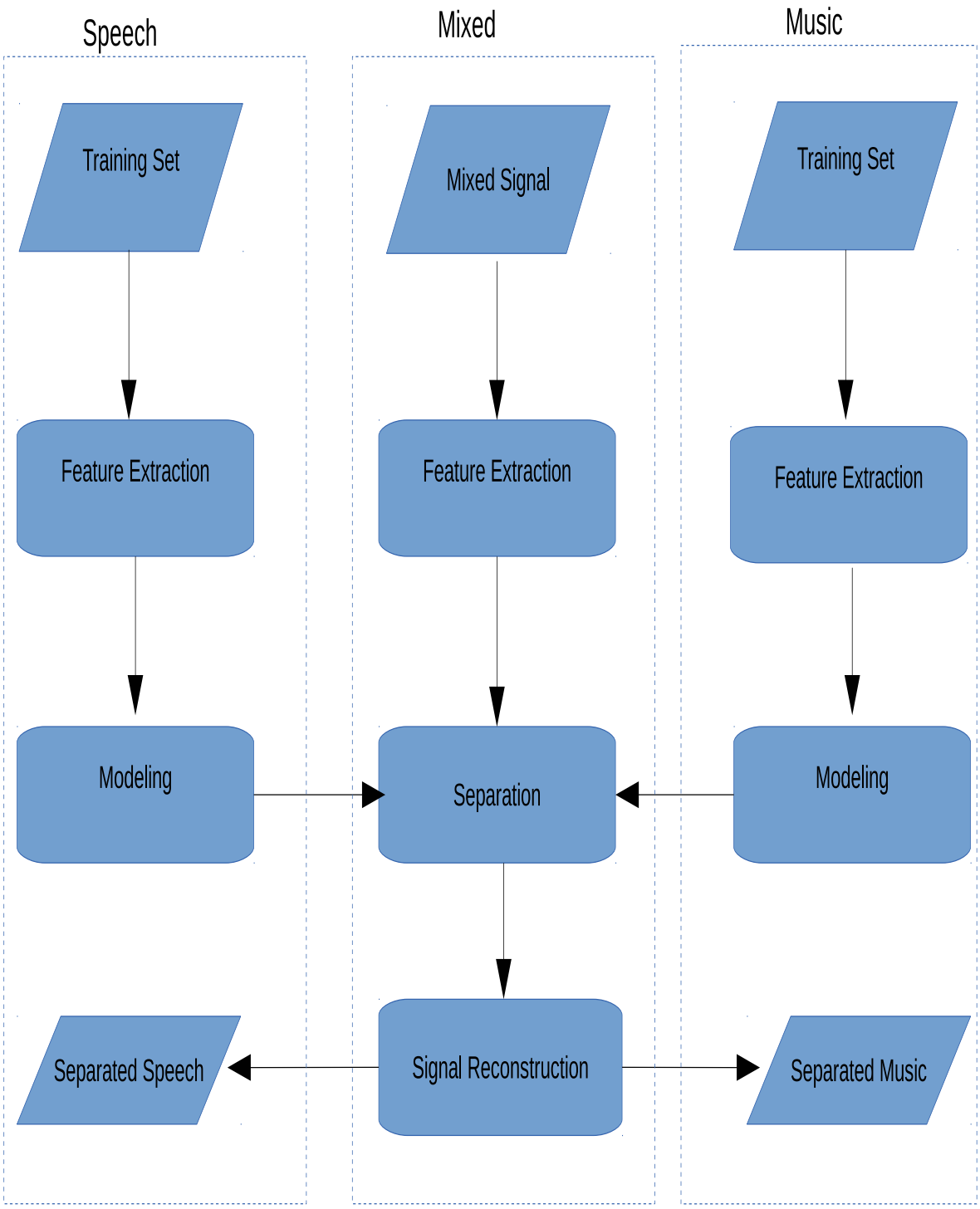

Figure 2.2. Visual representation of model based source separation methods.

- Reconstruction Strategy

Training Sets: There are two issues related to the training data sets in the source separation framework. First one is to use training data for each source. It is not necessary to use the training examples for both of the sources. However, it is not realistic to perform the single-channel source separation task without any pre-trained models. At least, for one of the sources, some training data is needed to learn the models. If training data is available for one of the sources, the other source signal can be estimated from the mixture signal. Second issue is the type of the training data as compared to the target signal data. The best one is to use the same type of data as the target data. However, it is not always possible to provide such type of data. Therefore, the type of training data is an important concern for the separation method.

Feature Extraction: Time or frequency domain representations can be used to represent the sources. Power or magnitude spectrum of the signals can be used to represent the sources in the frequency domain. The important issue related to the feature selected for the separation is the compatibility of the feature to the modeling approach and the availability of the reconstruction techniques with the features. As a non-negative feature, power or magnitude spectrum is appropriate and can be used with NMF or mixture based modeling approaches.

Modeling Techniques: As a modeling technique;

- Mixture Model (MM)
- Hidden Markov Model (HMM)
- Non-negative Matrix Factorization (NMF)

can be used to learn the prior information about the source signals. With mixture model based approaches, the mixture components are learned from the training data and for each time frame, it is assumed that one of the mixture components generates the data. However, in NMF model, it is assumed that the data vector is represented using the non-negative weighted sum of the non-negative templates which are estimated

from the training data vectors.

Separation Method: The choice of the separation method mainly depends on the modeling techniques of the sources signals. For example, if NMF method is used for modeling of both source signals, by using the same approach as training of NMF model, the separation method can be implemented. The source signals are represented using different modeling techniques in this thesis. Therefore, the separation method is more complex as compared to representing the source signals with the same models.

Reconstruction Strategy: Spectrum of the source signals whose parameters are found in the mixed signal using the separation method must be estimated. The reconstruction of the sources in time domain is necessary for not only measuring the separation performance measures but also for some applications. Therefore, by using the estimated spectrum of the source signals with the phase of the mixed signal, the source signals can be reconstructed in the time domain.

2.2.1. Separation Performance Measures

The evaluation method for speech-music separation aim measuring the amount of distortion between the original signal, s, and its separated version, $\hat{s}$. There are two types of effects in the recovered speech signal and they are measured using the following criteria:

- Speech to Music Ratio (SMR): measures the amount of residual of the music source in the separated speech signal.
- Source to Artifact Ratio (SAR): measures the amount of distortion due to the separation method in the separated speech signal.

In order to calculate two criteria of the separation algorithm, the effects in the separated speech signal have to be defined. There are three components in the recovered speech signal and can be listed as follows:

- Speech signal which proportional to the speech source signal represented as $\alpha_1 s$.
- Music signal contained in the recovered speech signal which can be regarded as an interference and is proportional to the music source signal, $\alpha_2 m$.
- Noise signal is generated by the the separation method and represented as n_1.

With the assumption of uncorrelatedness between speech and music source signals, we can make the following definitions to calculate the SMR and SAR values in dB. The same definitions can be used for the music signal by reversing the speech and music signals in the following definitions.

- Estimated Sources:

$$\hat{s} = \alpha_1 s + \alpha_2 m + n_1 \tag{2.7}$$

$$\hat{m} = \beta_1 s + \beta_2 m + n_2 \tag{2.8}$$

- Source Coefficients:

$$\alpha_1 = \langle s, \hat{s} \rangle \quad \alpha_2 = \langle m, \hat{s} \rangle \tag{2.9}$$

$$\beta_1 = \langle s, \hat{m} \rangle \quad \beta_2 = \langle m, \hat{m} \rangle \tag{2.10}$$

  where $\langle ., . \rangle$ represents dot product of the signals.
- Performance Measures:

$$SMR = 20 \log_{10} |\frac{\alpha_1}{\alpha_2}| \frac{\|s\|}{\|m\|} \tag{2.11}$$

$$SAR = 20 \log_{10} \frac{\|\hat{s} - n_1\|}{\|n_1\|} \tag{2.12}$$

2.3. Related Work

Many researchers studied single-channel source separation for mixture of speech from two speakers [14, 15] but there are only a few studies on single-channel speech-music separation [13, 16–19]. When we have more than one mixture of the sources from multiple channels, we can apply BSS techniques, which use assumptions about the sources such as independence [20]. However, in the case of single-channel mixtures, model-based approaches are needed to separate the sources. Model-based approaches are used to separate sound mixtures that contain the same class of sources such as speech from different people [21, 22] or music from different instruments [23, 24].

The pre-trained models in the single-channel separation task is used firstly by Roweis [25]. In [25], for each speaker an HMM is fit using patches of narrow-band spectrograms as the pattern vectors. In this model, the emission densities model the typical spectral patterns produced by each talker, while the transition probabilities encourage continuity. To separate a new single recording which is a mixture of known speakers, the pre-trained HMMs are combined into a factorial HMM (FHMM) architecture [26].

In [27], the performance of the HMM and GMM modelling techniques are compared. The motivation of using GMM is to take into account the diverse structure of sounds through multiple power spectral densities (PSD)s. The HMMs permit to take into account the *a priori* time dependencies between the modelled PSDs, through the state dependency structure.

The source separation technique presented in [23] suggests the use of Gaussian scaled mixture models (GSMMs) to model the statistical behavior of sources. In this technique, the speaker dependent models are formed by GSMM parameters trained from the sample data. The GSMM incorporates a supplementary scale parameter which aims at better taking into account non-stationarity of the sources.

Blouet *et al.* [17] compares the performances of three different code-book based

source separation techniques in speech-music separation task. GSMM-Based, auto-regressive (AR)-Based and amplitude factor-based code-books are developed. AR-Based approach is used because of the fact that spectral envelope of speech signals in the short-time Fourier transform (STFT) domain are efficiently characterized by AR models, which have been used for speech enhancement in [28]. Amplitude factor source separation technique [29] proposes to model each STFT frame of each source as a sum of elementary components modelled as zero-mean complex Gaussian distribution with known PSD and scaled by amplitude factors.

Tsai *et al.* [30] proposed to adapt music and voice models directly from the recording. In the first phase each recording is automatically segmented in a succession of vocal and non-vocal parts. Then, an adapted music model is learned on the non-vocal parts. Finally, using the adapted music model as a prior, an adapted voice model is trained from the vocal parts. Ozerov *et al.* [31] also used the same strategy in singing voice separation. However, Ozerov *et al.* used GMM-based source separation and they proposed to use maximum likelihood linear regression (MLLR) [32] to adapt the source models.

Schmidt and Olsson [14] applied sparse NMF (SNMF) algorithm to the speech separation problem. Speech signals are represented in Mel spectrum magnitude domain as suggested in [33] and it was assumed that spectrogram for each speaker can be sparsely represented in an over-complete basis dictionary, that is, each data point is a linear combination of few columns of the dictionary matrix $\mathbf{D}$. This corresponds to the sparsity of the code matrix $\mathbf{E}$. In order to learn the dictionary matrices for each speaker and the separation using these speaker-dependent matrices, SNMF algorithm proposed in [34] is used. The summary of the method can be seen in Figure 2.3. Moreover, phoneme-dependent dictionaries are also proposed to be used in separation. In this approach, the training data is first segmented according to phoneme labels obtained by a speech recognition software based on a HMM and the overall dictionary is constructed by concatenating the individual phoneme dictionaries. As compared to our work, in Chapter 6 we also propose to use phone-based speech models in the separation. However, whereas Schmidt and Olsson in [14] used phone models in order

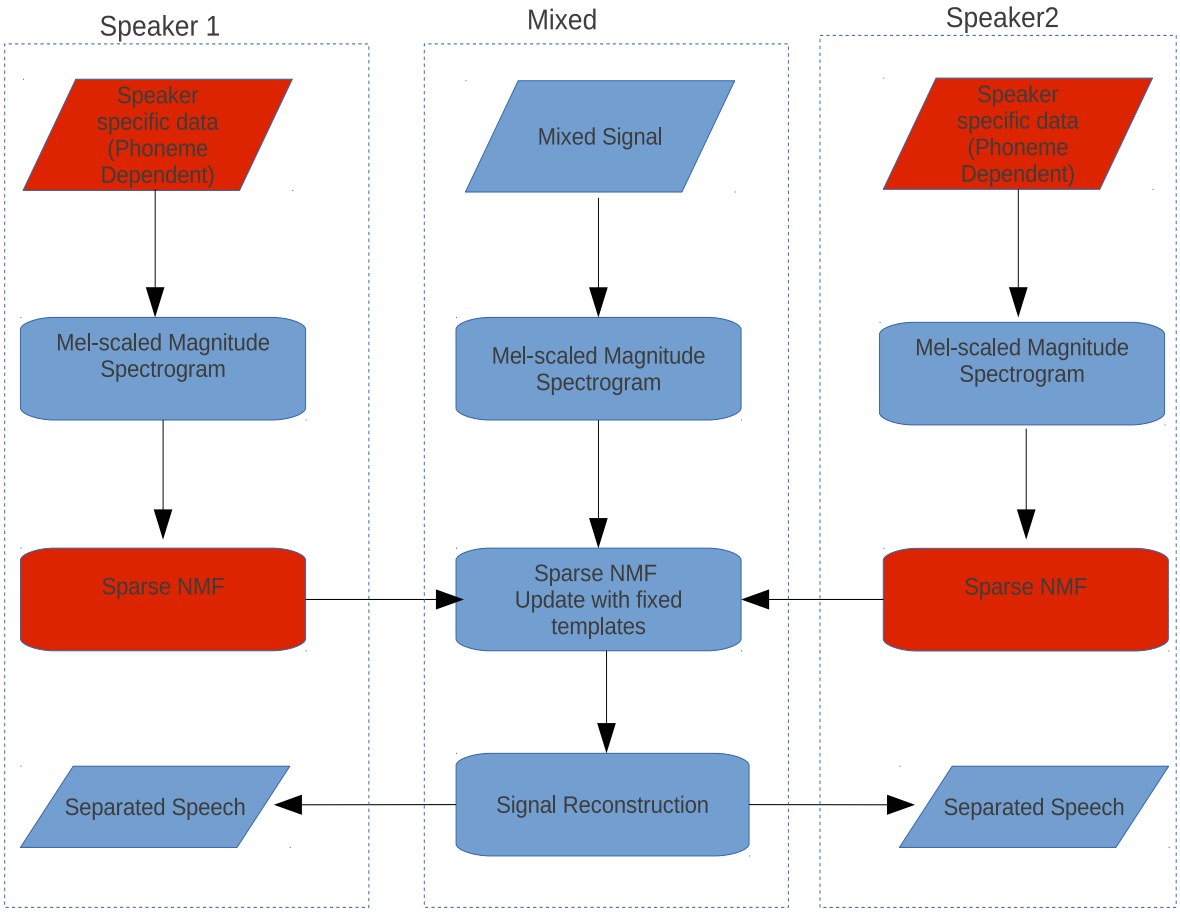

Figure 2.3. Sparse NMF based speech-speech separation.

to construct the overall speech template matrix, we propose to use a different phone-model for each time frame in the separation phase.

Virtanen [24] proposed an unsupervised sound source separation algorithm which combines NMF with temporal continuity and sparseness objectives. The method proposed a cost function, which is the sum of the squared differences between the gains in adjacent frames. This is a simple and efficient way of including the temporal continuity objective into the separation framework. The method is applied to separate the sound sources in music signals and the experiments showed that the temporal continuity criterion improves the detection accuracy of the pitched sounds and improves their signal-to-noise ratio (SNRs) slightly.

Raj *et al.* [18] used the NMF method for compensating the music signal for an

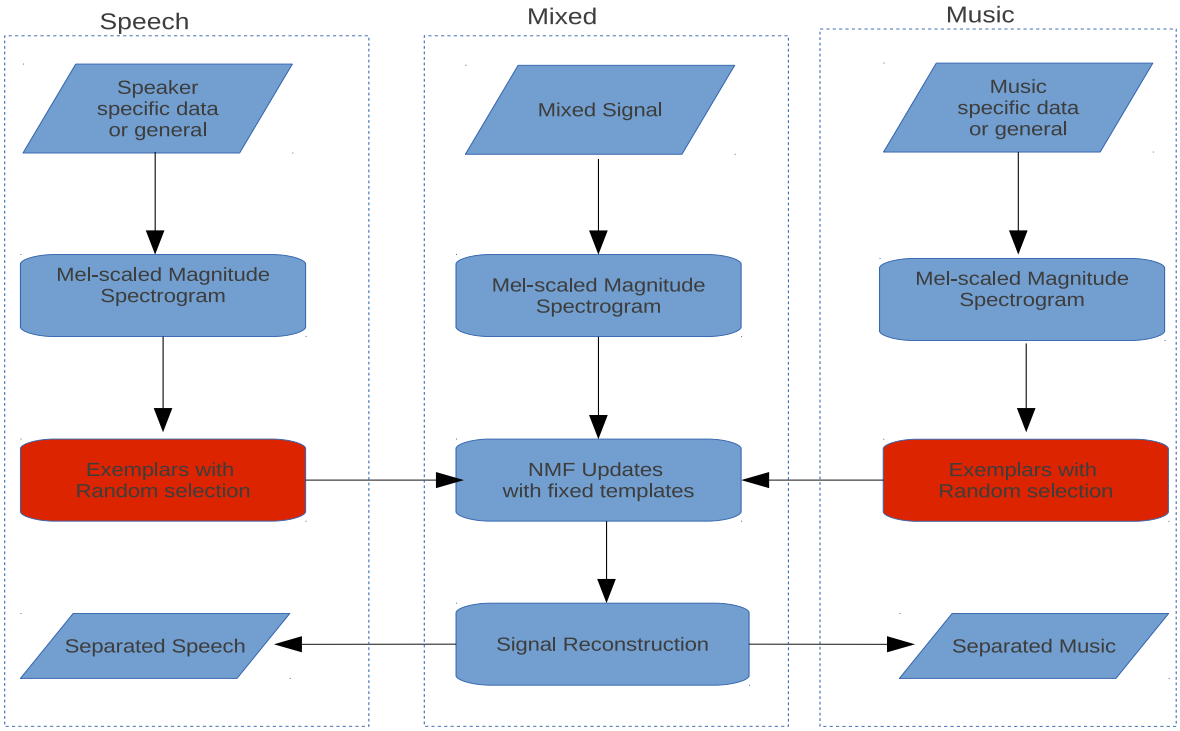

Figure 2.4. Exemplar based speech-music separation system.

ASR system for the first time. The summary of the method can be seen in Figure 2.4. They showed that NMF-based approaches are capable of generating enhanced signals that significantly improve the speech recognition performance. In [18], over-complete dictionaries consisting of random exemplars of the training data is used in the separation. The proposed system is tested on Wall Street Journal database which is artificially mixed with music signal. Experimental results in [18] show that although the compensation requires bases drawn from the music and speech signal, it works well when the identity of the music or speaker are unknown. As compared to our work with [18], though speech and music signals are modeled using NMF method in [18], we used a mixture model for the music signal and an NMF model for the speech signal. Moreover, in [18], the template vectors of the speech and music signals are chosen from the training vectors. In other words, NMF training is not applied for obtaining the speech and music templates.

Raj *et al.* [13] used phone-dependent NMF models for speech-music separation. The summary of the method can be seen in Figure 2.5. Actually, in [13], the templates

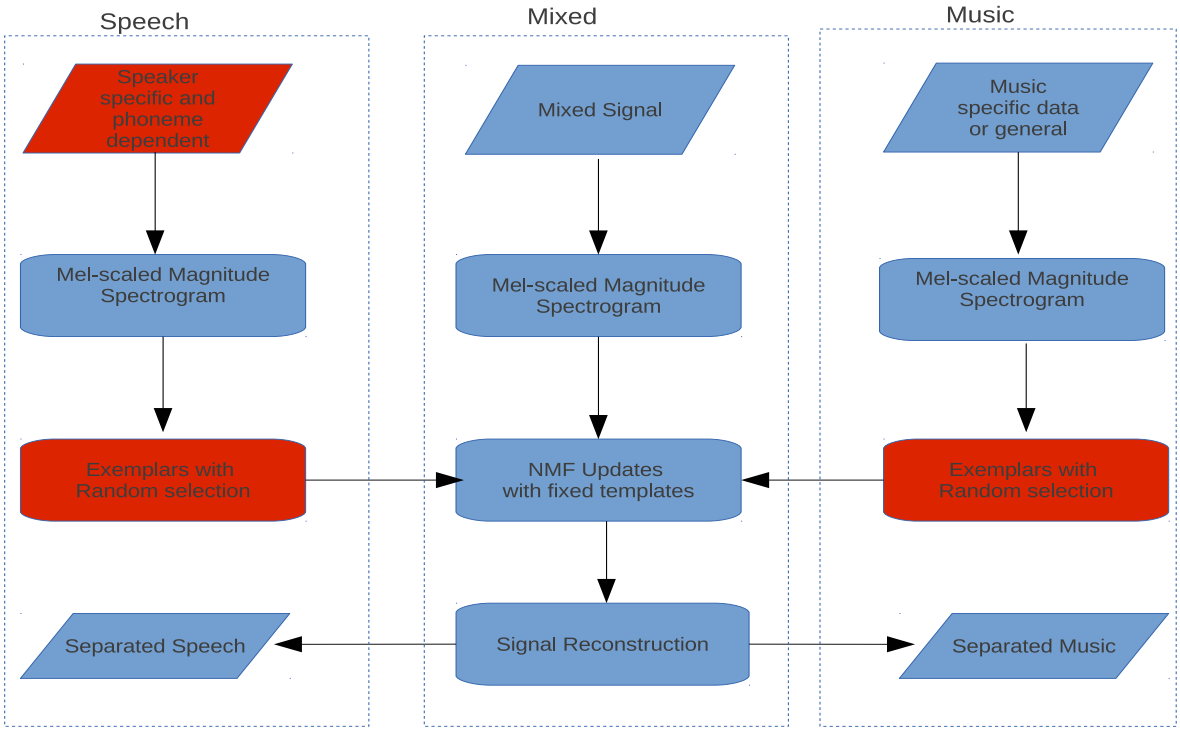

Figure 2.5. Phoneme-dependent exemplar based speech-music separation system.

for the speech and music signals are chosen from the training vectors which is called as 'exemplar' approach. The NMF updates are only used in the separation phase by fixing the exemplar-based speech and music templates. The phone-dependent approach in [13] is similar to our work in Chapter 6. However, in [13], the effect of the method to speech recognition is not analyzed. In other words, there was no performance improvement with objective criteria.

In [3], time-domain basis functions are used to model the source signals. Basis functions are trained using independent component analysis (ICA) method. Moreover, the separation is achieved using ML approach. Different from [3], we do not use training data for speech signal and we model the source signals in frequency domain. Another important difference between [3] and the current study is the evaluation of the separation performance measures. In [3], the separation performance is evaluated with source-to-interference (SIR) Ratio. In the current study, the separation performance is not only evaluated with separation measures such as SIR, source-to-artifact (SAR) ratio and source-to-distortion (SDR) ratio, but also with speech recognition performance

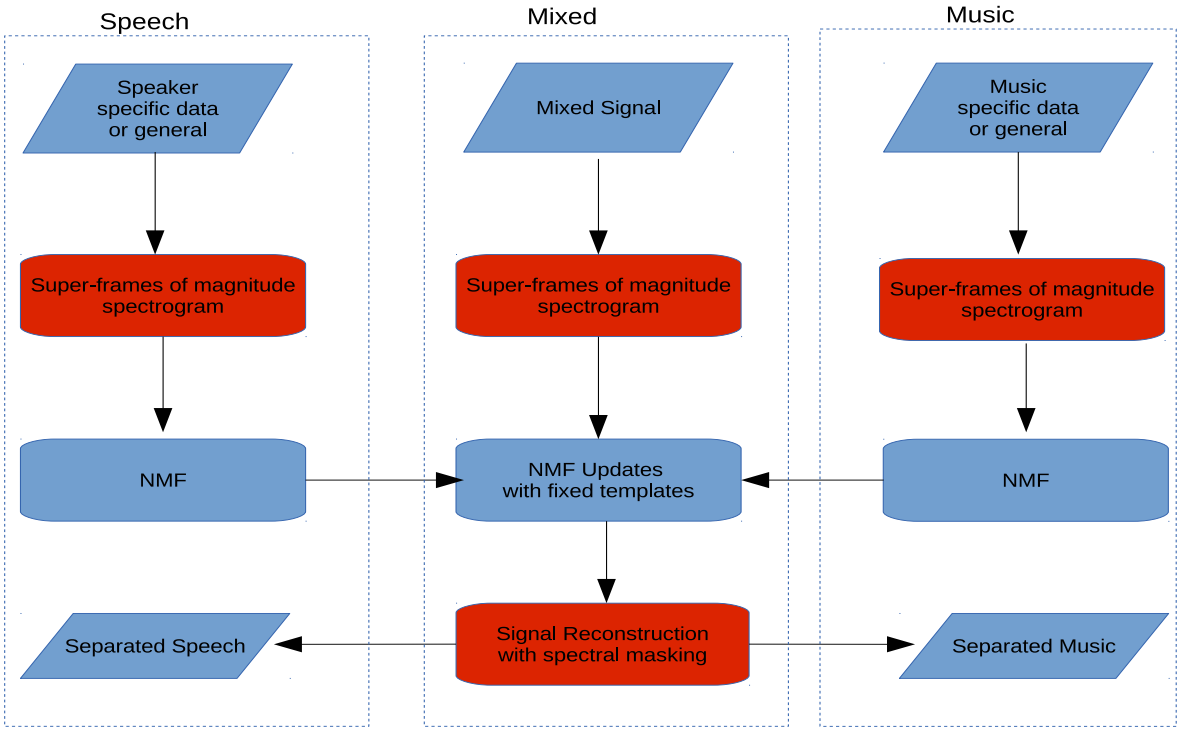

Figure 2.6. Speech-music separation system with super-frames and spectral masks.

measure.

In [19], NMF models are trained for speech and music signals. In the separation phase estimated source signal spectrum is used to obtain a spectral mask to reconstruct the source signals. The summary of the method can be seen in Figure 2.6. As compared to our study, we used the spectral mask with $p = 1$ which corresponds to the linear ratio mentioned in [19]. In [35], Grais and Erdoğan improved the method in [19] by applying the sliding window approach to obtain the spectral frames of the signals. Instead of using the frames of the magnitude spectrum of the signal, the concatenated frames are used to represent the signal. In [36], Grais and Erdoğan proposed to use the excitation matrices estimated in the training phase of the NMF based speech-music separation method as a prior model to HMM model for each source. Temporal dependency between the source frames is taken into account in this way. Source dynamics is used in the separation.

Although in some cases NMF results sparse representation of the given data,

it is a side effect of NMF and it is not controlled by the algorithm it is controlled by data.Therefore, Eggert and Korner [34] introduced SNMF. The motivation was to control the sparsity of the weight matrix, $\mathbf{V}$. The idea is that we have an over-complete representation for the data matrix, that is, the number of basis vectors is more than the dimension of the data matrix.

Hoyer [37] proposed another SNMF technique. His goal was to find a decomposition in which hidden components (weights) are sparse, meaning that they have probability densities which are highly peaked at zero and have heavy tails. The method differs from Eggert and Korner's method [34] in update method. In Hoyer's update rules, weights were not multiplicatively updated. He used projected gradient descent technique to update weights matrix and they projected any non-negative values to zero to satisfy non-negativity constraint. Hoyer [38] improved his previous technique so that sparsity can be imposed on the weights and basis vectors and the amount of sparsity can be controlled using the sparsity measure

Although NMF provides a useful tool for analyzing data, it ignores potential dependencies across successive columns of its input $\mathbf{V}$. A regularly repeating pattern that spans multiple columns of its input $\mathbf{V}$ would have to be represented by NMF using multiple bases that describe the entire sequence. Since this is a regularly repeating pattern it would be more satisfying if it was represented by a single basis function that could span the pattern length. In order to solve these problems, a convolutive extension to NMF (CNMF) which allows to extract cross-column patterns as single bases is proposed in [39] and [40].

Sparse CNMF (SCNMF) is an extension of CNMF with imposed sparseness constraint. SCNMF was introduced by O'Grady and Pearlmutter [41] and they follow Hoyer's [37] approach to find update rules.

3. NMF BASED SINGLE-CHANNEL SOURCE SEPARATION

3.1. Overview of NMF Model

NMF is a matrix factorization technique that forces the entries of the matrix factors to be non-negative. In this study, the data matrix, **S**, represents the spectrum of the signals. In data analysis point of view, the aim of the factorization is to find the approximation to the data matrix in an efficient way. In mathematical formulation, with NMF, we want to find an approximate factorization to a non-negative data matrix **S** as

$$\mathbf{S} \approx \mathbf{DE} \tag{3.1}$$

where all elements of **D** and **E** matrices are also non-negative. While **D** matrix contains template vectors to span the data matrix column space, **E** matrix contains corresponding excitations to represent the columns of the data matrix using the template vectors. The matrix factorization of the data matrix is shown in Figure 3.1. Each time frame of the data matrix is sum of non-negative weighted columns of the template matrix. The excitation matrix contains the non-negative weights values for each time frame and template vector.

The NMF method was firstly used by Lee and Seung [42] to represent images of face as an alternative to vector quantization (VQ) and principal component principal component analysis (PCA) technique. Their motivation of using NMF was to obtain a parts-based representation to the face images. The corresponding template vectors were parts of a face and these vectors were combining additively to form a face image. They also proposed an elegant way of finding matrix factors for a given data matrix. It was called multiplicative update rules and they solved the matrix factorization problem as minimization of the cost function (CF). The CF, which is defined as the KL divergence

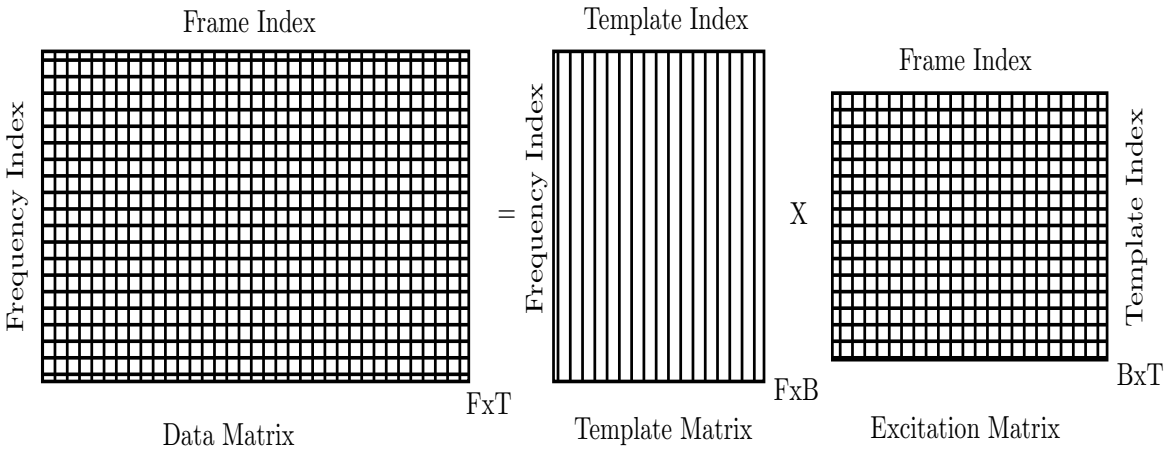

Figure 3.1. Visual representation of non-negative matrix factorization.

between the data matrix, **S**, and its approximation, **DE**, can be written as follows:

$$CF_{KL} = D_{KL}(\mathbf{S}||\mathbf{DE}) = -\sum_{f,t}\left\{S_{ft}\log([DE]_{ft}) - S_{ft}\log(S_{ft}) - ([DE]_{ft}) + S_{ft}\right\} \quad (3.2)$$

where f and t represents the index of frequency bins and time frames in spectral representation of the signals, respectively. The CF corresponds to KL divergence between the data matrix, **S**, and its approximation matrix, **DE**. This CF is not convex for both **D** and **E** matrices but it is convex for **D** or **E** separately. Therefore, it is not surprising that update rules will have two steps and in each step, **D** and **E** are updated separately. The corresponding update rules are:

$$\mathbf{D} = \mathbf{D} \otimes \frac{(\frac{\mathbf{S}}{\mathbf{DE}})\mathbf{E}^T}{\mathbf{1}\mathbf{E}^T} \quad (3.3)$$

$$\mathbf{E} = \mathbf{E} \otimes \frac{\mathbf{D}^T(\frac{\mathbf{S}}{\mathbf{DE}})}{\mathbf{D}^T\mathbf{1}} \quad (3.4)$$

where $\otimes$ represents element-wise multiplication of two matrices of the same size and **1** represents a matrix whose entries are equal to 1. In these equations, all divisions are element-wise and T represents transpose of a matrix. In [43], Lee and Seung also proved the convergence of update rules which guarantees that the CF is decreasing in

every step of the update equations.

In [6], it was investigated that the IS divergence between the data matrix, $\mathbf{S}$, and its approximation matrix, ($\mathbf{DE}$), as

$$CF_{IS} = D_{IS}(\mathbf{S}||\mathbf{DE}) = \sum_{f,t} \left\{ \frac{S_{ft}}{[DE]_{ft}} - \log(S_{ft}) + \log([DE]_{ft}) - 1 \right\}. \tag{3.5}$$

The corresponding update rules for IS divergence are:

$$\mathbf{D} = \mathbf{D} \otimes \frac{(\frac{\mathbf{S}}{(\mathbf{DE})^2})\mathbf{E}^T}{\frac{\mathbf{1}}{\mathbf{DE}}\mathbf{E}^T} \tag{3.6}$$

$$\mathbf{E} = \mathbf{E} \otimes \frac{\mathbf{D}^T(\frac{\mathbf{S}}{(\mathbf{DE})^2})}{\mathbf{D}^T\frac{\mathbf{1}}{\mathbf{DE}}}. \tag{3.7}$$

For the source separation point of view, KL and IS divergence cases use the magnitude and power spectrograms of the signals, respectively.

3.2. Probabilistic Interpretation of NMF

The interpretation of NMF as a low rank matrix approximation is sufficient for the derivation of a useful inference algorithm; yet this arguably does not provide the complete picture about the assumptions underlying the statistical properties of data matrix, $\mathbf{S}$. Therefore, Cemgil [5] described the NMF from a statistical perspective as a hierarchical probabilistic model. In [5], it was shown that the original multiplicative update equations of NMF appear as an EM algorithm for ML estimation of a conditionally Poisson model via data augmentation. Cemgil [5] also developed Bayesian extensions that facilitate more powerful modeling and allow more sophisticated inference, such as Bayesian model selection.

Fevotte and Cemgil [6] developed an interpretation of NMF methods based on Euclidean distance, KL and IS divergences in a probabilistic framework. They formu-

lated EM, Markov Chain Monte Carlo (MCMC) and Variational Bayes algorithms for these three different distance measures.

In [44], underlying probabilistic generative signal model of the NMF and the nonnegative update equations as a quasi gradient optimization are described in detail. Moreover, Gamma chain prior [45] on the template, **D**, and excitation, **E**, matrices are imposed as prior structures and it is shown that the resulting algorithm outperforms existing NMF strategies.

In [46], a Bayesian NMF model to separate tonal and percussive signals from a single-channel audio signal is proposed. The template and excitation matrices, **D** and **E** are divided into two partitions and assigned different prior distributions such that they encode a tonal and a percussive signal. The developed method in [46] estimates all parameters and hyper-parameters during inference, so there is no need for an additional training step in order to learn the basis vectors or their parameters. The method is evaluated to separate the musical instruments, flute and drums, from each other.

Virtanen and Cemgil [47] proposed a prior model based on the mixtures of Gamma distributions for each sound class to be separated. The method is used to separate speech from different speakers in a single-channel. In this model, hyper-parameters of the Gamma mixtures are trained using a training corpus given for each speaker. Using such scheme in separation allows adapting the spectral basis vectors of the sound sources during actual operation, when the exact characteristics of the sources are not known.

3.2.1. KL-NMF

In KL-NMF model, we factorize the data matrix as a multiplication of two non-negative matrices as

$$\mathbf{S} \approx \mathbf{DE} \tag{3.8}$$

where **S** represents the magnitude spectrogram of the signal to be modeled. NMF model assumes that for each time-frequency entry, the data component is generated by a number of Poisson sources whose intensities depends on the corresponding row of the template matrix and the column of the excitation matrix. We can show this mathematically as,

$$s_{fbt} \sim PO(s_{fbt}; D_{fb}E_{bt}) \tag{3.9}$$

$$S_{ft} = \sum_{b=1}^{B} s_{fbt} \tag{3.10}$$

$$s_{ft} \sim PO(s_{ft}; \sum_{b} D_{fb}E_{bt}) \tag{3.11}$$

where Poisson density of the random variable s is given as

$$\mathcal{PO}(s;\lambda) = \exp(s\log\lambda - \lambda - \log\Gamma(s+1)).$$

We call the variable $s_b = \{s_{fbt}\}$ as the b-th latent source where b represents the template vector index and B represents the number of template vectors. The probabilistic graphical model which represents the generative process for the NMF model is shown in Figure 3.2.

We can analytically marginalize out the latent sources $s = \{s_1, \ldots, s_B\}$ to obtain the marginal likelihood

$$\log p(\mathbf{S}|\mathbf{D},\mathbf{E}) = \log \sum_{s} p(\mathbf{S}|s)p(s|\mathbf{D},\mathbf{E}) = \log \prod_{ft} PO(S_{ft}; \sum_{b} D_{fb}E_{bt}) \tag{3.12}$$

$$\log p(\mathbf{S}|\mathbf{D},\mathbf{E}) = \sum_{f,t}(\; S_{ft} \log \sum_{b}(D_{fb}E_{bt}) - (\sum_{b} D_{fb}E_{bt}) - \log\Gamma(S_{ft}+1)) \tag{3.13}$$

The log-likelihood of the observed data, **S**, conditioned on template and excitation

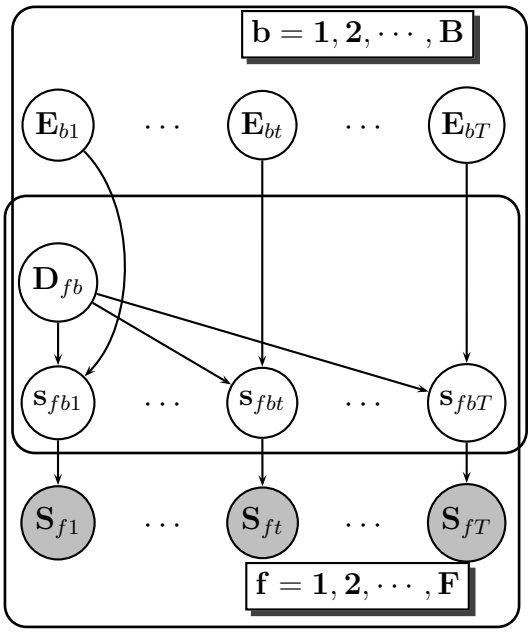

Figure 3.2. Graphical model for NMF.

matrices can be written as

$$L_S(\mathbf{D}, \mathbf{E}) = \log \sum_s p(\mathbf{S}|s)p(s|\mathbf{D}, \mathbf{E}) \geq \sum_s q(s) \log \frac{p(\mathbf{S}, s|\mathbf{D}, \mathbf{E})}{q(s)} = B_{EM}[q] \qquad (3.14)$$

where $q(s)$ represents the posterior distribution of the latent sources. We can show that the lower bound, $B_{EM}[q]$, is tight for the exact posterior of the latent sources,

$$\arg\max_{q(s)} B_{EM}[q] = p(s|\mathbf{S}, \mathbf{D}, \mathbf{E}) \qquad (3.15)$$

Hence the log-likelihood can be maximized iteratively

$$q(s)^{(n)} = p(s|\mathbf{S}, \mathbf{D}^{(n-1)}, \mathbf{E}^{(n-1)}) \qquad (3.16)$$

$$(\mathbf{D}^{(n)}, \mathbf{E}^{(n)}) = \arg\max_{\mathbf{D},\mathbf{E}} \langle \log p(s, \mathbf{S}|\mathbf{D}, \mathbf{E}) \rangle_{q(s)^{(n)}} \qquad (3.17)$$

where n shows iteration index.

E step: First let us write the joint distribution

$$p(\mathbf{S}, s|\mathbf{D}, \mathbf{E}) = p(\mathbf{S}|s, \mathbf{D}, \mathbf{E})p(s|\mathbf{D}, \mathbf{E}) = p(\mathbf{S}|s)p(s|\mathbf{D}, \mathbf{E}) \tag{3.18}$$

We know that $S_{ft} = \sum_b s_{fbt}$, therefore

$$p(S_{ft}|s_{fbt}) = \begin{cases} 1 & \text{if } S_{ft} = \sum_b s_{fbt} \\ 0 & \text{else} \end{cases} \tag{3.19}$$

Using the Poisson distribution, s_{fbt} is of the type

$$p(s_{fbt}|D_{fb}E_{bt}) = \mathcal{PO}(s_{fbt}; D_{fb}E_{bt}) = \exp(s_{fbt}\log(D_{fb}E_{bt}) - (D_{fb}E_{bt}) - \log\Gamma(s_{fbt}+1)) \tag{3.20}$$

Utilizing these information, (3.18) is extended in logarithmic fashion as

$$\begin{aligned} \log p(\mathbf{S}, s|\mathbf{D}, \mathbf{E}) &= \sum_{f,t}\Bigg\{\log\delta(S_{ft} - \sum_b s_{fbt}) + \sum_b \Big[s_{fbt}\log(D_{fb}E_{bt}) - (D_{fb}E_{bt}) \\ &\quad - \log\Gamma(s_{fbt}+1)\Big]\Bigg\} \end{aligned} \tag{3.21}$$

We also know that, due to the Poisson superposition property,

$$\log p(\mathbf{S}|\mathbf{D}, \mathbf{E}) = \sum_{f,t}\left(S_{ft}\log\big(\sum_b D_{fb}E_{bt}\big) - \Big(\sum_b D_{fb}E_b n\Big) - \log\Gamma(S_{ft}+1)\right) \tag{3.22}$$

which leads to the following exact posterior equation:

$$\begin{aligned} \log p(\mathbf{s}|\mathbf{S}, \mathbf{D}, \mathbf{E}) &= \log p(\mathbf{S}, \mathbf{s}|\mathbf{D}, \mathbf{E}) - \log p(\mathbf{S}|\mathbf{D}, \mathbf{E}) \\ &= \sum_{f,t}\Bigg\{\log\delta(S_{ft} - \sum_b s_{fbt}) + \sum_b s_{fbt}\log(D_{fb}E_{bt}) - \sum_b (D_{fb}E_{bt}) \\ &\quad - \sum_b \log\Gamma(s_{fbt}+1) - S_{ft}\log\sum_b (D_{fb}E_{bt}) + \Big(\sum_b D_{fb}E_{bt}\Big) \\ &\quad + \log\Gamma(S_{ft}+1)\Bigg\} \end{aligned}$$

$$\begin{aligned}\log p(\mathbf{s}|\mathbf{S},\mathbf{D},\mathbf{E}) &= \sum_{f,t}\Big\{\log\delta(S_{ft}-\sum_b s_{fbt}) + \log\Gamma(S_{ft}+1) - \sum_b \log\Gamma(s_{fbt}+1) \\ &\quad + \sum_b s_{fbt}\log D_{fb}E_{bt} - \sum_b s_{fbt}\log(\sum_b D_{fb}E_{bt})\Big\} \\ &= \sum_{f,t}\Big\{\log\delta(S_{ft}-\sum_b s_{fbt}) + \log\Gamma(S_{ft}+1) - \sum_b \log\Gamma(s_{fbt}+1) \\ &\quad + \sum_b s_{fbt}\log\frac{D_{fb}E_{bt}}{\sum_b D_{fb}E_{bt}}\Big\}\end{aligned}$$

Introducing the posterior probability variable for the latent source, s_{fbt} as:

$$p_{fbt} = \frac{D_{fb}E_{bt}}{\sum_b D_{fb}E_{bt}}, \qquad \sum_b p_{fbt} = 1 \tag{3.23}$$

we see that the resulting posterior of the previous equation is a type of Multinomial distribution as,

$$\log p(s|\mathbf{S},\mathbf{D},\mathbf{E}) = \sum_{f,t}\log\mathcal{M}(s_{f(1:B)t}; p_{f(1:B)t}) \tag{3.24}$$

which is expressed as

$$\mathcal{M}(s;S,p) = \binom{S}{s_1 s_2 \ldots s_B} p_1^{s_1}p_2^{s_2}\ldots p_B^{s_B}\delta(S-\sum_b s_b) = \delta(S-\sum_b s_b)S!\prod_b \frac{p_b^{s_b}}{s_b!} \tag{3.25}$$

Finally, we can use the standard result of the marginal mean:

$$\langle s_b\rangle = Sp_b \tag{3.26}$$

M step: We can calculate the expectation of the joint likelihood under the posterior

distribution, Q, to be optimized

$$\begin{aligned} Q &= \langle \log p(\mathbf{S}, s|\mathbf{D}, \mathbf{E}) \rangle_{p(s|\mathbf{S},\mathbf{D},\mathbf{E})} \\ &= \sum_{f,t} \Bigg\{ \langle \log \delta(S_{ft} - \sum_b s_{fbt}) \rangle + \sum_b \Big[\langle s_{fbt} \rangle \log(D_{fb} E_{bt}) - (D_{fb} E_{bt}) \\ &\quad - \Big\langle \log \Gamma(s_{fbt} + 1) \Big\rangle \Big] \Bigg\} \end{aligned}$$

From (3.23) and (3.26), we know that

$$\langle s_{fbt} \rangle = S_{ft} \frac{D_{fb} E_{bt}}{\sum_b D_{fb} E_{bt}} \tag{3.27}$$

Finally, taking the derivatives of Q with respect to each element of D_{fb} and E_{bt}, we find the following update rules:

$$\frac{\partial Q}{\partial D_{fb}} = -\sum_t E_{bt} + \frac{\sum_t \langle s_{fbt} \rangle}{D_{ft}} = 0 \tag{3.28}$$

$$D_{fb} = \frac{\sum_t \langle s_{fbt} \rangle}{\sum_t E_{bt}} = D_{fb} \frac{\frac{\sum_t S_{ft} E_{bt}}{\sum_b D_{fb} E_{bt}}}{\sum_t E_{bt}} \tag{3.29}$$

$$\frac{\partial Q}{\partial E_{bt}} = -\sum_f D_{fb} + \frac{\sum_f \langle s_{fbt} \rangle}{E_{bt}} = 0 \tag{3.30}$$

$$E_{bt} = \frac{\sum_f \langle s_{fbt} \rangle}{\sum_f D_{fb}} = E_{bt} \frac{\frac{\sum_f D_{fb} S_{ft}}{\sum_b D_{fb} E_{bt}}}{\sum_f D_{fb}} \tag{3.31}$$

Comparing Equations (3.3 and 3.4) with (3.29 and 3.31) of the classic NMF, we see that they are equivalent. This is because there is a correspondence between choosing the generalized KL divergence measure in the classic setting and choosing Poisson distribution in the statistical perspective [5].

3.2.2. IS-NMF

In IS-NMF model, we factorize the data matrix as a multiplication of two non-negative matrices as

$$\mathbf{P} \approx \mathbf{DE} \tag{3.32}$$

where P represents the power spectrogram of the signal to be modeled. IS-NMF model assumes that for each time-frequency entry, the complex spectrum of the signal (S) is generated by a number of complex Gaussian sources whose variances depends on the corresponding row of the template matrix and corresponding column of the excitation matrix. By modeling complex spectrum of the signal, maximization of the likelihood of the complex spectrum of the signal with complex Gaussian sources corresponds to minimization of the IS divergence between the power spectrogram of the signal with its approximation [6]. We can show this mathematically as,

$$s_{fbt} \sim N_c(s_{fbt}; 0, D_{fb} E_{bt}) \tag{3.33}$$

$$S_{ft} = \sum_{b=1}^{B} s_{fbt} \tag{3.34}$$

$$s_{ft} \sim N_c(s_{ft}; 0, \sum_b D_{fb} E_{bt}) \tag{3.35}$$

We call the variables $s_i = \{s_{fbt}\}$ as the latent sources. The probabilistic graphical model which represents the generative process for the NMF model is shown in Figure 3.2. We can analytically marginalize out the latent sources $s = \{s_1, \ldots, s_B\}$ to obtain the marginal likelihood

$$\log p(\mathbf{S}|\mathbf{D}, \mathbf{E}) = \log \sum_s p(\mathbf{S}|s)p(s|\mathbf{D}, \mathbf{E}) = \log \prod_{ft} N_c(S_{ft}; 0, \sum_b D_{fb}E_{bt}) \tag{3.36}$$

$$\log p(\mathbf{S}|\mathbf{D}, \mathbf{E}) = \sum_{f,t} \Big(-\log(\sum_b D_{fb}E_{bt}) - \frac{|S_{ft}|^2}{\sum_b D_{fb}E_{bt}} \Big) \tag{3.37}$$

E step: First let us write the joint distribution

$$\log p(\mathbf{s}, \mathbf{S}|\mathbf{D}, \mathbf{E}) = \sum_{f,t} \Big\{ \sum_b \big[-\log(D_{fb}E_{bt}) - \frac{|s_{fbt}|^2}{D_{fb}E_{bt}} \big] + \log \delta(S_{bt} - \sum_b D_{fb}E_{bt}) \Big\} \tag{3.38}$$

Posterior distribution of the latent sources:

$$\begin{aligned} p(s_{fbt}|S_{ft}, D, E) &= N_c(s_{fbt}; \mu_{fbt}, \Sigma_{fbt}) && (3.39)\\ \Sigma_{fbt} &= \frac{D_{fb}E_{bt}}{\sum_b D_{fb}E_{bt}} \sum_{i \neq b} D_{fi}E_{it} && (3.40)\\ \mu_{fbt} &= \frac{D_{fb}E_{bt}}{\sum_b D_{fb}E_{bt}} S_{ft} && (3.41) \end{aligned}$$

Marginal Expectation of the latent sources:

$$\begin{aligned} \langle |s_{fbt}|^2 \rangle &= \Sigma_{fbt} + |\mu_{fbt}|^2 && (3.42)\\ \langle |s_{fbt}|^2 \rangle &= \frac{D_{fb}E_{bt}}{\sum_b D_{fb}E_{bt}} \sum_{i \neq b} D_{fi}E_{it} + (\frac{D_{fb}E_{bt}}{\sum_b D_{fb}E_{bt}})^2 P_{ft} && (3.43) \end{aligned}$$

M step: We can calculate the expectation of the joint likelihood under the posterior distribution, Q, to be optimized

$$Q =^c \sum_{f,t} \Big\{ \sum_b \big[-\log(D_{fb}E_{bt}) - \frac{\langle |s_{fbt}|^2 \rangle}{D_{fb}E_{bt}} \big] \Big\} \tag{3.44}$$

$$\frac{\partial Q}{\partial D_{fb}} = \sum_t (-\frac{1}{D_{fb}} - \frac{\langle |s_{fbt}|^2 \rangle}{D_{fb}^2} E_{bt}) = 0 \tag{3.45}$$

$$D_{fb} = \frac{1}{T} \sum_t \frac{\langle |s_{fbt}|^2 \rangle}{E_{bt}} \tag{3.46}$$

$$\frac{\partial Q}{\partial E_{bt}} = \sum_f (-\frac{1}{E_{bt}} - \frac{\langle |s_{fbt}|^2 \rangle}{E_{bt}^2 D_{fb}}) = 0 \tag{3.47}$$

$$E_{bt} = \frac{1}{F} \sum_f \frac{\langle |s_{fbt}|^2 \rangle}{D_{fb}} \tag{3.48}$$

Comparing Equations (3.6 and 3.7) with (3.46 and 3.48) of the classic NMF, we see that they are not equivalent. This is because there is not a correspondence between choosing the generalized IS divergence measure in the classic setting and choosing complex Gaussian distribution in the statistical perspective [6].

3.3. Speech-Music Separation with NMF Models

In NMF based speech-music separation systems, during training phase, the magnitude (KL Case) or power (IS Case) spectrum of the speech and music signals are used to train an NMF model for each source as:

$$\mathbf{S} = \mathbf{D}^s \mathbf{E}_t^s \quad \text{and} \quad \mathbf{M} = \mathbf{D}^m \mathbf{E}_t^m. \tag{3.49}$$

The template and excitation matrices can be calculated via multiplicative update rules [42] efficiently. In the separation phase, by concatenating the individual template matrices, $\mathbf{D}^s$ and $\mathbf{D}^m$, an overall pre-trained template matrix is obtained and used as a model. Using the magnitude or power spectrum of the mixed signal as the observation

signal representation, the excitation matrix for each source is calculated by solving the equation

$$\mathbf{X} = [\mathbf{D}^s \ \mathbf{D}^m][(\mathbf{E}^s)^T \ (\mathbf{E}^m)^T]^T \tag{3.50}$$

where $\mathbf{E}^s$ and $\mathbf{E}^m$ represents the excitation matrix for speech and music sources in the mixed signal respectively. After finding the excitation matrix for each source, the reconstruction of the speech and music signals can be done using estimated intensity or variance parameters of the speech and music sources and the observation values. We can estimate the reconstructed spectrum as the joint posterior of the source signals as

$$(\widehat{\mathbf{S}}, \widehat{\mathbf{M}}) = \arg\max_{\mathbf{S},\mathbf{M}} p(\mathbf{S}, \mathbf{M}|\mathbf{X}, \mathbf{D}^s, \mathbf{E}^s, \mathbf{D}^m, \mathbf{E}^m).$$

This corresponds to the estimation of the magnitude or power spectrum of the speech and music sources as

$$\widehat{\mathbf{S}} = \mathbf{X} \otimes \frac{\mathbf{D}^s\mathbf{E}^s}{(\mathbf{D}^s\mathbf{E}^s + \mathbf{D}^m\mathbf{E}^m)}. \tag{3.51}$$

$$\widehat{\mathbf{M}} = \mathbf{X} \otimes \frac{\mathbf{D}^m\mathbf{E}^m}{(\mathbf{D}^s\mathbf{E}^s + \mathbf{D}^m\mathbf{E}^m)}. \tag{3.52}$$

where all matrix divisions are element-wise. This is also known as the Wiener filtering approach and was used in NMF based speech-music separation in [18]. Since NMF methods find an approximation to the magnitude or power spectrogram of the mixed signal, the error term between the approximated and the truth values of the spectrograms is not assigned to any source signals (speech or music). This problem can be solved by estimating the source spectrograms jointly using the mixed signal spectrogram. This enables the perfect reconstruction of the target source signals. The overall NMF based speech-music separation method is shown in Figure 3.3.

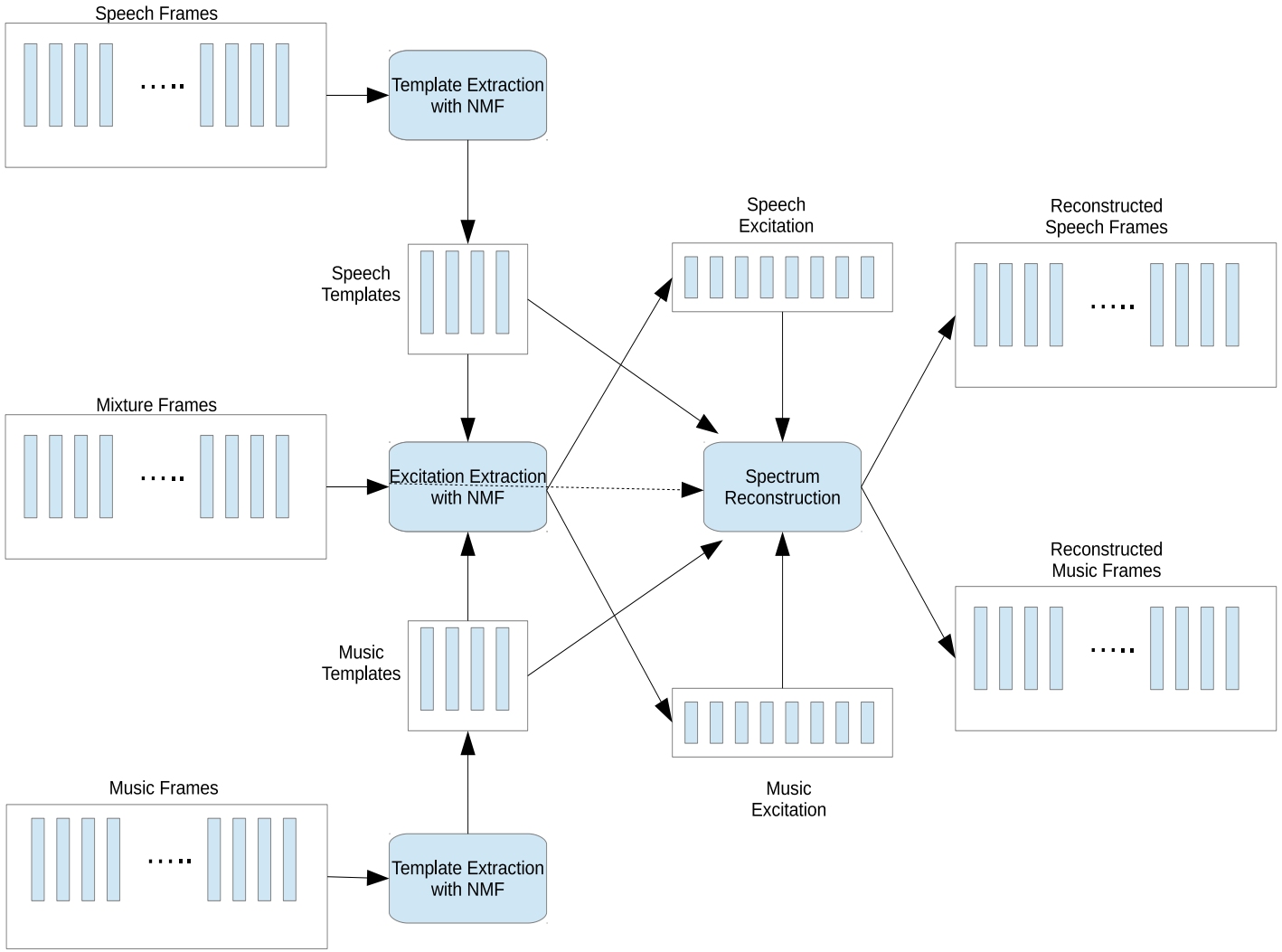

Figure 3.3. NMF based speech-music separation system.

The steps of the separation process in the case of previously trained template matrices for the speech and music sources can be summarized for KL and IS divergences as follows:It should be noted that magnitude or power spectrum of the mixed signal are used in the separation process for KL and IS divergences respectively.

Summary of KL Divergence Based Speech-Music Separation:

(i) Compute the posterior cell probability of the speech source:

$$p^s_{fbt} = \frac{D^s_{fb} E^s_{bt}}{\sum_b D^s_{fb} E^s_{bt} + \sum_k D^m_{fk} V^m_{kt}} \tag{3.53}$$

(ii) Compute the sufficient statistics for the speech source:

$$\langle s_{fbt} \rangle = X_{ft} p^s_{fbt} \tag{3.54}$$

(iii) Compute the excitation matrix for the speech signal:

$$E^s_{bt} = \frac{\sum_f \langle s_{fbt} \rangle}{\sum_f D^s_{fb}} \tag{3.55}$$

(iv) Compute the posterior cell probability of the music source:

$$p^m_{fkt} = \frac{D^m_{fk} E^m_{kt}}{\sum_b D^s_{fb} E^s_{bt} + \sum_k D^m_{fk} V^m_{kt}} \tag{3.56}$$

(v) Compute the sufficient statistics for the music source:

$$\langle m_{fkt} \rangle = X_{ft} p^m_{fkt} \tag{3.57}$$

(vi) Compute the excitation matrix for the music signal:

$$E^m_{kt} = \frac{\sum_f \langle m_{fkt} \rangle}{\sum_f D^m_{fk}} \tag{3.58}$$

(vii) Iterate over Equation 3.53 to Equation 3.58

(viii) Reconstruct the source signals using Equations 3.51 and 3.52.

Summary of IS Divergence Based Speech-Music Separation:

(i) Compute the posterior mean and variance of the speech source:

$$\Sigma_{fbt} = \frac{D^s_{fb}E^s_{bt}}{\sum_b D^s_{fb}E^s_{bt} + \sum_b D^m_{fk}E^m_{kt}}\left(\sum_{i\neq b} D^s_{fi}E^s_{it} + \sum_k D^m_{fk}E^m_{kt}\right) \tag{3.59}$$

$$\mu_{fbt} = \frac{D^s_{fb}E^s_{bt}}{\sum_b D^s_{fb}E^s_{bt} + \sum_{b'} D^m_{fb'}E^m_{b't}}X_{ft} \tag{3.60}$$

(ii) Compute the sufficient statistics for the speech source:

$$\langle |s_{fbt}|^2 \rangle = \Sigma_{fbt} + |\mu_{fbt}|^2 \tag{3.61}$$

(iii) Compute the excitation matrix for the speech signal:

$$E^s_{bt} = \frac{1}{F}\sum_f \frac{\langle |s_{fbt}|^2 \rangle}{D^s_{fb}} \tag{3.62}$$

(iv) Compute the posterior mean and variance of the music source:

$$\Sigma_{fkt} = \frac{D^m_{fk}E^m_{kt}}{\sum_b D^s_{fb}E^s_{bt} + \sum_k D^m_{fk}E^m_{kt}}\left(\sum_b D^s_{fb}E^s_{bt} + \sum_{i\neq k} D^m_{fi}E^m_{it}\right) \tag{3.63}$$

$$\mu_{fkt} = \frac{D^m_{fk}E^m_{kt}}{\sum_b D^s_{fb}E^s_{bt} + \sum_k D^m_{fk}E^m_{kt}}X_{ft} \tag{3.64}$$

(v) Compute the sufficient statistics for the music source:

$$\langle |m_{fkt}|^2 \rangle = \Sigma_{fkt} + |\mu_{fkt}|^2 \tag{3.65}$$

(vi) Compute the excitation matrix for the music signal:

$$E_{kt}^{m} = \frac{1}{F} \sum_{f} \frac{\langle |m_{fkt}|^2 \rangle}{D_{fk}^{m}} \tag{3.66}$$

(vii) Iterate over Equation 3.59 to Equation 3.66

(viii) Reconstruct the source signals using Equations 3.51 and 3.52.

3.4. Experimental Results

The ultimate goal of the speech-music separation in our study is to increase the ASR performance. Therefore, we analyze the performance of the method using ASR performance measure, word accuracy Rate (WAcc) which is defined as 100-WER. However, in order to relate the separation quality which characterizes the separation performance to ASR tasks, we also report SMR and SAR values as proposed in [48].

3.4.1. Speech Recognition System and Test Set

For speech recognition tests, we used a CMU-Sphinx HMM-based continuous density speech recognizer which is trained to recognize Turkish Broadcast News speech which is sampled at 16 kHz. The gender-dependent acoustic models are trained using Mel-frequency cepstral coefficients (MFCCs) and their deltas and double-deltas calculated in 25ms frames. The test set contains 240 utterances distributed approximately uniformly across 8 speakers. The total length of the test set is about 70 minutes and the average length of the utterances is about 18 seconds. The test system is summarized in Figure 3.4.

The test utterances are mixed with 10 different jingles at $0, 5, 10, 15$ and 20 dB SMR levels to create the test set. The average length of the jingles is 7 seconds. The background music signal is generated by repeating the jingles up to the length of the speech. The jingles are taken from real broadcast news jingles. In this study, we assume, which jingle is used to generate the background music is known as a prior.

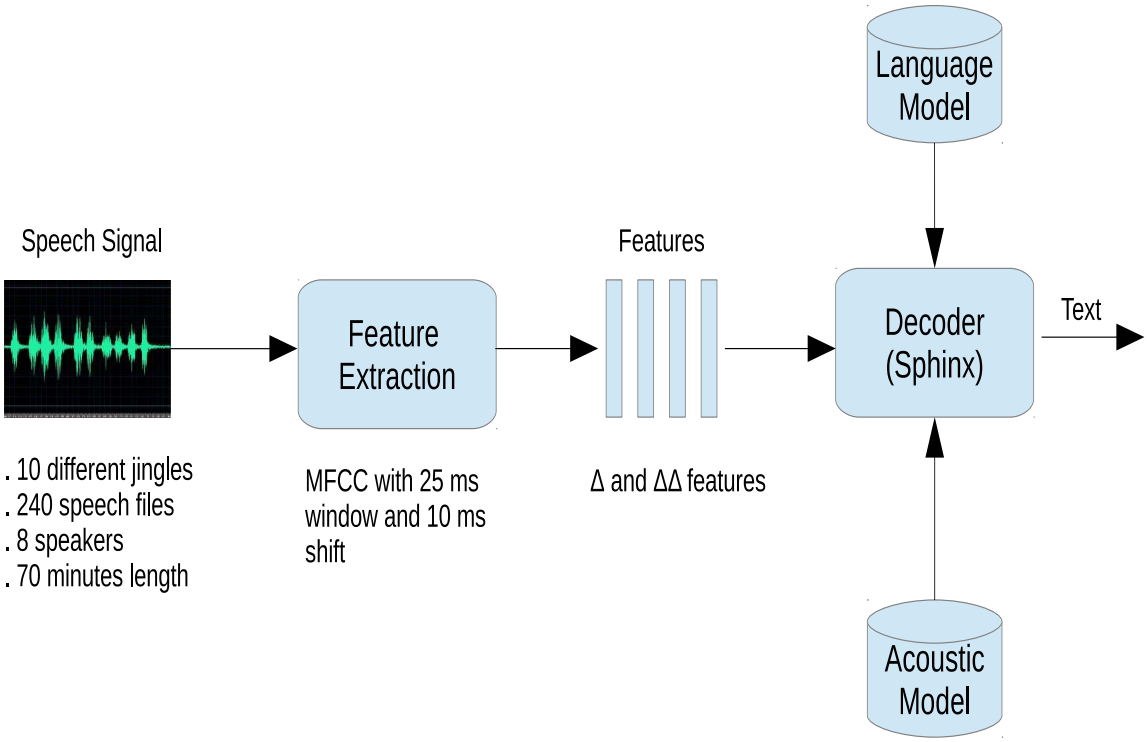

Figure 3.4. ASR system and test set for NMF based separation.

WAcc values of the clean speech data and the mixed data without any separation method are shown in Table 3.1 The magnitude or power spectrum are computed using 1024-point length frames and 512 point frame shift is used.

Table 3.1. Baseline WAcc values.

Baseline	Input SMR Values				
Results	0dB	5dB	10dB	15dB	20dB
Clean	75.7	75.7	75.7	75.7	75.7
Mixed	1.5	6.5	25.0	47.5	64.4

3.4.2. Training Data and Models

Four types of speech and music data set are used to train the NMF models for the speech and music signals.

For speech signal, the different models can be listed as follows and the properties of the models are summarized in Table 3.2 and shown in Figure 3.6. The number of

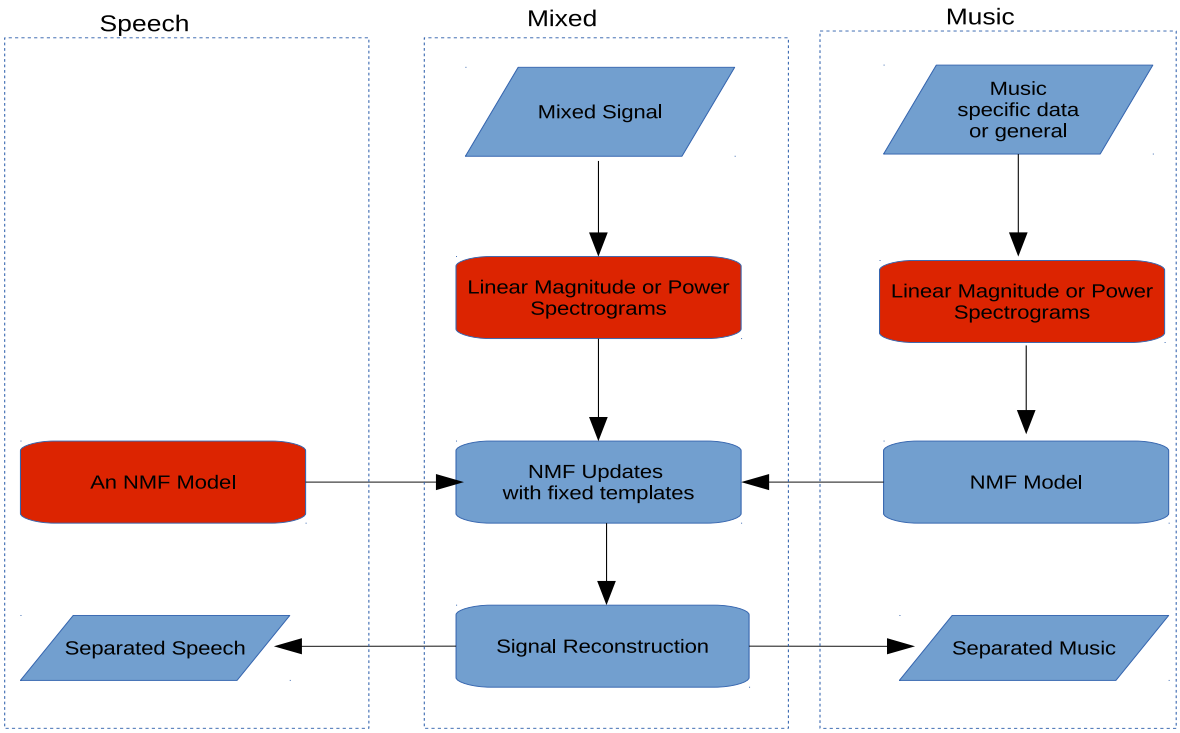

Figure 3.5. NMF based speech-music separation with 'None' model.

bases for both sources are chosen by making many different trials.

- 'Self' case refers to the training data of the target speaker which is the same as the mixed signal which has to be separated.
- 'Other' case refers to the training data from 3 different people who are from the same gender as the target speaker.
- 'All' case refers to the training data from target speaker and 3 people with the same gender as the target speaker.
- 'None' case refers to no training data is used for modeling the speech signal. In this case, we use an NMF model with untrained template matrix. The template and excitation matrices are estimated from the mixed signal, simultaneously. Speech-music separation method with 'None' speech model is presented in Figure 3.5.

For music signal, the training models can be listed as follows and summarized in Table 3.3 and shown in Figure 3.7.

Table 3.2. Speech training data set properties.

Data Set	# of Speakers	Definition of the set	Length (min.)	# of Bases
Self	1	The same speaker	2	30
All	4	Including Speaker	8	30
Other	3	Excluding Speaker	6	30
None	0	No speech data	0	30

- 'Original' case refers to the jingle itself which is used to create the background music signal of the mixed signal. In 'Original' case, the frames of the jingle are used as the template vectors of NMF model.
- 'Self' case refers to the jingle itself which is used to create the background music signal of the mixed signal. The templates are trained from the jingle frames.
- 'Other' case refers to the training data from 9 different jingle which are not used in background music generation.
- 'All' case refers to the jingles which includes the jingle that is used for the background music generation.

Table 3.3. Music training data set properties.

Data Set	# of Jingle	Definition of the set	Length (sec.)	# of Bases
Original	1	The same jingle	7	# of frames in the jingle
Self	1	The same jingle	7	30
All	10	Including jingle	120	30
Other	9	Excluding jingle	116	30

3.4.3. Experimental Analysis

In order to analyze the effect of 16 different combination of training data sets for speech and music signals, SMR, SAR and WAcc values are represented in Tables 3.4, 3.5,3.6, 3.7, 3.8 and 3.9.

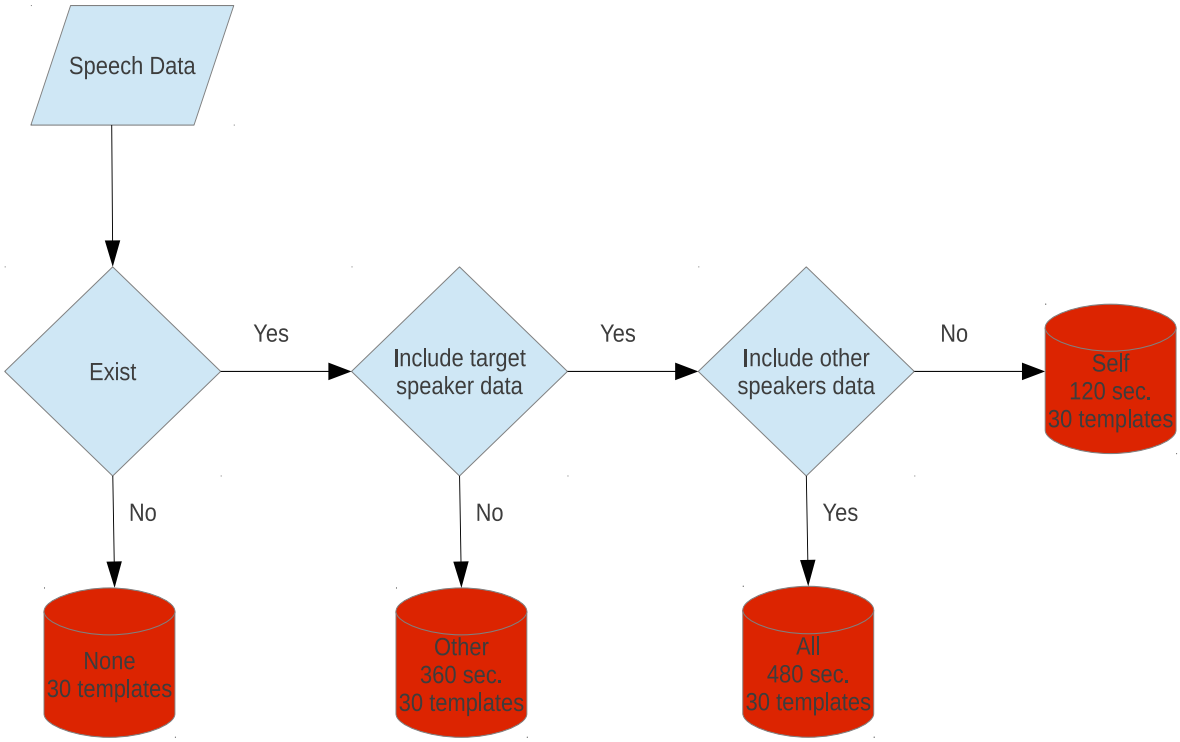

Figure 3.6. Training data types for speech signal with NMF methods.

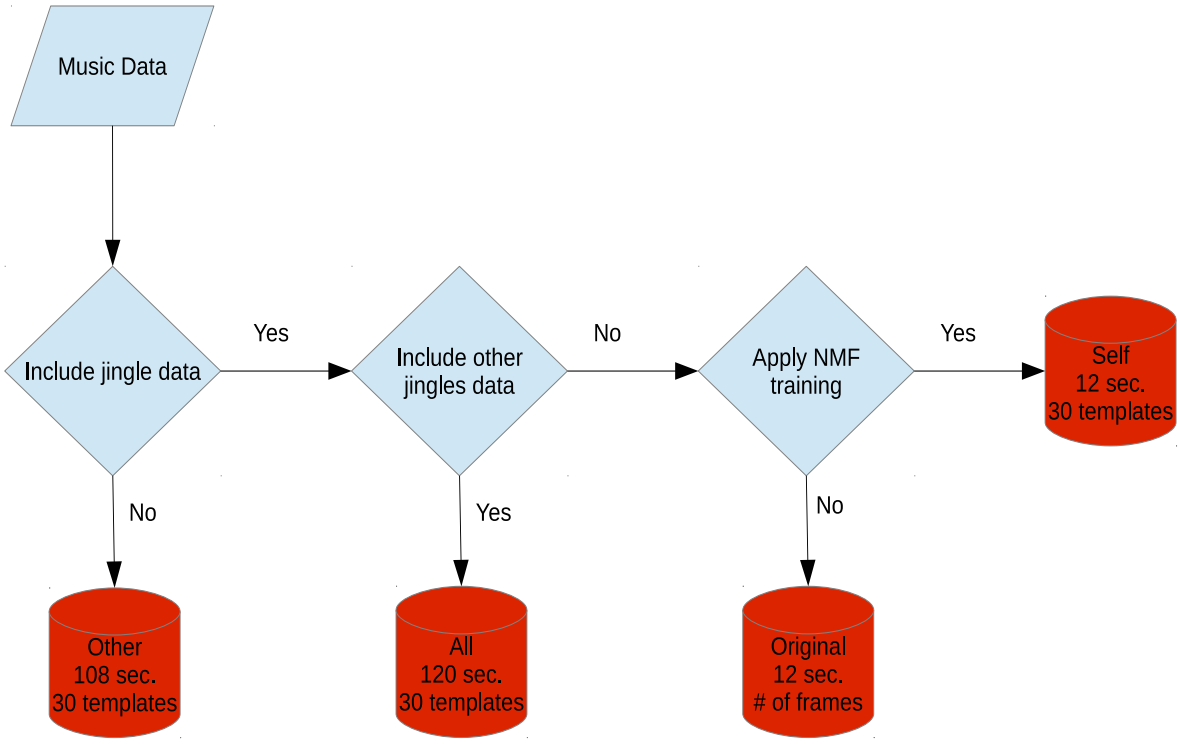

Figure 3.7. Training data types for music signal with NMF methods

KL SMR Value Analysis:

(i) For all of the speech models, 'Original' model for the music signal generates higher SMR Values than other models ('Self', 'All' and 'Other').

(ii) For all of the speech models except 'None' model, 'All' and 'Self' models for the music signal gives similar SMR values.

(iii) For 'None' model for the speech signal, 'All' and 'Other' models for the music signal yields similar SMR values.

Table 3.4. Output SMR values of KL-NMF methods with different training data sets.

Output SMR (dB)		Input SMR (dB)				
Speech	Music	0dB	5dB	10dB	15dB	20dB
None	Other	2.1	13.6	23.9	35.5	45.4
	All	2.9	14.7	26.4	36.9	46.5
	Self	9.9	19.6	32.4	38.0	47.0
	Original	17.9	25.4	38.7	41.4	49.9
Other	Other	8.3	17.7	26.2	35.9	44.8
	All	9.8	19.0	27.9	36.7	45.5
	Self	9.9	18.9	30.3	36.5	45.3
	Original	14.6	22.6	34.1	38.6	46.9
All	Other	8.4	17.9	26.4	36.1	45.0
	All	9.8	19.1	28.1	36.9	45.7
	Self	10.0	19.1	30.5	36.8	45.5
	Original	14.9	22.9	34.5	39.0	47.3
Self	Other	9.6	18.8	27.2	36.6	45.4
	All	11.2	20.2	28.9	37.5	46.1
	Self	11.0	19.9	31.2	37.2	45.8
	Original	15.3	23.2	34.5	39.0	47.2

KL SAR Value Analysis:

(i) For all of the speech models, 'Original' and 'Self' models for the music signal generates higher SAR Values than other models ('All' and 'Other').

(ii) For 'None' model for the speech signal, 'All' and 'Other' models for the music signal yields similar SAR values. Even, for high input SMR values, 'Other' model gives better SAR values than 'All' model.

Table 3.5. Output SAR values of KL-NMF methods with different training data sets.

Output SAR (dB)		Input SMR (dB)				
Speech	Music	0dB	5dB	10dB	15dB	20dB
None	Other	8.2	10.0	12.2	14.9	16.7
	All	8.1	9.8	12.2	14.2	15.7
	Self	10.0	12.0	15.1	16.5	18.4
	Original	9.2	11.2	14.5	15.9	17.8
Other	Other	10.3	12.6	14.5	16.7	18.3
	All	10.3	12.7	14.5	16.2	17.5
	Self	10.7	13.1	16.0	17.7	19.8
	Original	10.8	13.1	16.2	18.0	20.2
All	Other	10.2	12.7	14.7	16.9	18.6
	All	10.3	12.9	14.9	16.7	18.2
	Self	10.7	13.2	16.3	18.0	20.2
	Original	10.9	13.3	16.4	18.2	20.6
Self	Other	9.9	12.2	14.0	16.0	17.5
	All	10.0	12.2	14.0	15.7	17.0
	Self	10.5	12.8	15.6	17.3	19.3
	Original	10.6	12.9	15.9	17.6	19.8

KL WAcc Value Analysis:

(i) With 'None' speech model (see Figure 3.8), 'All' and 'Other' music models gives worse speech recognition performance at '20dB' input SMR value. 'None' model provides recognition improvements with 'Self' and 'Original' models. For all of the other model combinations and input SMR values, the separation method improves the recognition performance as compared to the no separation case.

(ii) For all of the speech models, 'Original' model for the music signal generates higher WAcc Values than other models ('Self','All' and 'Other'). 'Original' model

provides better recognition results as compared to 'Self' model. See Figure 3.9.

(iii) For 'Original' model for the speech signal, 'Self', 'All' and 'Other' models for the music signal yields similar WAcc values (see Figure 3.10). Although using a speech model improves the recognition accuracy, all type of the speech models provides similar performance.

Table 3.6. Output WAcc values of KL-NMF methods with different training data sets.

WAcc Values (%)		Input SMR (dB)				
Speech	Music	0dB	5dB	10dB	15dB	20dB
None	Other	1.2	7.2	21.3	42.2	54.5
	All	1.4	8.0	25.5	44.1	55.3
	Self	10.7	24.6	49.2	54.5	64.5
	Original	17.2	28.0	51.1	53.3	61.3
Other	Other	9.9	25.3	45.1	62.7	70.8
	All	11.5	28.5	50.3	64.3	71.1
	Self	14.3	31.6	58.8	64.4	71.9
	Original	27.5	43.0	62.8	66.5	71.4
All	Other	9.0	26.8	45.4	63.6	70.4
	All	11.3	29.0	50.4	65.3	71.4
	Self	14.2	31.5	59.9	64.0	71.6
	Original	28.1	43.6	63.8	67.4	72.0
Self	Other	9.4	25.1	45.0	60.3	68.0
	All	11.1	28.2	48.7	61.5	68.9
	Self	14.5	31.9	57.7	62.2	69.6
	Original	27.5	41.2	61.3	63.9	69.6

IS SMR Value Analysis:

(i) For all of the speech models, 'Original' model for the music signal generates higher SMR Values than other models ('Self', 'All' and 'Other').

(ii) For all of the speech models except 'None' model, 'All' and 'Self' models for the music signal gives similar SMR values.

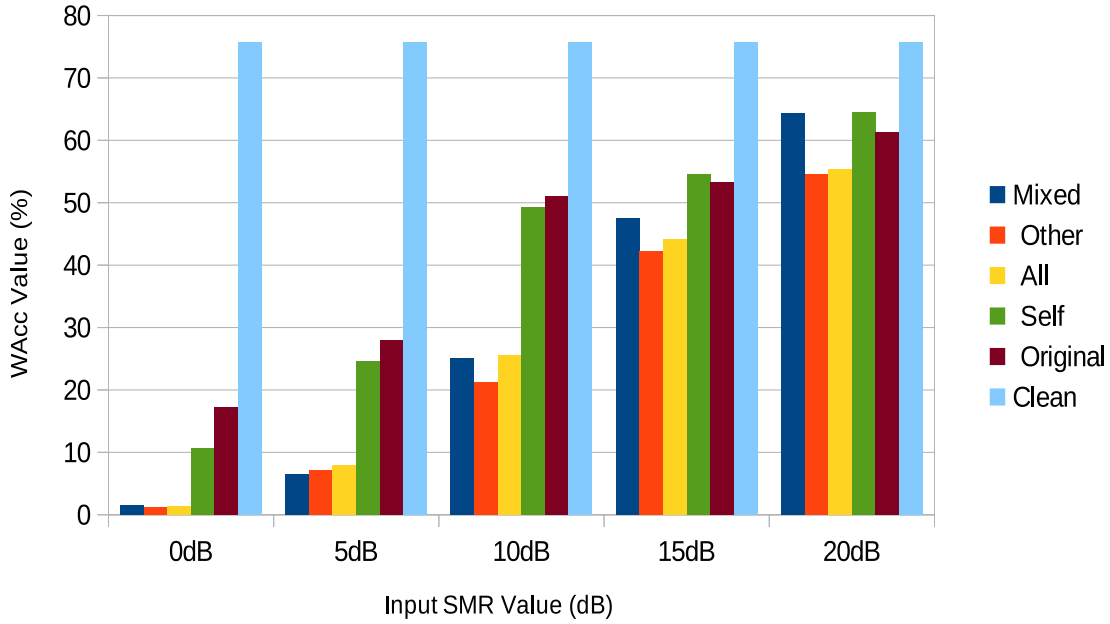

Figure 3.8. ASR result with 'None' speech model.

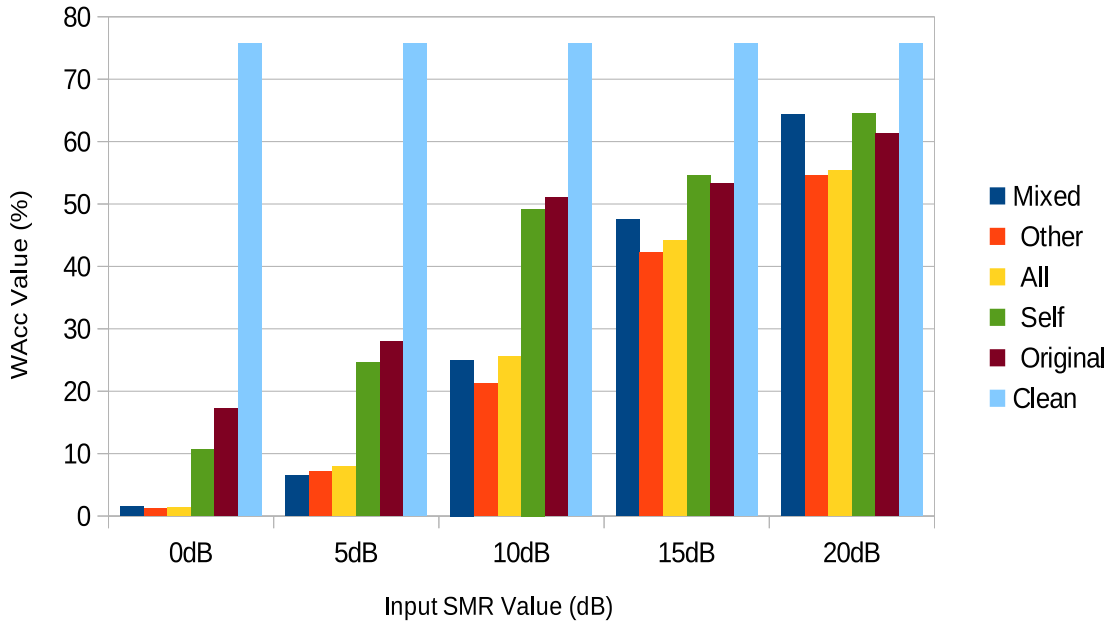

Figure 3.9. ASR result with 'Original' and 'Self' music models.

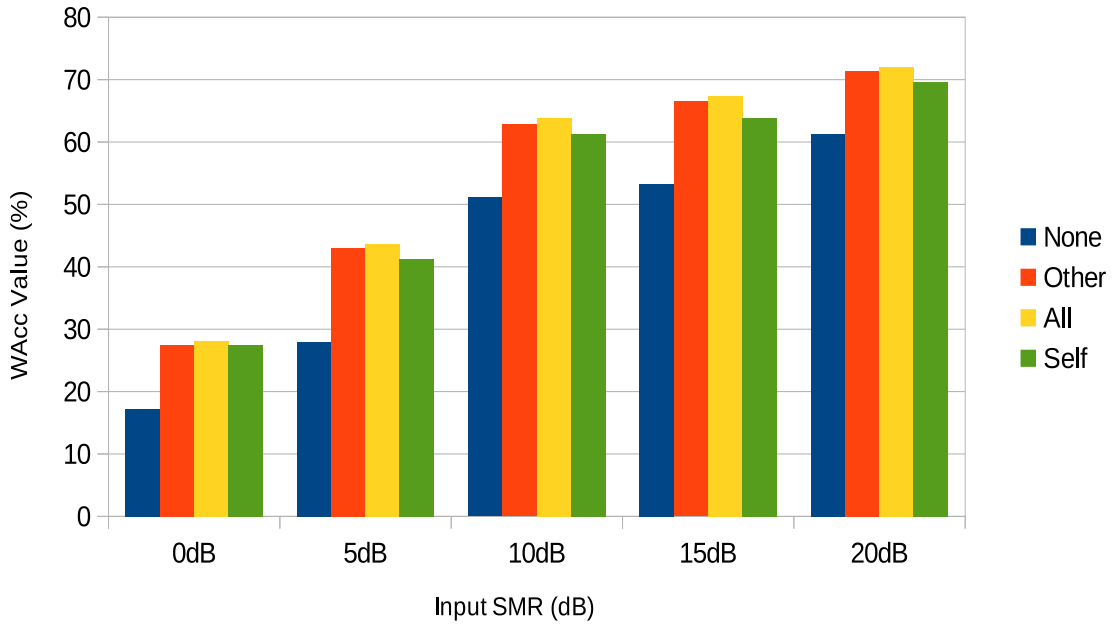

Figure 3.10. ASR result with speech data types with 'Original' music model.

(iii) For 'None' model for the music signal, 'All' and 'Other' models for the music signal yields similar SMR values.

IS SAR Value Analysis:

(i) For all of the speech models, 'Original' and 'Self' models for the music signal generates higher SAR Values than other models ('All' and 'Other')and their SAR values are close to each other.

(ii) For 'None' model for the music signal, 'All' and 'Other' models for the music signal yields similar SAR values.

IS WAcc Value Analysis:

(i) For all of the model combinations and input SMR values, the separation method improves the recognition performance as compared to the no separation case.

(ii) For all of the speech models, 'Original' model for the music signal generates higher WAcc Values than other models ('Self','All' and 'Other').

(iii) For 'Original' model for the music signal, 'Self','All' and 'Other' models for the

Table 3.7. Output SMR values of IS-NMF methods with different training data sets.

Output SMR (dB)		Input SMR (dB)				
Speech	Music	0dB	5dB	10dB	15dB	20dB
None	Other	1.9	13.0	22.6	33.6	43.4
	All	3.1	14.2	24.7	34.8	44.4
	Self	8.7	18.0	30.2	36.1	45.1
	Original	13.4	21.6	34.4	38.5	47.1
Other	Other	7.8	17.0	25.5	35.1	44.1
	All	9.0	18.1	26.9	35.8	44.7
	Self	9.0	17.9	29.0	35.5	44.4
	Original	12.2	20.3	31.6	36.9	45.5
All	Other	7.7	17.0	25.6	35.3	44.4
	All	9.0	18.2	27.2	36.1	45.0
	Self	9.1	18.1	29.3	35.8	44.7
	Original	12.6	20.7	32.2	37.3	45.9
Self	Other	8.5	17.5	25.9	35.4	44.4
	All	9.9	18.7	27.5	36.2	45.0
	Self	9.7	18.4	29.5	35.9	44.7
	Original	12.7	20.7	32.0	37.2	45.7

speech signal yields similar WAcc values.

(iv) For all of the speech models, the 'Self' model for the music signal gives better speech recognition performance than 'All' and 'Other' cases at $0, 5$ and 10 dB input SMR values.

Overall Experimental Results Analysis:

When we analyze the speech recognition performances of all of the model combinations at different input SMR values, it should be emphasized that IS-NMF outperforms KL-NMF methods. Another important observation is that, in higher input SMR values, the difference between the model combinations and methods turns out to be negligible. It should be noted that, in our experimental study, many different number of bases for

Table 3.8. Output SAR values of IS-NMF methods with different training data sets.

Output SAR (dB)		Input SMR (dB)				
Speech	Music	0dB	5dB	10dB	15dB	20dB
None	Other	6.8	9.6	12.7	16.6	19.8
	All	6.8	10.0	13.7	17.4	20.7
	Self	9.3	12.1	16.7	18.5	21.7
	Original	9.7	12.4	17.4	18.7	21.7
Other	Other	8.5	11.7	14.4	17.6	20.4
	All	8.8	12.0	14.8	17.6	20.1
	Self	9.4	12.3	16.3	18.4	21.6
	Original	10.3	13.0	17.3	19.0	22.0
All	Other	8.3	11.5	14.3	17.7	20.7
	All	8.6	11.8	14.7	17.7	20.4
	Self	9.3	12.3	16.4	18.6	21.8
	Original	10.2	13.1	17.5	19.1	22.2
Self	Other	8.6	11.7	14.3	17.5	20.4
	All	8.9	12.1	14.9	17.8	20.6
	Self	9.5	12.4	16.5	18.6	21.7
	Original	10.3	13.1	17.4	19.1	22.1

the both sources are used. However, it is concluded that except 'Original' case for the music signal, 30 template vectors are enough to model the sources.

Table 3.9. Output WAcc values of IS-NMF methods with different training data sets.

WAcc Values (%)		Input SMR (dB)				
Speech	Music	0dB	5dB	10dB	15dB	20dB
None	Other	1.4	9.8	26.8	50.6	62.6
	All	2.1	14.0	37.2	55.7	66.1
	Self	14.7	30.7	56.5	59.4	68.1
	Original	31.4	42.6	59.2	62.8	69.8
Other	Other	9.9	26.2	44.1	62.9	69.5
	All	12.3	28.8	50.2	64.1	70.7
	Self	17.4	34.1	61.4	64.6	72.0
	Original	39.6	49.2	64.3	67.2	72.2
All	Other	9.4	25.2	43.5	62.0	70.0
	All	11.8	28.8	50.9	64.5	70.3
	Self	16.5	33.9	60.4	64.8	71.3
	Original	39.3	49.1	64.6	67.1	72.2
Self	Other	11.0	26.3	45.0	62.1	69.0
	All	30.2	30.2	51.5	64.2	70.1
	Self	18.4	35.2	61.8	64.2	71.6
	Original	38.9	49.1	64.1	66.7	72.6

4. MUSIC MODELING FOR SPEECH-MUSIC SEPARATION

Background music removal is a serious problem for automatic broadcast news transcription systems. For such systems, it can be assumed that the background music is generated by using a catalog which contains the jingles used to create the background music signal. This scenario is shown in Figure 4.1. In other words, the background music is composed of the jingles in the catalog. In our proposed system, it is assumed that a speech-music segmentation system can partition incoming audio as speech, music and speech-music mixture. Hence, which jingle is used to create the background music can be detected using the music parts of the audio. The proposed system is shown in Figure 4.3. However, for this study, we assume, which jingle of the catalog is used

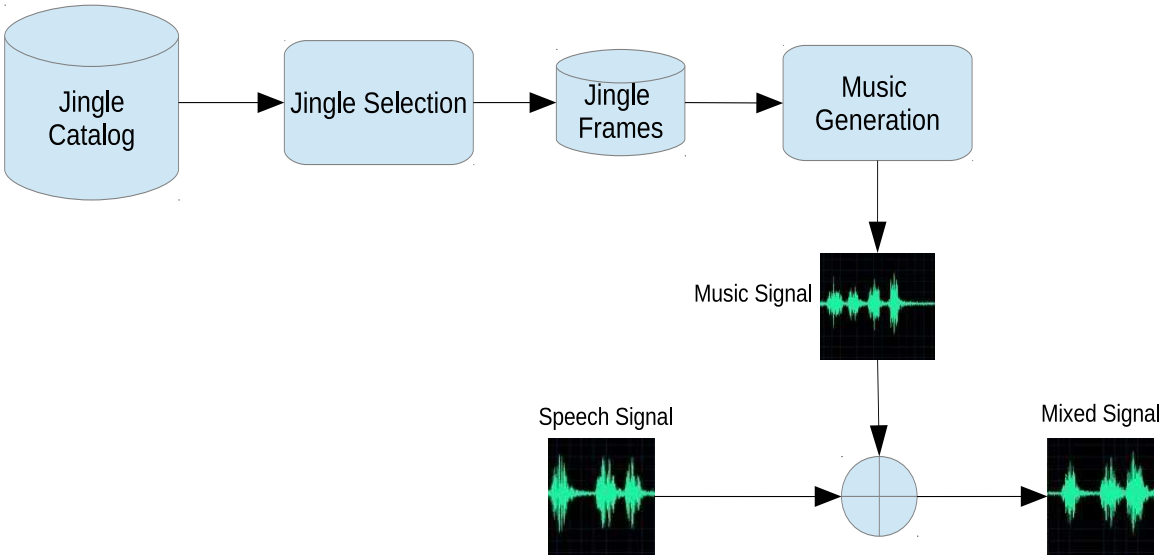

Figure 4.1. Background music generation scenario.

to create the background music is known as a prior. The framework for this scenario is shown in Figure 4.3. Although a prior speech model which can be trained using the speech part of the segmented audio can be used in the separation phase, up to the Chapter 5, any pre-trained speech model is not used in the separation process. In other words, the speech source is extracted from the mixed signal without any prior knowledge about the structure of the speech signal. The proposed separation strategy

is summarized in Figures 4.2 and 4.5. The difference of the proposed strategy from the traditional NMF based speech-music separation methods can be understood by comparing Figures 3.3 and 4.5.

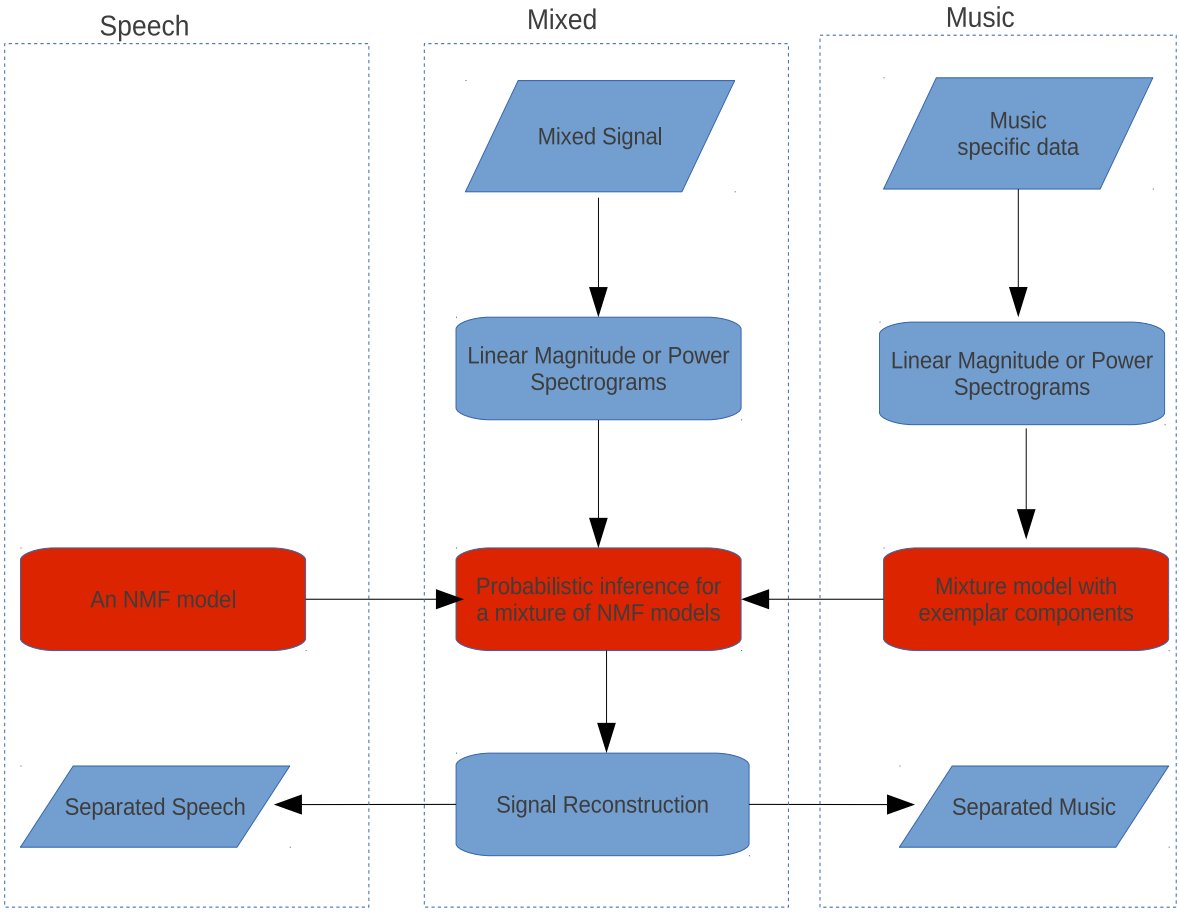

Figure 4.2. Mixture based speech-music separation method overview.

Since we assumed that for broadcast news problem, the background jingle is known, it is feasible to use the jingle itself in the separation process as a music model. The frames of jingle are used as the template vectors. The background music is assumed to be generated by using the random parts of the jingle. In other words, which part of the jingle is played for each time frame is not known. Moreover, a gain or frequency filtering can be applied to create the background music from the jingle frames. Therefore, there are two problems in this assumption:

- The jingle frame identity for each time frame must be estimated.

- Gain parameter (which corresponds to the volume change in the background music) or frequency filtering parameter must be estimated.

The generation process of the background music from the jingle is shown in Figure 4.4. The background music is generated using the jingle frames so it is feasible to use the jingle frames as the mixture components. It is necessary to find the identity of the jingle frame and the scaling parameters (gain or frequency) for each time frame to separate the speech from the background music signal.

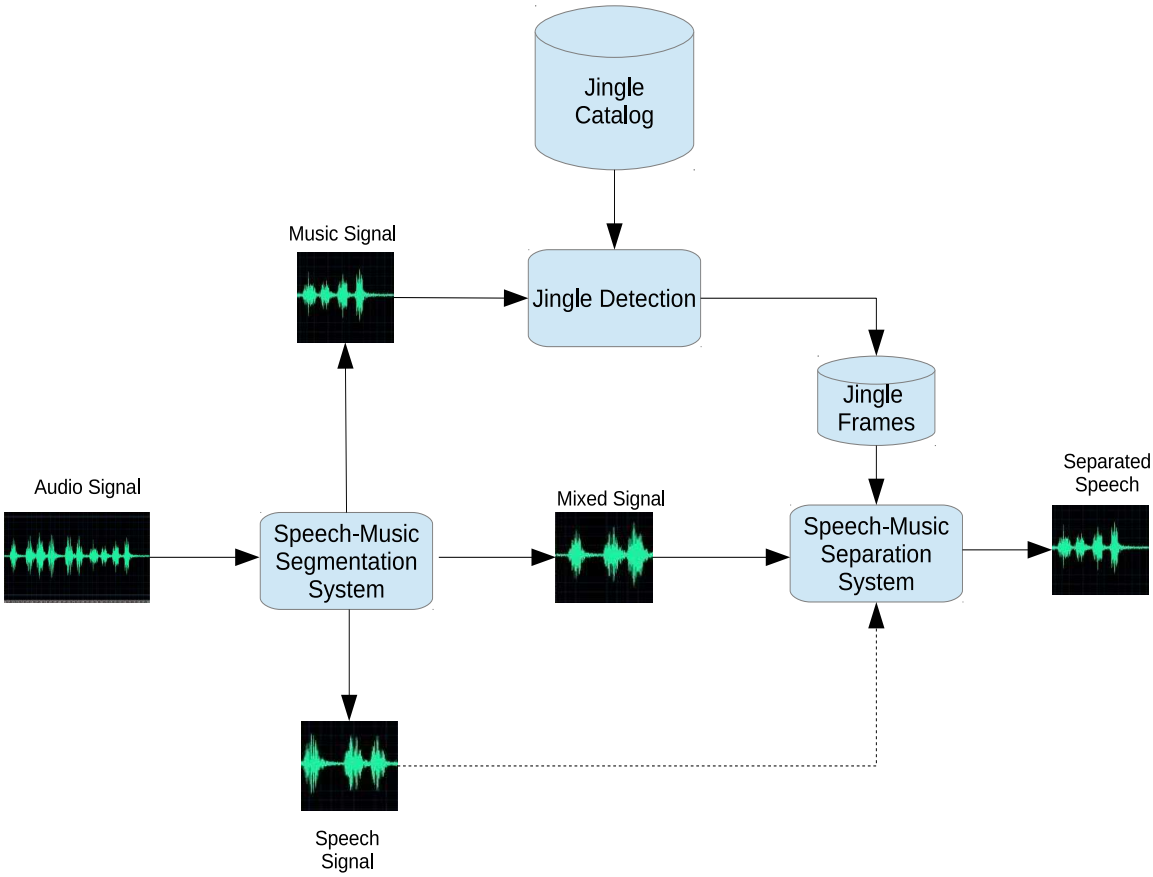

Figure 4.3. Catalog based speech-music separation system framework.

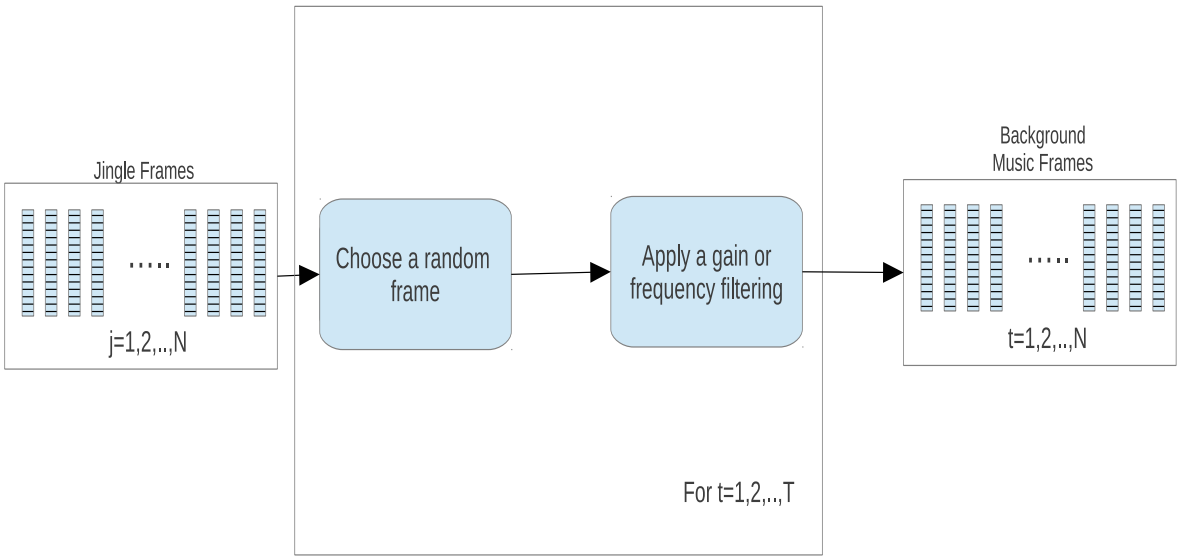

Figure 4.4. Background music generation process from jingle frames.

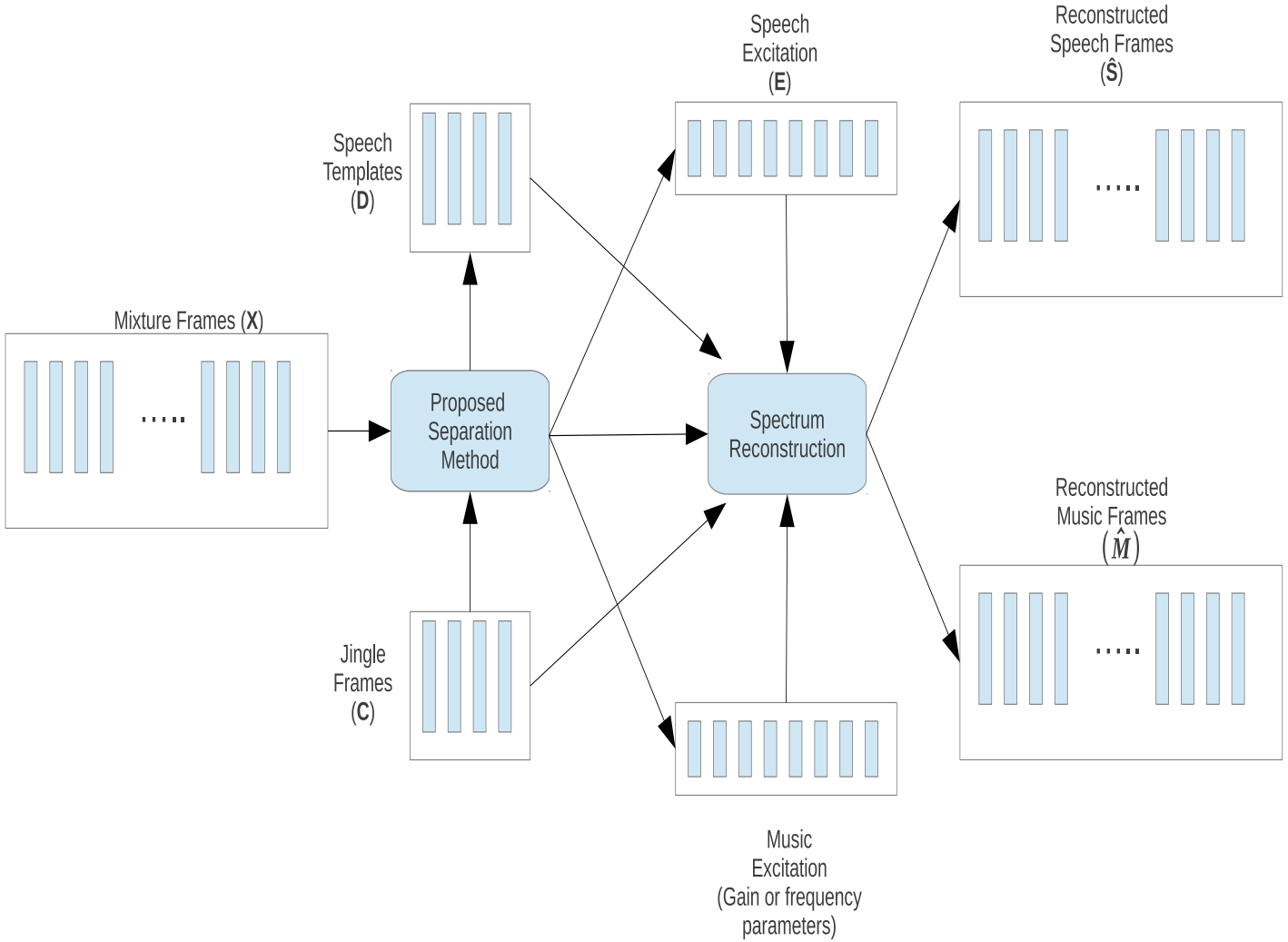

Figure 4.5. Proposed speech-music separation system.

4.1. Mixture of NMF Model

4.1.1. Baseline Model Description

In this model, we can express each time-frequency entry of the spectrum of the mixed signal at time t and frequency bin f as

$$X_{ft} = S_{ft} + M_{ft}$$

where $\mathbf{S}$ and $\mathbf{M}$ represents the magnitude or complex spectrum of the speech and music signals, respectively. We assume an NMF based generative model, which uses a Poisson [5] or a complex Gaussian [6] observation model, for the spectrum of the speech signal.

In Poisson model, the magnitude spectrogram of the speech signal is assumed to be generated by latent Poisson sources. Maximization of the likelihood of the magnitude spectrum of the signal with Poisson sources corresponds to the minimization of the KL divergence between the magnitude spectrogram of the signal with its NMF approximation [5].

However, in complex Gaussian model, the latent sources are complex Gaussians and they generate the complex spectrum of the speech signal. Moreover, maximization of the likelihood of the complex spectrum of the signal with complex Gaussian sources corresponds to the minimization of the IS divergence between the power spectrogram of the signal with its NMF approximation [6]. In other words, in IS case, we are using the complex spectrum of the signal ($\mathbf{X}$) in probabilistic model. However, maximizing the likelihood of the data with complex spectrum yields the minimization of the IS divergence between the power spectrum of the signal ($|\mathbf{X}|^{\mathbf{2}}$) and its approximation.

In this probabilistic model, each time-frequency entry of the spectrum of the

speech signal is assumed to be generated by B Poisson or complex Gaussian sources as

$$S_{ft} = \sum_{b=1}^{B} s_{fbt}.$$

Each latent Poisson and complex Gaussian source model is given by

$$\begin{aligned} s_{fbt} &\sim \mathcal{PO}(s_{fbt}; D_{fb}E_{bt}) \\ s_{fbt} &\sim \mathcal{N}_c(s_{fbt}; 0, D_{fb}E_{bt}) \end{aligned}$$

where Poisson density of the random variable s is given as

$$\mathcal{PO}(s; \lambda) = \exp(s \log \lambda - \lambda - \log \Gamma(s+1))$$

and complex Gaussian density of the random variable s is given as

$$\mathcal{N}_c(s; \mu, \Sigma) = |\pi\Sigma|^{-1} \exp(-(s-\mu)^H \Sigma^{-1}(s-\mu)).$$

In this representation, **D** and **E** matrices contain the parameters of the spectrum of the speech signal. In NMF model, **D** contains template vectors of the magnitude or power spectrograms of the speech signal and **E** contains the corresponding excitations of the template vectors. The template matrix represents the prior information about the source signal and the excitation matrix represents the variance of the prior information in the mixed signal to be separated. However, in this study, no prior information about the speech signal is used for the separation and both template and excitation matrices are estimated from the mixed signal. We use Poisson or complex Gaussian mixture

models in the generative model of the spectrogram of the music signal respectively

$$M_{ft} = m_{ft} \tag{4.1}$$

$$m_{ft}|r_t \sim \prod_{j=1}^{N} \mathcal{PO}(m_{ft}; C_{fj}h_f v_t)^{[r_t=j]} \tag{4.2}$$

$$m_{ft}|r_t \sim \prod_{j=1}^{N} \mathcal{N}_c(m_{ft}; 0, C_{fj}h_f v_t)^{[r_t=j]} \tag{4.3}$$

where r_t represents the active jingle frame and $[r_t = j]$ represents the indicator function, which is 1 when j-th frame of the jingle is used as the background music frame and its value is 0, otherwise. In Equations (4.1) and (4.3), C_{fj} represents the spectrogram corresponding to the f-th frequency bin and the j-th frame of the jingle. h_f represents filtering parameter for frequency bin f and v_t represents the gain parameter for time frame t. The goal here is to model the gain changes (fade-in, fade-out) and filtering (equalization). Each jingle frame is drawn independently from a set of jingle frames as

$$r_t = j \in \{1, 2, .., N\} \text{ with probability } \pi_j$$

where π represents probability distribution on the jingle frames and N represents the number of jingle frames. In this model, it should be emphasized that, which jingle from the catalog is used, is assumed to be known as a prior. The difference from the speech model is that, the intensity parameter of the Poisson model or variance parameter of the complex Gaussian model is chosen from the magnitude or power spectrograms of a set of previously obtained jingle frames. Each time-frequency entry of the obtained jingle frame is changed via a scaling parameter to model the fade-in or fade-out in time domain or equalization in frequency domain. The scaling parameter consists of the gain parameter for each time frame and the filtering parameter for each frequency bin.

The overall graphical model corresponding to the generation of the mixture of the speech and music signals is shown in Figure 4.4. The upper side of the graphical model generates the spectrogram of the speech part of the mixture whereas the lower side

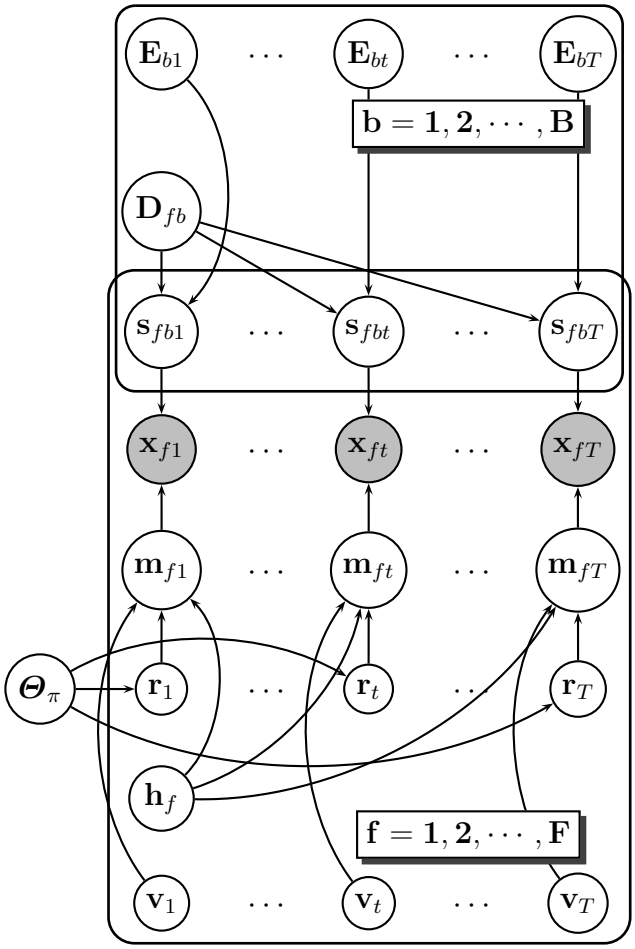

Figure 4.6. Graphical model for speech-music mixture.

generates the spectrogram of the music part. When we examine the overall probabilistic model, the distributions of the observed data conditioned on the model parameters are mixture of Poissons or complex Gaussians for KL and IS models, respectively.

4.1.2. EM Algorithm for Poisson Case

After describing the probabilistic model, the appropriate inference methodology must be developed for estimating the parameters of the latent speech and music sources to be reconstructed. Since the probabilistic model contains the latent sources and parameters, EM approach can be used as an inference method.

First, in E-step, the expectation of the joint log-likelihood of the latent sources and data under the posterior distribution of the latent sources must be calculated. If

r_t is given, that means we know which frame of the jingle generated the t-th frame of the background music. This case corresponds to the classical NMF model of the speech spectrum except that one music source is added to the speech sources to generate the spectrum of the music. Therefore, the calculation of the posterior distribution of the speech and music sources given the active jingle frame is identical to the calculation of the posteriors of the speech sources in the classical NMF models [5]. The derivations of posteriors of the speech sources for Poisson model is described in [5]. We extend the classical NMF derivations for the mixture of NMF case in this study.

E step: First let us write the joint distribution of the mixture of NMF model

$$
\begin{aligned}
q(s,m,r|\mathbf{X},\Theta) &= q(s,m|\mathbf{X},r,\Theta)q(r|\mathbf{X},\Theta) && (4.4)\\
q(s,m|\mathbf{X},r,\Theta) &= \frac{p(s,m,\mathbf{X}|r)}{p(\mathbf{X}|r)} && (4.5)\\
q(s,m|\mathbf{X},r,\Theta) &= \exp(\log(p(s,m,\mathbf{X}|r)) - \log(p(\mathbf{X}|r))) && (4.6)
\end{aligned}
$$

$$
\phi = p(s,m,\mathbf{X}|r,\Theta) = p(\mathbf{X}|s,m,r)p(s|\mathbf{D},\mathbf{E})p(m|r,h,v) \qquad (4.7)
$$

where Θ represents the model parameters $(\mathbf{D},\mathbf{E},h,v,\pi)$.

The conditional joint log-likelihood of the data given the active jingle index with the latent sources can be written as:

$$
\begin{aligned}
\log\phi &= \sum_{f,t}\Bigg[\sum_{b}(-D_{fb}E_{bt} + s_{fbt}\log(D_{fb}E_{bt}) - \log\Gamma(s_{fbt}+1)) && (4.8)\\
&- C_{fj}h_f v_t + m_{ft}\log(C_{fj}h_f v_t) - \log\Gamma(m_{ft}+1) + \log\delta(X_{ft} - \sum_{b} s_{fbt} - m_{ft})\Bigg]
\end{aligned}
$$

The conditional marginal log-likelihood of the data given the active jingle index is:

$$
\log p(\mathbf{X}|r,\Theta) = \sum_{f,t}\left[X_{ft}\log(\sum_{b} D_{fb}E_{bt} + C_{fj}h_f v_t) - \log\Gamma(X_{ft}+1)\right] \qquad (4.9)
$$

For the latent speech and music sources in Poisson model, it is known that if the observation is the sum of the values of Poisson sources, the posterior distribution over

the sources given the observation is a multinomial distribution [5]. Since we have a different multinomial distribution for each active jingle frame, j, the overall posterior distribution over the latent sources is a mixture of multinomials. For each j, the posterior distribution of the latent sources is a conditional multinomial distribution as follows:

$$\begin{aligned}
q(s,m|\mathbf{X},r,\Theta) &= \exp(\log(p(s,m,\mathbf{X}|r)) - \log(p(\mathbf{X}|r))) \\
&= \exp\Bigg\{ \sum_{f,t} \Big[\sum_{b} -D_{fb}E_{bt} + s_{fbt}\log(D_{fb}E_{bt}) - \log\Gamma(s_{fbt}+1)\Big] \\
&\quad -C_{fj}h_f v_t + m_{ft}\log(C_{fj}h_f v_t) - \log\Gamma(m_{ft}+1) \\
&\quad + \log\delta(X_{ft} - \sum_{b} s_{fbt} - m_{ft}) \\
&\quad -\Big[(\sum_{b} s_{fbt} + m_{ft})\log(\sum_{b} D_{fb}E_{bt} + C_{fj}h_f v_t) - \log\Gamma(X_{ft}+1)\Big]\Bigg\} \\
&= \exp\Bigg\{ \sum_{f,t} \Big[\sum_{b} s_{fbt}\log(\frac{D_{fb}E_{bt}}{\sum_b D_{fb}E_{bt} + C_{fj}h_f v_t})\Big] \\
&\quad +m_{ft}\log(\frac{C_{fj}h_f v_t}{\sum_b D_{fb}E_{bt} + C_{fj}h_f v_t})\Bigg\} \\
&= \mathcal{M}(s^j_{f1t},..,s^j_{fBt},m^j_{ft};X_{ft},p^j_{f1t},...,p^j_{fBt},p^j_{ft})
\end{aligned}$$

where $\mathcal{M}$ represents the multinomial distribution. The parameters (p^j_{fbt}, p^j_{ft}) of the multinomial distribution represents the conditional posterior probability of b-th speech source (s^j_{fbt}) and the music source (m^j_{ft}) in frequency bin f and time frame t conditioned on j-th jingle frame can be found as follows:

$$p^j_{fbt} = \frac{D_{fb}E_{bt}}{\sum_b D_{fb}E_{bt} + C_{fj}h_f v_t} \tag{4.10}$$

$$p^j_{ft} = \frac{C_{fj}h_f v_t}{\sum_b D_{fb}E_{bt} + C_{fj}h_f v_t}. \tag{4.11}$$

The last latent sources are active jingle frame indexes and now the posterior distribution of the active jingle frame indexes are computed. The posterior probability of the active

jingle frame index, j, at time t in Poisson model can be found as follows:

$$p(r_t = j|\mathbf{X}, \Theta) = \frac{p(\mathbf{X}|r_t = j, \Theta)p(r_t = j, \Theta)}{\sum_j p(\mathbf{X}|r_t = j, \Theta)p(r_t = j, \Theta)} \quad (4.12)$$

$$= \frac{\prod_f \mathcal{PO}(X_{ft}; C_{fj}h_f v_t + \sum_b D_{fb}E_{bt})\pi_j}{\sum_j \left[\prod_f \mathcal{PO}(X_{ft}; C_{fj}h_f v_t + \sum_b D_{fb}E_{bt})\pi_j\right]}. \quad (4.13)$$

It should be noted that, the posterior distribution over the jingle index frames at time t are dependent only on the current observation (X_t). In other words, the posterior distribution can be computed for each time frame independently.

Now, the sufficient statistics of the latent sources is calculated to be used in the M-step. In Poisson model, the sufficient statistics which correspond to the conditional marginal expectations of the latent sources can be calculated as follows:

$$\langle s^j_{fbt} \rangle = X_{ft} p^j_{fbt} \quad (4.14)$$

$$\langle m^j_{ft} \rangle = X_{ft} p^j_{bt}. \quad (4.15)$$

The expected value of active jingle frame r_t being equal to j at time frame t is

$$\langle [r_t = j] \rangle = p(r_t|X) \quad (4.16)$$

M Step: After calculating the expectations, we can find the model parameters that maximize the likelihood of the data. In M-Step, the expected value of the joint log-likelihood of the data and the latent sources under the posterior distribution of the latent sources, which is represented as Q, is calculated and used for finding the maximizing model parameters. The details of the calculations can be found in Appendix B.1. We compute the parameters of the speech spectrum, **D** and **E** matrices. Each entry of the template matrix, **D**, can be calculated as

$$D_{fb} = \frac{\sum_{t,j} \langle [r_t = j] \rangle \langle s^j_{fbt} \rangle}{\sum_t E_{bt}} \quad (4.17)$$

Now, we find the each entry of the excitation matrix of the speech spectrogram, **E**,

using the following equation

$$E_{bt} = \frac{\sum_{f,j}\langle[r_t = j]\rangle\langle s^j_{fbt}\rangle}{\sum_f D_{fb}} \tag{4.18}$$

$$\tag{4.19}$$

We want to find the filtering parameter for each frequency bin, h_f, and gain parameter for each time frame, v_t. The filtering parameter for each frequency bin can be found using

$$h_f = \frac{\sum_{t,j}\langle[r_t = j]\rangle\langle m^j_{ft}\rangle}{\sum_{t,j}\langle[r_t = j]\rangle C_{fj} v_t} \tag{4.20}$$

The gain parameter for time t can be found using

$$v_t = \frac{\sum_{f,j}\langle[r_t = j]\rangle\langle m^j_{ft}\rangle}{\sum_{f,j}\langle[r_t = j]\rangle C_{fj} h_f} \tag{4.21}$$

The resulting mixture of NMF model-based update rules and traditional NMF updates for speech-music separation for Poisson observation model are presented in the following tables. For traditional NMF method, the magnitude spectrogram of the jingle matrix $\mathbf{C}$ is used as the template matrix ($\mathbf{D}^m$) for the music signal and corresponding excitation matrix is represented with $\mathbf{E}^m$. For mixture of NMF based approach, instead of excitation matrix, $\mathbf{E}^m$, the gain and frequency filtering parameters are used for representing modeling the music signal in the mixed signal. $\mathbf{D}^s$ and $\mathbf{E}^s$ represent the template and the corresponding excitation matrices for the speech signal, respectively. The overview of the EM algorithm can be seen in Figure 4.7.

The reconstruction of the source signals with mixture of NMF models are obtained in a similar fashion to Section 3.3.

$$\widehat{\mathbf{S}} = \mathbf{X} \otimes \frac{\mathbf{D}^s\mathbf{E}^s}{(\mathbf{D}^s\mathbf{E}^s + (\mathbf{CR}) \otimes (hv^T))}. \tag{4.22}$$

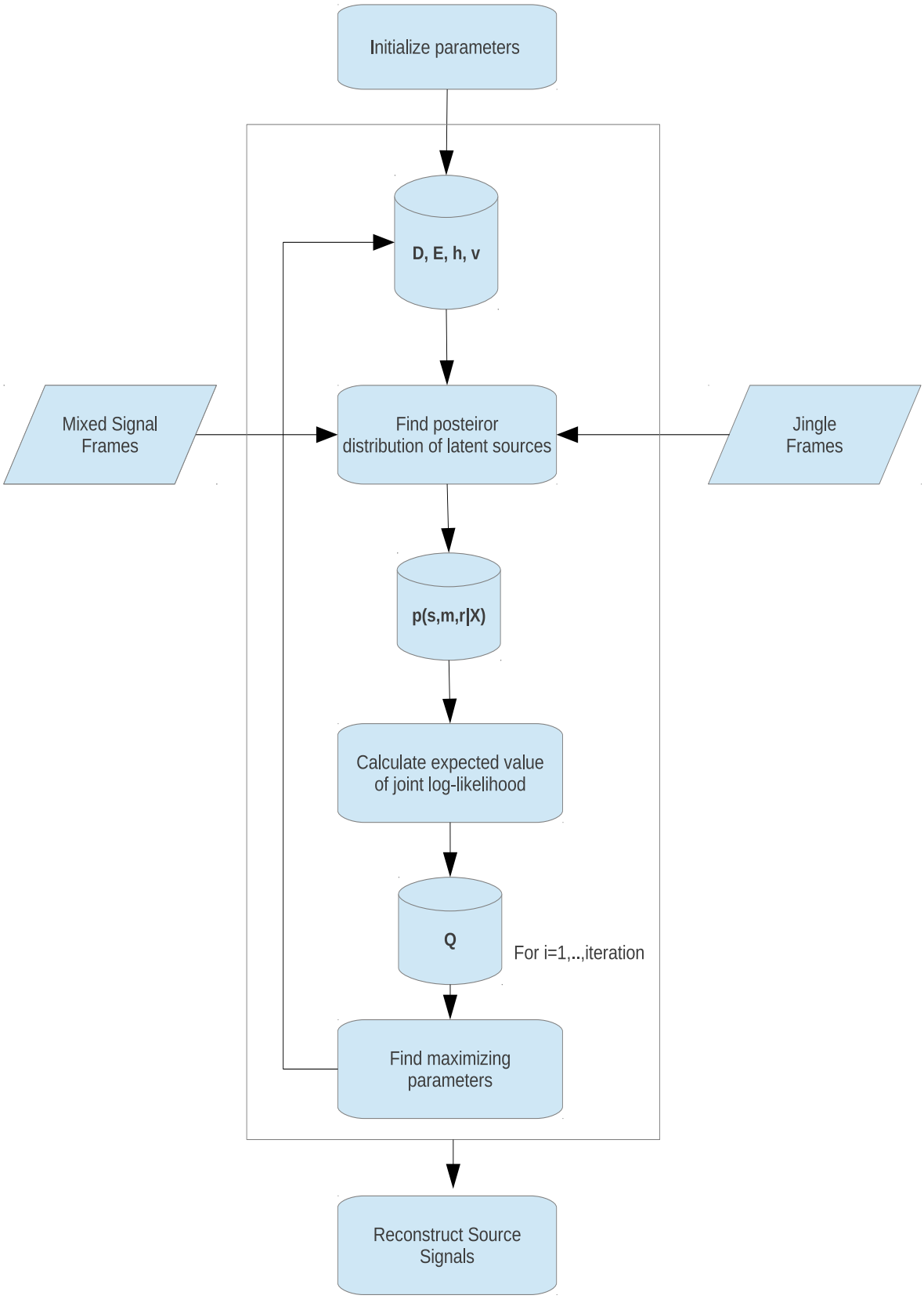

Figure 4.7. EM algorithm summary for speech-music separation.

$$\widehat{\mathbf{M}} = \mathbf{X} \otimes \frac{(\mathbf{CR}) \otimes (hv^T)}{(\mathbf{D}^s\mathbf{E}^s + (\mathbf{CR}) \otimes (hv^T))}. \tag{4.23}$$

where $\mathbf{R}$ represents the posterior probabilities of the jingle indexes for all time frames.

Summary of KL-NMF Based Speech-Music Separation with Known Jingle:

(i) Compute the posterior cell probability of the speech source:

$$p_{fbt} = \frac{D^s_{fb}E^s_{bt}}{\sum_b D^s_{fb}E^s_{bt} + \sum_j C_{fj}E^m_{jt}} \tag{4.24}$$

(ii) Compute the sufficient statistics for the speech source:

$$\langle s_{fbt} \rangle = \frac{X_{ft}(D^s_{fb}E^s_{bt})}{\sum_b D^s_{fb}E^s_{bt} + \sum_j C_{fj}E^m_{jt}} \tag{4.25}$$

(iii) Compute the template matrix for the speech signal:

$$D^s_{fb} = \frac{\sum_t \langle s_{fbt} \rangle}{\sum_t E^s_{bt}} \tag{4.26}$$

(iv) Compute the excitation matrix for the speech signal:

$$E^s_{bt} = \frac{\sum_f \langle s_{fbt} \rangle}{\sum_f D^s_{fb}} \tag{4.27}$$

(v) Compute the posterior cell probability of the music source:

$$p_{fkt} = \frac{C_{fk}E^m_{kt}}{\sum_b D^s_{fb}E^s_{bt} + \sum_j C_{fj}E^m_{jt}} \tag{4.28}$$

(vi) Compute the sufficient statistics for the music source:

$$\langle m_{fkt} \rangle = \frac{X_{ft}(C_{fk}E^m_{kt})}{\sum_b D^s_{fb}E^s_{bt} + \sum_j C_{fj}E^m_{jt}} \tag{4.29}$$

(vii) Compute the excitation matrix for the music signal:

$$E_{kt}^{m} = \frac{\sum_f \langle m_{fkt} \rangle}{\sum_f C_{fk}} \tag{4.30}$$

(viii) Iterate over Equation 4.24 to Equation 4.30

(ix) Reconstruct the source signals using Equations 3.51 and 3.52.

KL Mixture of NMF Based Speech-Music Separation with Known Jingle:

(i) Compute the posterior probability of active jingle frame index:

$$p(r_t|X) = \frac{\prod_{ft} \mathcal{PO}(X_{ft}; C_{fj} h_f v_t + \sum_b D_{fb} E_{bt}) \pi_j}{\sum_j \left[\prod_{ft} \mathcal{PO}(X_{ft}; C_{fj} h_f v_t + \sum_b D_{fb} E_{bt}) \pi_j \right]} = \langle [r_t = j] \rangle. \tag{4.31}$$

(ii) Compute the posterior cell probability of the speech source:

$$p_{fbt}^{j} = \frac{D_{fb} E_{bt}}{\sum_b D_{fb} E_{bt} + C_{fj} h_f v_t} \tag{4.32}$$

(iii) Compute the sufficient statistics for the speech source:

$$\langle s_{fbt}^{j} \rangle = \frac{X_{ft} (D_{fb}^{s} E_{bt}^{s})}{\sum_b D_{fb}^{s} E_{bt}^{s} + C_{fj} h_f v_t} \tag{4.33}$$

(iv) Compute the template matrix for the speech signal:

$$D_{fb}^{s} = \frac{\sum_{t,j} \langle [r_t = j] \rangle \left\langle s_{fbt}^{j} \right\rangle}{\sum_t E_{bt}^{s}} \tag{4.34}$$

(v) Compute the excitation matrix for the speech signal:

$$E_{bt}^{s} = \frac{\sum_{f,j} \langle [r_t = j] \rangle \left\langle s_{fbt}^{j} \right\rangle}{\sum_f D_{fb}^{s}} \tag{4.35}$$

(vi) Compute the posterior cell probability of the music source:

$$p^{j}_{ft} = \frac{C_{fj}h_f v_t}{\sum_b D_{fb}E_{bt} + C_{fj}h_f v_t} \tag{4.36}$$

(vii) Compute the sufficient statistics for the music source:

$$\langle m^{j}_{ft} \rangle = \frac{X_{ft}(C_{fj}h_f v_t)}{\sum_b D^{s}_{fb}E^{s}_{bt} + C_{fj}h_f v_t} \tag{4.37}$$

(viii) Compute the gain parameter for the music signal:

$$v_t = \frac{\sum_{f,j}\langle [r_t = j]\rangle \langle m^{j}_{ft}\rangle}{\sum_{f,j}\langle [r_t = j]\rangle C_{fj}h_f} \tag{4.38}$$

(ix) Compute the filtering parameter for the music signal:

$$h_f = \frac{\sum_t \langle m_{ft}\rangle}{\sum_{t,j}\langle [r_t = j]\rangle C_{fj}v_t} \tag{4.39}$$

(x) Iterate over Equation 4.31 to Equation 4.39.

(xi) Reconstruct the source signals using Equations 4.22 and 4.23.

4.1.3. EM Algorithm for Complex Gaussian Case

The same strategy in Section 4.1.2 is used for obtaining the EM update equations. In complex Gaussian model case, the calculation of the posterior distribution of the speech and music sources given the active jingle frame is identical to the calculation of the posteriors of the speech sources in the classical IS-NMF model [6]. The derivations of posteriors of the speech sources for complex Gaussian model are described in [6].
E step: First let us write the joint distribution of the mixture of NMF model with complex Gaussian observation model. The conditional joint log-likelihood of the data given the active jingle index with the latent sources can be written as:

$$\log \phi \;=\; \sum_{f,t} \Big[\sum_b -\log(D_{fb}E_{bt}) - \frac{|s_{fbt}|^2}{D_{fb}E_{bt}} \Big] - \log(C_{fj}h_f v_t) - \frac{|m_{ft}|^2}{C_{fj}h_f v_t}$$

The conditional marginal log-likelihood of the data given the active jingle index is:

$$\log p(\mathbf{X}|r,\Theta) = \sum_{f,t} \big[-\log(\sum_{b} D_{fb}E_{bt} + C_{fj}h_f v_t) - \frac{|X_{ft}|^2}{\sum_b D_{fb}E_{bt} + C_{fj}h_f v_t}\big] \quad (4.40)$$

Similarly, for the latent speech and music sources in complex Gaussian model, it is known that if the observation is the sum of the values of complex Gaussian sources, the posterior distribution over the sources given that observation is a complex Gaussian distribution [6]. Since we have a different complex Gaussian distribution for each active jingle frame, j, the overall posterior distribution over the latent sources is a mixture of complex Gaussians. For each jingle frame j, the conditional posterior distribution of the latent speech and music sources can be written as follows:

$$\begin{aligned} p(s_{fbt}|X,r_t) &= \mathcal{N}_c(s^j_{fbt};\mu^j_{fbt},\Sigma^j_{fbt}) \\ p(m_{ft}|X,r_t) &= \mathcal{N}_c(m^j_{ft};\mu^j_{ft},\Sigma^j_{ft}). \end{aligned}$$

The conditional posterior mean and variance of b-th speech source in frequency bin f and time frame t conditioned on j-th jingle frame are (Details are in Appendix B.2):

$$\Sigma^j_{fbt} = \frac{D_{fb}E_{bt}}{\sum_b D_{fb}E_{bt} + C_{fj}h_f v_t}(\sum_{i\neq b} D_{fi}E_{it} + C_{fj}h_f v_t) \quad (4.41)$$

$$\mu^j_{fbt} = \frac{D_{fb}E_{bt}}{\sum_b D_{fb}E_{bt} + C_{fj}h_f v_t} X_{ft} \quad (4.42)$$

The last latent sources are active jingle frame indexes and now the posterior distribution of the active jingle frame indexes are computed. The posterior probability of the active jingle frame, j, at time t in Gaussian model is:

$$p(r_t|X) = \frac{\prod_{ft}\mathcal{N}_c(X_{ft};0,C_{fj}h_f v_t + \sum_b D_{fb}E_{bt})\pi_j}{\sum_j \big[\prod_{ft}\mathcal{N}_c(X_{ft};0,C_{fj}h_f v_t + \sum_b D_{fb}E_{bt})\pi_j\big]}. \quad (4.43)$$

Now, the sufficient statistics of the latent sources is calculated to be used in the M-step. In complex Gaussian model, the sufficient statistics which corresponds to the expected

power of the latent complex sources can be calculated as follows:

$$\langle |s^j_{fbt}|^2 \rangle = \Sigma^j_{fbt} + |\mu^j_{fbt}|^2 \tag{4.44}$$

$$\langle |m^j_{ft}|^2 \rangle = \Sigma^j_{ft} + |\mu^j_{ft}|^2. \tag{4.45}$$

The expected value of active jingle frame r_t being equal to j at time frame t is

$$\langle [r_t = j] \rangle = p(r_t|X) \tag{4.46}$$

M Step: After calculating the expectations, we can find the model parameters that maximize the likelihood of the data. In M-Step, the expected value of the joint log-likelihood of the data and the latent sources under the posterior distribution of the latent sources, which is represented as Q, is calculated and used for finding the maximizing model parameters. The details of the calculations can be found in Appendix B.2. We compute the parameters of the power spectrum of the speech signal, **D** and **E** matrices. Each entry of the template matrix, **D**, can be calculated as

$$D_{fb} = \frac{1}{T} \sum_{t,j} \langle [r_t = j] \rangle \frac{\langle |s^j_{fbt}|^2 \rangle}{E_{bt}} \tag{4.47}$$

Now, we find each entry of the excitation matrix of the speech spectrogram, **E**, using the following equation

$$E_{bt} = \frac{1}{F} \sum_{f,j} \langle [r_t = j] \rangle \frac{\langle |s^j_{fbt}|^2 \rangle}{D_{fb}} \tag{4.48}$$

We want to find the filtering parameter for each frequency bin, h_f, and gain parameter for each time frame, v_t. The filtering parameter for each frequency bin can be found using

$$h_f = \frac{1}{T} \sum_{t,j} \langle [r_t = j] \rangle \frac{\langle |m^j_{ft}|^2 \rangle}{C_{fj} v_t}. \tag{4.49}$$

The gain parameter for time t can be found using

$$v_t = \frac{1}{F}\sum_{b,j}\langle[r_t = j]\rangle\frac{\langle|m_{ft}^j|^2\rangle}{C_{fj}h_f} \tag{4.50}$$

The resulting mixture of NMF model-based update rules and traditional NMF updates for speech-music separation for complex Gaussian observation model are presented in the following summaries. For traditional NMF method, the power spectrogram of the jingle matrix $\mathbf{C}$ is used as the template matrix ($\mathbf{D}^m$) for the music signal and corresponding excitation matrix is represented with $\mathbf{E}^m$. For mixture of NMF based approach, instead of excitation matrix, $\mathbf{E}^m$, the gain and frequency filtering parameters are used for modeling the music signal in the mixed signal. $\mathbf{D}^s$ and $\mathbf{E}^s$ represent the template and the corresponding excitation matrices for the speech signal, respectively. In the following equations, $|\mathbf{X}|^{\mathbf{2}}$ represents the power spectrogram of the mixed signal. The reconstruction of the source signals with mixture of NMF models are obtained in a similar fashion to Section 3.3.

$$\widehat{\mathbf{S}} = |\mathbf{X}|^{\mathbf{2}}\otimes\frac{\mathbf{D}^s\mathbf{E}^s}{(\mathbf{D}^s\mathbf{E}^s + (\mathbf{CR})\otimes(hv^T))}. \tag{4.51}$$

$$\widehat{\mathbf{M}} = |\mathbf{X}|^{\mathbf{2}}\otimes\frac{(\mathbf{CR})\otimes(hv^T)}{(\mathbf{D}^s\mathbf{E}^s + (\mathbf{CR})\otimes(hv^T))}. \tag{4.52}$$

where $\mathbf{R}$ represents the posterior probabilities of the jingle indexes for all time frames.

Summary of IS-NMF Based Speech-Music Separation with Known Jingle:

(i) Compute the posterior distribution parameters of the speech source:

$$\Sigma_{fbt} = \frac{D_{fb}^s E_{bt}^s}{\sum_b D_{fb}^s E_{bt}^s + \sum_j C_{fj}E_{jt}^m}\left(\sum_{i\neq b} D_{fi}^s E_{it}^s + \sum_j C_{fj}E_{jt}^m\right) \tag{4.53}$$

$$\mu_{fbt} = \frac{D_{fb}^s E_{bt}^s}{\sum_b D_{fb}^s E_{bt}^s + \sum_j C_{fj}E_{jt}^m}X_{ft} \tag{4.54}$$

(ii) Compute the sufficient statistics for the speech source:

$$\langle |s_{fbt}|^2 \rangle = \Sigma_{fbt} + |\mu_{fbt}|^2 \tag{4.55}$$

(iii) Compute the template matrix for the speech signal:

$$D^s_{fb} = \frac{1}{T} \sum_t \frac{\langle |s_{fbt}|^2 \rangle}{E^s_{bt}} \tag{4.56}$$

(iv) Compute the excitation matrix for the speech signal:

$$E^s_{bt} = \frac{1}{F} \sum_f \frac{\langle |s_{fbt}|^2 \rangle}{D^s_{fb}} \tag{4.57}$$

(v) Compute the posterior distribution parameters of the music source:

$$\Sigma_{ukt} = \frac{C_{uk} E^m_{kt}}{\sum_b D^s_{fb} E^s_{bt} + \sum_j C_{fj} E^m_{jt}} \left(\sum_b D^s_{fb} E^s_{bt} + \sum_{i \neq k} C_{fi} E^m_{bt} \right) \tag{4.58}$$

$$\mu_{ukt} = \frac{C_{uk} E^m_{kt}}{\sum_b D^s_{fb} E^s_{bt} + \sum_j C_{fj} E^m_{jt}} X_{ft} \tag{4.59}$$

(vi) Compute the sufficient statistics for the music source:

$$\langle |m_{ukt}|^2 \rangle = \Sigma_{ukt} + |\mu_{ukt}|^2 \tag{4.60}$$

(vii) Compute the excitation matrix for the music signal:

$$E^m_{kt} = \frac{1}{F} \sum_f \frac{\langle |m_{fkt}|^2 \rangle}{C_{fk}} \tag{4.61}$$

(viii) Iterate over Equation 4.53 to Equation 4.61

(ix) Reconstruct the source signals using Equations 3.51 and 3.52.

IS Mixture of NMF Based Speech-Music Separation with Known Jingle:

(i) Compute the posterior probability of active jingle frame index:

$$p(r_t|X) = \frac{\prod_{ft} \mathcal{N}_c(X_{ft}; 0, C_{fj}h_f v_t + \sum_b D_{fb}E_{bt})\pi_j}{\sum_j \left[\prod_{ft} \mathcal{N}_c(X_{ft}; 0, C_{fj}h_f v_t + \sum_b D_{fb}E_{bt})\pi_j\right]} \tag{4.62}$$

(ii) Compute the posterior distribution parameters of the speech source:

$$\Sigma^j_{fbt} = \frac{D^s_{fb}E^s_{bt}}{\sum_b D^s_{fb}E^s_{bt} + C_{fj}h_f v_t}(\sum_{i\neq b} D^s_{fb}E^s_{bt} + C_{fj}h_f v_t) \tag{4.63}$$

$$\mu^j_{fbt} = \frac{D^s_{fb}E^s_{bt}}{\sum_b D^s_{fb}E^s_{bt} + C_{fj}h_f v_t}X_{ft} \tag{4.64}$$

(iii) Compute the sufficient statistics for the speech source:

$$\langle |s^j_{fbt}|^2 \rangle = \Sigma^j_{fbt} + |\mu^j_{fbt}|^2 \tag{4.65}$$

(iv) Compute the template matrix for the speech signal:

$$D^s_{fb} = \frac{1}{T}\sum_{t,j} \langle [r_t = j] \rangle \frac{\langle |s^j_{fbt}|^2 \rangle}{E^s_{bt}} \tag{4.66}$$

(v) Compute the excitation matrix for the speech signal:

$$E^s_{bt} = \frac{1}{F}\sum_{b,j} \langle [r_t = j] \rangle \frac{\langle |s^j_{fbt}|^2 \rangle}{D^s_{fb}} \tag{4.67}$$

(vi) Compute the posterior distribution parameters of the music source:

$$\Sigma^j_{ft} = \frac{C_{fj}h_f v_t}{\sum_b D_{fb}E_{bt} + C_{fj}h_f v_t}(\sum_b D_{fb}E_{bt}) \tag{4.68}$$

$$\mu^j_{ft} = \frac{C_{fj}h_f v_t}{\sum_b D_{fb}E_{bt} + C_{fj}h_f v_t}X_{ft} \tag{4.69}$$

(vii) Compute the sufficient statistics for the music source:

$$p^j_{ft} = \frac{C_{fj}h_f v_t}{\sum_b D_{fb}E_{bt} + C_{fj}h_f v_t} \tag{4.70}$$

(viii) Compute the sufficient statistics for the music source:

$$\langle |m_{ft}^j|^2 \rangle = \Sigma_{bt}^j + |\mu_{bt}^j|^2 \tag{4.71}$$

(ix) Compute the gain parameter for the music signal:

$$v_t = \frac{1}{F} \sum_{b,j} \langle [r_t = j] \rangle \frac{\langle |m_{ft}^j|^2 \rangle}{C_{fj} h_f} \tag{4.72}$$

(x) Compute the filtering parameter for the music signal:

$$h_f = \frac{1}{T} \sum_{t,j} \langle [r_t = j] \rangle \frac{\langle |m_{ft}^j|^2 \rangle}{C_{fj} v_t} \tag{4.73}$$

(xi) Iterate over Equation 4.62 to Equation 4.73

(xii) Reconstruct the source signals using Equations 4.51 and 4.52.

4.2. Gain Estimation Problem in Poisson Model

In Poisson observation model, it is experimentally shown that the estimation performance of the method with update Equation 4.38 for each time frame is very poor at the mixture frames which either have

(i) low input Music-to-Speech Ratio (MSR) or

(ii) active jingle frames with low energy.

In Figure 4.8, these two reasons for low estimation performance are shown in an example. In this example, the true gain parameter is constant at 1 for all frames. For example, for the first 5 frames though the input MSR values are high, the gain estimation error is very high due to the fact that the active jingle frame has low energy and so it is confused with other frames. This fact can be seen in Figure 4.9. When we analyze the gain parameter of the frames between the time indexes 20 and 25, it can be seen that though the active jingle frames have high energy, the gain estimation error is

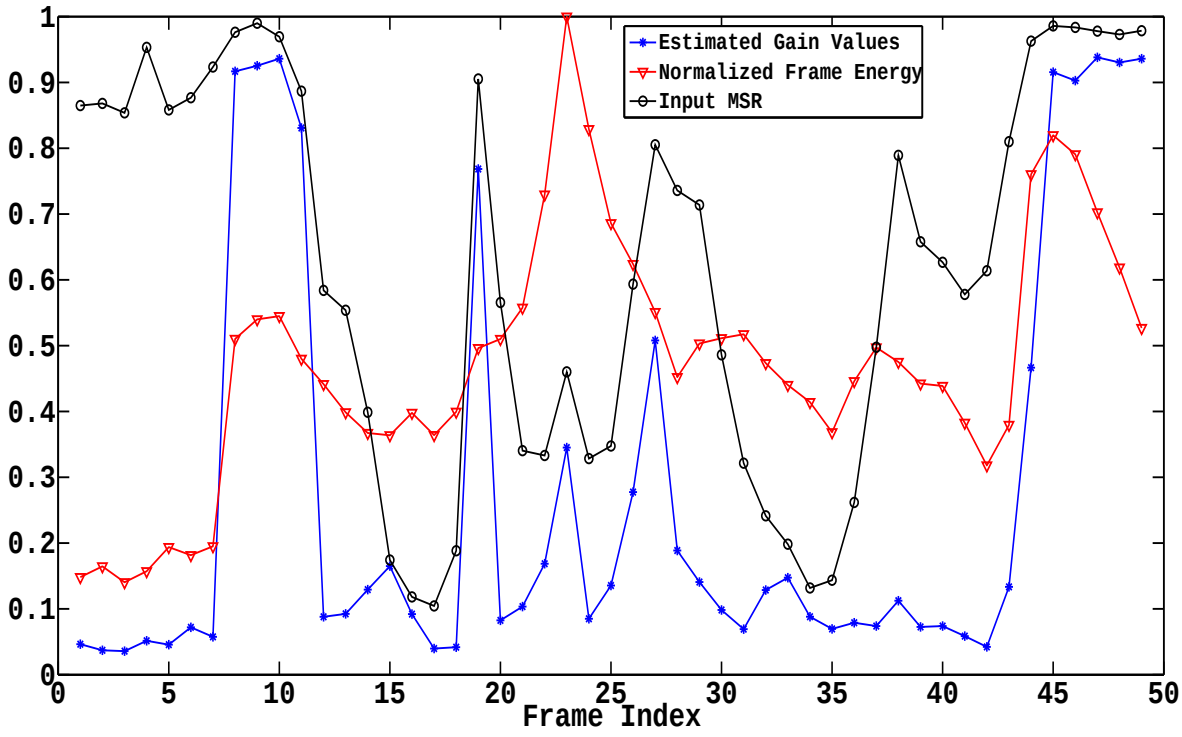

Figure 4.8. Gain estimation problem reasons: Low input MSR and low active frame energy.

high due to the fact that input MSR value is low and the speech signal suppresses the music signal. In fact, at these parts, the inference method cannot estimate the posterior probabilities of the catalog frames accurately. That is, the method cannot decide which frame of the jingle is active at these parts. This fact is shown in Figure 4.9. Although most of the maximum posterior probabilities (MPP) of the jingle frames are very low, 67% of the frames with MPP are the indeed active frames for this example.

Using this analysis about the gain estimation problem, we propose two different correction methods to enhance the gain estimation performance of the inference method. These methods are called "MAP Estimation Method" and "Piece-wise Constant Estimation Method".

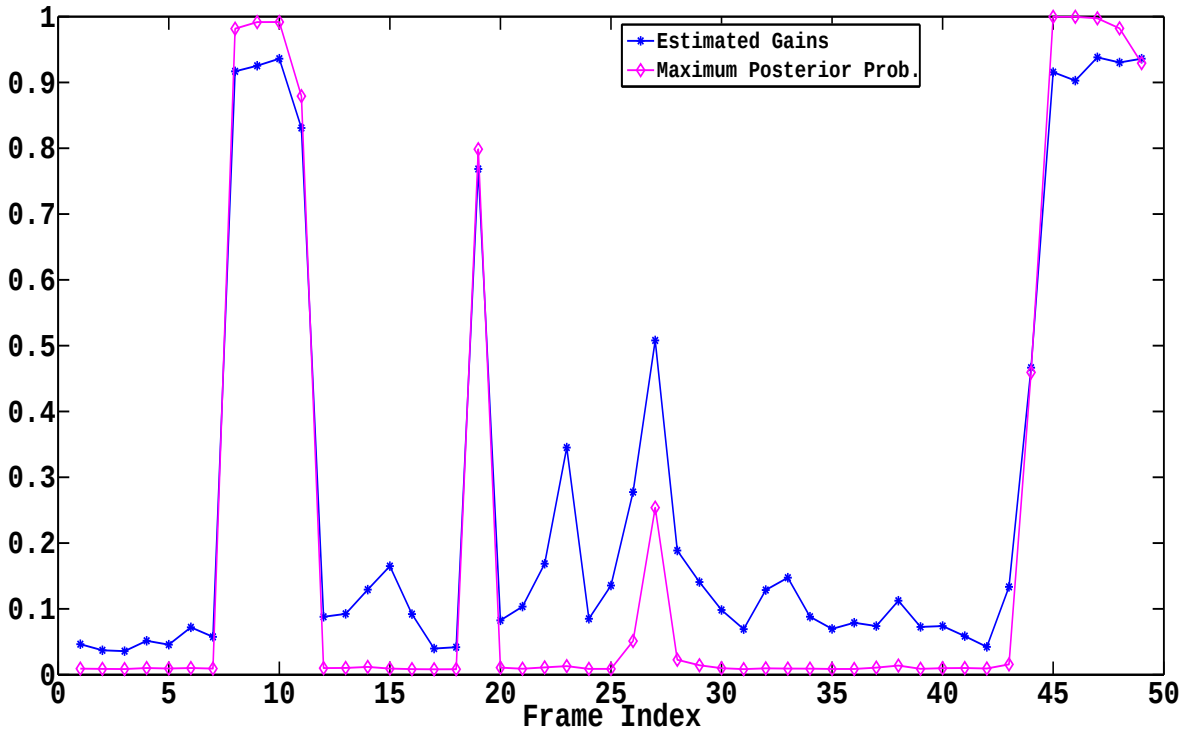

Figure 4.9. Relation between the gain parameter and the MAP values.

4.2.1. MAP Estimation Method

Experimentally, we observe that although MPP for most of the mixture frames are very low, the frames which have MPP are indeed the active frames. Therefore, after some iterations with the original posterior update Equation 4.31, the frames with the MPP can be chosen as the active frames. Then the posterior probability of these MPP frames are assigned to 1 so as to estimate the gain parameter more accurately. After this assignment, even though the posterior probabilities are not updated, other update rules are applied via reassigned posterior probabilities. This approach can be shown mathematically as follows:

$$r_t^* = \arg\max_{r_t} p(r_t|X,\theta) \tag{4.74}$$

$$p(r_t = j) = \begin{cases} 1 & \text{if } j = r_t^*, \\ 0 & \text{Otherwise} \end{cases} \tag{4.75}$$

4.2.2. Piece-wise Constant Estimation

When we analyze the gain estimation results obtained using the original update Equation 4.38 in Figure 4.9, it is observed that when the MPP of a frame is high enough, the gain parameter for this frame is estimated correctly. Therefore, we can use the gain parameter of the closest frame which has high MPP values as the gain parameter for the frame which has low MPP value. The question here is, what threshold value will be used for deciding whether the MPP of a frame is low or high? We decide on this threshold value using a development set which maximizes the separation performance. We call this estimation method, given in Algorithm 4.10, as 'Piece-wise Constant Estimation' (PCE) because the resultant gain parameter is a piece-wise constant version of the originally estimated gain parameter.

$g = empty$;
for $t = 1$ **to** T **do**
  if $MAP(t) > Threshold$ **then**
    Add t to g;
  endif
endfor
$L \Leftarrow length(g)$;
$i \Leftarrow 1$;
for $t = 1$ **to** T **do**
  if $i < L$ **and** $|t - g(i+1)| < |t - g(i)|$ **then**
    $i \Leftarrow i + 1$;
  endif
  $v_{es}(t) = v_{es}(i)$;
endfor

Figure 4.10. Piece-wise constant estimation (PCE) algorithm .

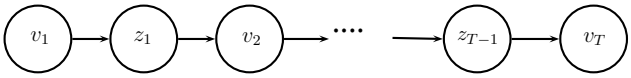

Figure 4.11. GMC graphical model for gain parameter.

4.2.3. Gamma Markov Chain for Gain Estimation

A Gamma Markov chain (GMC) [45], shown in Figure 4.11, is a prior structure for a chain of positive variables, where the correlation between consecutive variables is positive. In addition, each variable is conditionally conjugate, i.e., their prior and full conditional distributions are Gamma. A GMC of $v_{1:T}$ can be defined as

$$v_1 \sim \mathcal{G}(v_1; a_v, b_v/a_v) \tag{4.76}$$

$$z_t|v_t \sim \mathcal{IG}(z_t; a_z, a_z v_t) \tag{4.77}$$

$$v_{t+1}|z_t \sim \mathcal{G}(v_{t+1}; a_v, z_t/a_v) \tag{4.78}$$

where a_v, a_z, b_v are the hyper-parameters of the chain and $z_{1:T-1}$ are auxiliary variables introduced to have positive correlation and conjugacy properties simultaneously. a_v and a_z are the coupling hyper-parameters and they determine the degree of correlation between variables. $\mathcal{G}$ and $\mathcal{IG}$ represent Gamma and Inverse-Gamma distributions, respectively. The overall graphical model with the GMC on the gain values is shown in Figure 4.12.

The full joint distribution of the mixture-based model with GMC can be decomposed as:

$$\begin{aligned}\log \phi &= \log p(\mathbf{X}, s, m, r, v, z|\Theta) \\ &= \log p(\mathbf{X}|s, m) + \log p(s|\mathbf{D}, \mathbf{E}) + \log p(m|r, v) + \\ &\quad \log p(v, z|a_v, b_v, a_z) + \log p(r|\pi)\end{aligned}$$

where Θ represents the parameters of the latent speech and music sources and the hyper-parameters of the GMC parameters. Since the posterior distributions of the gain parameters, v, z and the hidden sources are coupled, we cannot compute the overall

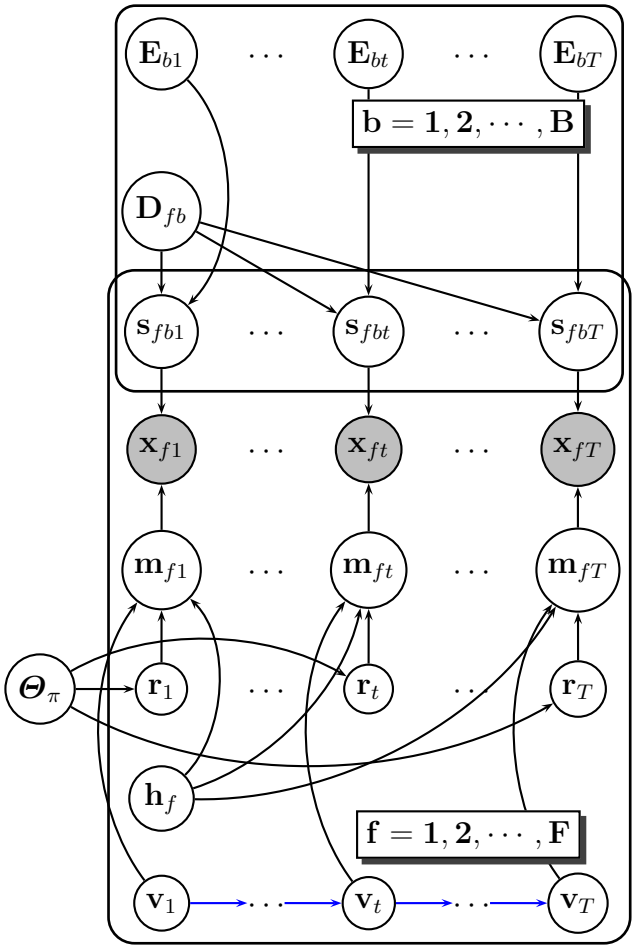

Figure 4.12. Graphical model for speech-music mixture with GMC on gain values.

joint posterior distribution exactly. In this case, we use the variational technique that factorizes the posterior distribution into the posteriors of the decoupled random variables as follows:

$$
\begin{aligned}
q(s,m,r) &\propto \exp(\langle \log p(\mathbf{X}, s, m, r, v, z|\Theta) \rangle_{q(v)q(z)}) \\
q(v) &\propto \exp(\langle \log p(\mathbf{X}, s, m, r, v, z|\Theta) \rangle_{q(s,m,r)q(z)}) \\
q(z) &\propto \exp(\langle \log p(\mathbf{X}, s, m, r, v, z|\Theta) \rangle_{q(s,m,r)q(v)})
\end{aligned}
$$

The joint posterior distribution of the latent speech and music sources and the jingle indexes is also a multinomial mixture model (MMM). However, the calculation of the parameters of the distribution differ from the original model which is described in Section 4.1.2. The overall posterior distribution can be decomposed conditioned on

the jingle frame index, r, as

$$\begin{aligned} q(s,m,r) &= q(s,m|r)q(r) \\ q(s,m|r) &= \mathcal{M}(s_{f1t},.,s_{fBt},m_{ft};X_{ft},p^j_{f1t},.,p^j_{fBt},p^j_{bt}) \end{aligned}$$

The parameters of this MMM can be computed using:

$$\begin{aligned} p^j_{fbt} &= \frac{D_{fb}E_{bt}}{(\sum_b D_{fb}E_{bt}) + C_{fj}h_f \exp(\langle \log v_t \rangle)} \\ p^j_{bt} &= \frac{C_{fj}h_f \exp(\langle \log v_t \rangle)}{(\sum_b D_{fb}E_{bt}) + C_{fj}h_f \exp(\langle \log v_t \rangle)} \\ q(r_t = j) &= \frac{\prod_{f,t} \mathcal{PO}(X_{ft}; \sum_b D_{fb}E_{bt} + C_{fj}h_f\langle v_t \rangle)\pi_j}{\sum_j \left[\prod_{f,t} \mathcal{PO}(X_{ft}; \sum_b D_{fb}E_{bt} + C_{fj}h_f\langle v_t \rangle)\pi_j\right]} \\ &= \langle [r_t = j] \rangle \end{aligned}$$

The only difference from Equations 4.32 and 4.31 is that instead of using v_t, its expectations are used for calculating the posteriors. The marginal expectation of the latent sources under the posterior distribution can be found using:

$$\langle s_{fbt} \rangle = X_{ft}(\sum_j \langle [r_t = j] \rangle p^j_{fbt}) \tag{4.79}$$

$$\langle m_{ft} \rangle = X_{ft}(\sum_j \langle [r_t = j] \rangle p^j_{bt}) \tag{4.80}$$

$$\tag{4.81}$$

Now, the posterior distribution of the gain parameter, v_t and the auxiliary variable, z_t, are calculated. The posterior of the gain parameter, v_t, is also Gamma-distributed due to the conjugacy of Poisson and Gamma distributions. The posterior distribution

of v_{t+1} conditioned on the auxiliary variable z_t is:

$$\begin{aligned} q(v_{t+1}) &\propto \mathcal{G}(v_{t+1}; \alpha^v_{t+1}, \beta^v_{t+1}) && (4.82)\\ \alpha^v_{t+1} &= a_v + \sum_u \langle m_{u(t+1)} \rangle && (4.83)\\ \beta^v_{t+1} &= (a_v \langle \frac{1}{z_t} \rangle + \sum_{b,j} C_{fj} f_u)^{-1} && (4.84) \end{aligned}$$

The sufficient statistics of the gain parameter, which are used for estimating the posteriors of the other parameters are:

$$\begin{aligned} \exp(\langle \log v_{t+1} \rangle) &= \exp(\Psi(\alpha^v_{t+1}))\beta^v_{t+1} && (4.85)\\ \langle v_{t+1} \rangle &= \alpha^v_{t+1}\beta^v_{t+1} && (4.86) \end{aligned}$$

where Ψ denotes the digamma function defined as $\Psi(\alpha) \equiv d \log \Gamma(a)/da$. We also need to compute the posterior distribution and the sufficient statistics of the inverse of the gain parameter which has an Inverse-Gamma distribution as follows:

$$\begin{aligned} \frac{1}{v_{t+1}} &\sim \mathcal{IG}(\frac{1}{v_{t+1}}; \alpha^v_{t+1}, \frac{1}{\beta^v_{t+1}}) && (4.87)\\ \langle \frac{1}{v_{t+1}} \rangle &= \frac{1}{(\alpha^v_{t+1} - 1)\beta^v_{t+1}} && (4.88) \end{aligned}$$

The posterior of auxiliary variable, z_t, is also Inverse Gamma-distributed due to the conjugacy of Poisson and Inverse Gamma distributions. The posterior distribution of z_t conditioned on the gain parameter, v_t is:

$$\begin{aligned} q(z_t) &\propto \mathcal{IG}(z_t; \alpha^z_t, \beta^z_t) && (4.89)\\ \alpha^z_t &= a_z \quad \text{and} \quad \beta^z_t = (\frac{1}{a_z} \langle \frac{1}{v_t} \rangle)^{-1} && (4.90) \end{aligned}$$

The sufficient statistics of the auxiliary variable, which are used for estimating the posterior of the gain parameter are:

$$\langle z_t \rangle = \frac{\beta^z_t}{\alpha^z_t - 1} \quad (4.91)$$

We also need to compute the posterior distribution and the corresponding sufficient statistics of the inverse of the auxiliary variable which has a Gamma distribution as follows:

$$\frac{1}{z_t} \sim \mathcal{G}(\frac{1}{z_t}; \alpha_t^z, \frac{1}{\beta_t^z}) \tag{4.92}$$

$$\langle \frac{1}{z_t} \rangle = \frac{\alpha_t^z}{\beta_t^z}. \tag{4.93}$$

4.3. Gain Estimation Problem in Complex Gaussian Model

In the previous section, we analyzed the gain estimation problem of mixture-based method with Poisson model. In [9], we proposed using a GMC which is a prior structure for a chain of positive variables to enhance the gain estimation of the proposed method. In this part, we analyze the gain estimation problem of the mixture-based method with a complex Gaussian model. In Figure 4.13, the estimated gain parameter and correctly identified active frames are plotted to show the relation between them. Since the gain value of each frame is estimated independently, the abrupt change in gain values of consecutive frames is possible and can be seen in Figure 4.13.

In order to prevent the abrupt changes in the gain values, as similar to Poisson case, Inverse-Gamma-Markov-Chain (IGMC) is used to impose the correlation between consecutive gain values for the mixture-based method with complex Gaussian model.

An IGMC [45], which is shown in Figure 4.14, is a prior structure for a chain of positive variables, where the correlation between consecutive variables is also positive. In addition, each variable is conditionally conjugate, i.e., their prior and full conditional distributions are Inverse-Gamma. An IGMC of $v_{1:T}$ can be defined as

$$z_1 \sim \mathcal{IG}(z_1; a_z, \frac{b}{a_z}) \tag{4.94}$$

$$z_t|v_{t-1} \sim \mathcal{IG}(z_t; a_z, \frac{v_{t-1}}{a_z}) \tag{4.95}$$

$$v_t|z_t \sim \mathcal{IG}(v_t; a_v, \frac{z_t}{a_v}) \tag{4.96}$$

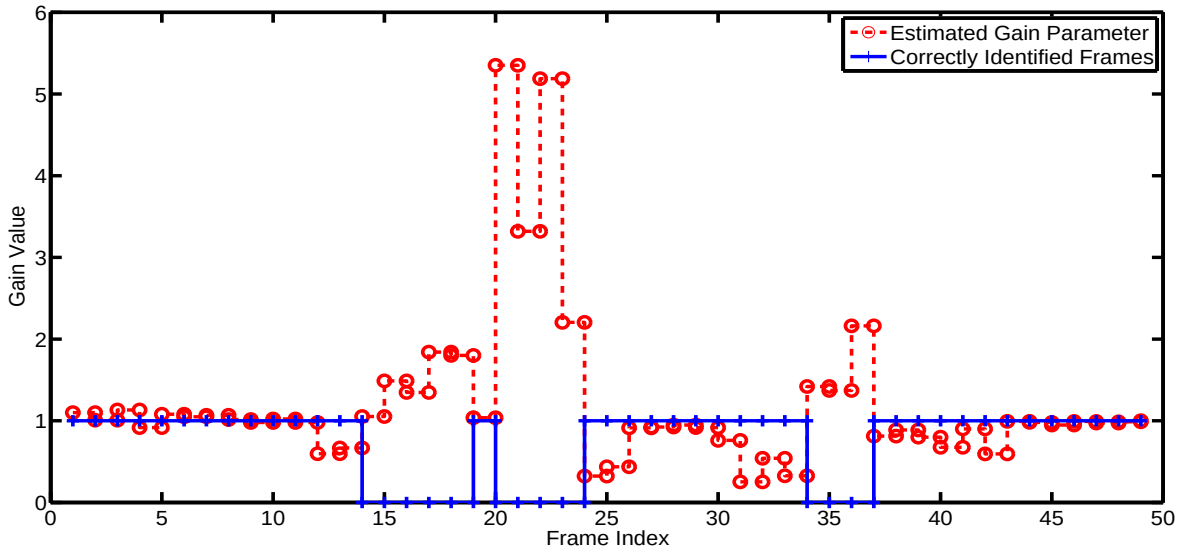

Figure 4.13. Estimated gain values and correctly identified active frames.

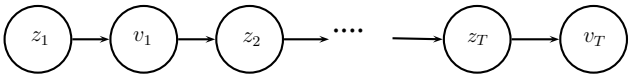

Figure 4.14. GMC graphical model for gain parameter.

where a_v, a_z, b are the hyper-parameters of the chain and $z_{1:T}$ are auxiliary variables introduced for having positive correlation and conjugacy properties simultaneously. a_v and a_z are the coupling hyper-parameters and they determine the degree of correlation between variables. The full joint distribution of the mixture-based model with IGMC can be decomposed as:

$$\begin{aligned} \log \phi &= \log p(\mathbf{X}, s, m, r, v, z|\Theta) \\ &= \log p(\mathbf{X}|s, m) + \log p(s|\mathbf{D}, \mathbf{E}) + \log p(m|r, v) + \\ &\quad \log p(v, z|a_v, b_v, a_z) + \log p(r|\pi) \end{aligned}$$

where Θ represents the parameters of the latent speech and music sources and the hyper-parameters of the GMC parameters. Since the posterior distributions of the gain parameters, v, z and the latent sources are coupled, we cannot compute the overall joint posterior distribution exactly. Therefore, we use the variational technique that factorizes the posterior distribution into the posteriors of the decoupled random

variables as follows:

$$q(s,m,r) \propto \exp(\langle\phi\rangle_{q(v)q(z)}) \tag{4.97}$$

$$q(v) \propto \exp(\langle\phi\rangle_{q(s,m,r)q(z)}) \tag{4.98}$$

$$q(z) \propto \exp(\langle\phi\rangle_{q(s,m,r)q(v)}) \tag{4.99}$$

The joint posterior distribution of the latent speech and music sources and the jingle indexes is also a CGMM. However, the calculation of the parameters of the distribution differ from the original model which is described in Section 4.1.3. The parameters of this CGMM can be computed using Equations 4.62-4.64 and Equation 4.43 by only replacing v_t with $(\langle\frac{1}{v_t}\rangle)^{-1}$. The sufficient statistics of the latent complex gaussian sources under the posterior distribution can be found using Equation 4.44.

Now, the posterior distribution of the gain parameter, v_t and the auxiliary variable, z_t, are calculated. The posterior of the gain parameter, v_t, is also inverse-Gamma-distributed due to the conjugacy of variance of Gaussian distribution and inverse-Gamma distribution. The posterior distribution of v_t conditioned on the auxiliary variable z_t is:

$$q(v_t) \propto \mathcal{IG}(v_t;\alpha_t^v,\beta_t^v) \tag{4.100}$$

$$\alpha_t^v = a_z + a_v + F \tag{4.101}$$

$$\beta_t^v = (a_z\langle\frac{1}{z_{t+1}}\rangle + a_v\langle\frac{1}{z_t}\rangle + \sum_{f,j}\langle[r_t=j]\rangle\frac{\langle|m_{ft}^j|^2\rangle}{C_{fj}f_u})^{-1} \tag{4.102}$$

The sufficient statistic of the gain parameter, which is used for estimating the posteriors of the other parameters is:

$$\langle v_t\rangle = \frac{1}{\beta_t^v(\alpha_{t+1}^v-1)} \tag{4.103}$$

We also need to compute the posterior and the sufficient statistics of the inverse of the

gain parameter which has an Gamma distribution as follows:

$$\frac{1}{v_t} \sim \mathcal{G}(\frac{1}{v_t}; \alpha_t^v, \beta_t^v) \tag{4.104}$$

$$\langle \frac{1}{v_t} \rangle = \alpha_t^v \beta_t^v \tag{4.105}$$

The posterior of auxiliary variable, z_t, is also Inverse Gamma-distributed due to the conjugacy of Complex Gaussian and Inverse Gamma distributions. The posterior distribution of z_t conditioned on the gain parameter, v_t is:

$$q(z_t) \propto \mathcal{IG}(z_t; \alpha_t^z, \beta_t^z) \tag{4.106}$$

$$\alpha_t^z = a_v + a_z \tag{4.107}$$

$$\beta_t^z = (a_z \langle \frac{1}{v_{t-1}} \rangle + a_v \langle \frac{1}{v_t} \rangle)^{-1} \tag{4.108}$$

The sufficient statistic of the auxiliary variable, which are used to estimate the posterior of the gain parameter is:

$$\langle z_t \rangle = \frac{1}{(\alpha_t^z - 1)\beta_t^z} \tag{4.109}$$

We also need to compute the posterior and the sufficient statistics of the inverse of the auxiliary variable which has a Gamma distribution as follows:

$$\frac{1}{z_t} \sim \mathcal{G}(\frac{1}{z_t}; \alpha_t^z, \beta_t^z) \tag{4.110}$$

$$\langle \frac{1}{z_t} \rangle = \beta_t^z \alpha_t^z \tag{4.111}$$

4.4. Temporal Continuity between jingle frames

Up to this point, it is assumed that the background music is generated by choosing an active frame among the jingle frames independent from each other. However, it is more realistic to choose the active frame as dependent on the previous chosen active frame for composing a background music. This strategy enables us to create more

realistic background music. That is, most of the time, the next frame in the background music is the next frame in the jingle. Actually, this model corresponds to applying a Markovian probabilistic structure [49–51] to choose the active jingle frames. This fact yields a different probability structure on active jingle indexes and can be shown mathematically as:

$$r_t = \begin{cases} r_{t-1} + 1 & \text{mod(N) with probability } w, \\ j \in \{1, 2, .., N\} & \text{each with probability} (1-w)/N \end{cases}$$

where r_t represents the active jingle frame at time t and w represents the probability of choosing the next frame of the jingle as active frame. This probabilistic model is shown in Figure 4.4.

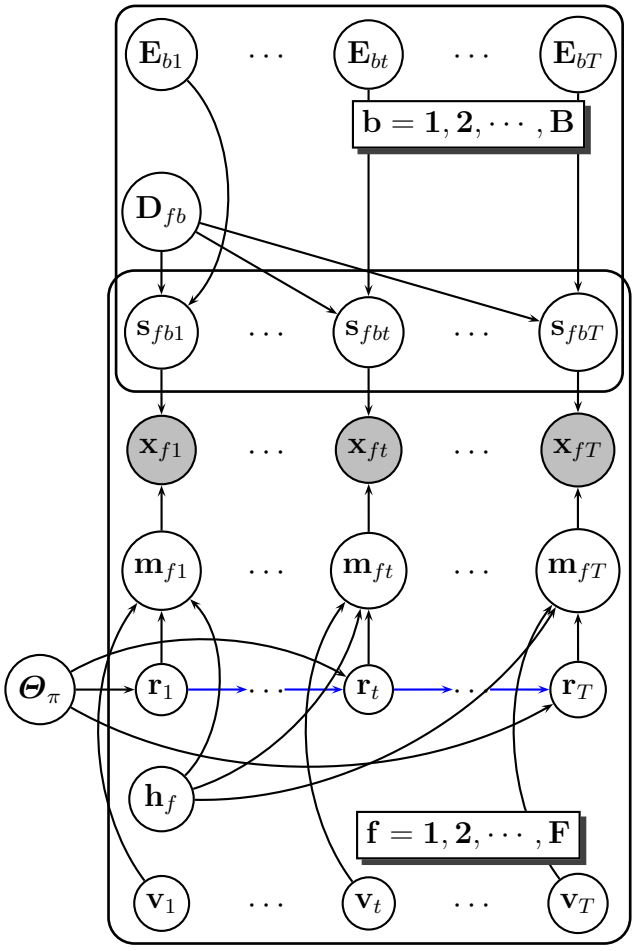

Figure 4.15. Graphical model for speech-music mixture with temporal continuity between jingle frames.

Since we use a probabilistic strategy for speech-music separation problem in this study, this strategy can easily be extended such that it can benefit from the temporal continuity information in the separation process. As in the mixture model, EM method will be used. Different from the mixture model, instead of computing the posterior distribution of the jingle frames independently for each time frame, we have to consider the temporal continuity constraint in this computation. Therefore, with temporal dependent jingle frames, the posterior distribution of the jingle frames must be estimated using the Baum-Welch (BW) algorithm in E step of the method. The remaining update rules are the same as the update rules of the mixture model. The BW algorithm is as follows:

$$p(r_t = j|\mathbf{x}, \theta) = \frac{\alpha_j(t)\beta_j(t)}{\sum_j \alpha_j(t)\beta_j(t)}$$

where

$$\begin{aligned} \alpha_j(t) &= p(\mathbf{x_1}, ..., \mathbf{x_t}, r_t = j) \\ \beta_j(t) &= p(\mathbf{x_{t+1}}, ..., \mathbf{x_T}|r_t = j) \end{aligned}$$

where T represents the total number of frames in mixed signal x.

4.5. Experimental Results

4.5.1. Speech Recognition System and Test Set

For speech recognition tests, we used CMU-Sphinx, an HMM-based continuous density speech recognizer which is trained to recognize Turkish Broadcast News speech. The gender-dependent acoustic models are trained using MFCCs and their deltas and double-deltas calculated in 25ms frames with 10ms shift of the 16 kHz clean speech data. For each gender, 40 hours of speech data is used to train context dependent phone models. The vocabulary size of the recognition system is about 30k. The test set contains 1232 utterances distributed reasonably uniform across 8 speakers. The

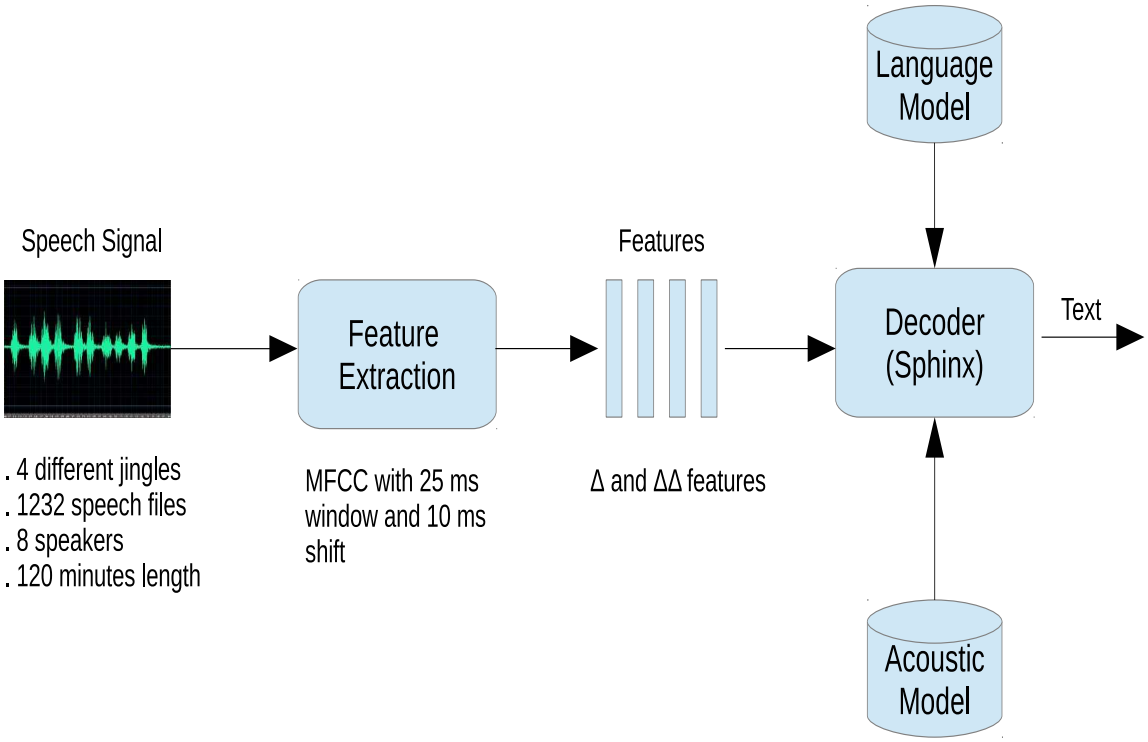

Figure 4.16. ASR system and test set for mixture based separation.

total length of the test set is about 2 hours. The test utterances are mixed with 4 different jingles, length of 4 seconds at different SMR levels to create the test set. For this study, the identity of the jingle is assumed to be known as a prior. The test system is summarized in Figure 4.16.

The background music signal is generated by repeating the jingles up to the length of the speech. The average length of the speech sentences is 6 seconds. The jingles are taken from the broadcast news jingles. The magnitude spectrogram is computed using 1024-point length frames with 50% overlapping. We use this larger window and shift size to reduce the computational complexity of the separation algorithm. The number of speech bases is fixed at 30. The speech recognition performance is measured using WAcc. The baseline speech recognition results with clean and mixed speech are presented in Table 4.1.

Table 4.1. Baseline WAcc values.

Baseline Results	Input SMR Values				
	0dB	5dB	10dB	15dB	20dB
Clean	75.1	75.1	75.1	75.1	75.1
Mixed	0.4	2.6	15.3	40.9	60.4

4.5.2. Evaluation Plan

In our experimental study, the effects of three major factors on the mixture of NMF based separation method are tested. These factors are:

- Divergence Measure (KL or IS): The aim is to compare the effect of divergence measures on the separation performance.
- Temporal Dependency between the jingle frames (Independent (I) or Markovian Dependency (M)): The aim is to analyze the effect of imposing temporal continuity between the jingle frames on the separation performance.
- Gain Estimation Strategy (Ground Truth of the gains (T), the original method (O) and Gamma Chains (G)):

As a complete example of naming, 'IS-M-G' represents separation with IS divergence with Markovian structure between the jingle frames and IGMC is used as the gain estimation method.

4.5.3. Comparison of Observation Model Performances

In this section, we compare the separation performances of the proposed mixture-based approaches with true (known) gain values, which are called as 'IS-I-T' and 'KL-I-T' methods. As a reference, the separation performances of the traditional NMF methods, which are called as 'IS-NMF' and 'KL-NMF', are also measured.

In this part, we use every frame of the spectrogram of the jingle itself as the mixture component in the mixture of NMF model or a template vector in NMF model.

However, it should be noted that no prior speech information is used in the experiments. The SMR, SAR and WAcc values of the methods are shown in Tables 4.2, 4.3 and 4.4, respectively.

Table 4.2. Output SMR values of NMF and mixture based methods.

Separation	Input SMR Values				
Method	0dB	5dB	10dB	15dB	20dB
KL-NMF	**22.1**	**28.5**	**35.3**	**42.7**	**50.6**
IS-NMF	16.5	23.8	31.4	39.3	47.5
KL-I-T	17.9	24.1	30.6	37.9	45.8
IS-I-T	15.9	23.4	31.1	38.9	46.8

Table 4.3. Output SAR values of NMF and mixture based methods.

Separation	Input SMR Values				
Method	0dB	5dB	10dB	15dB	20dB
KL-NMF	10.8	13.4	15.9	18.3	20.4
IS-NMF	11.3	14.6	17.6	20.6	23.4
KL-I-T	10.9	14.2	17.1	20.1	23.0
IS-I-T	**12.1**	**15.2**	**18.1**	**21.1**	**24.1**

Table 4.4. Output WAcc values of NMF and mixture based methods.

Separation	Input SMR Values				
Method	0dB	5dB	10dB	15dB	20dB
KL-NMF	25.6	42.8	56.7	63.8	68.5
IS-NMF	33.7	53.9	64.3	70.2	72.4
KL-I-T	30.9	48.2	60.6	67.5	71.2
IS-I-T	**36.8**	**55.5**	**65.4**	**71.2**	**72.4**

In [7], it was shown that the ASR results with KL-I-T method is better than KL-NMF method. When we examine the results in Tables 4.2, 4.3 and 4.4, we can make the same conclusion for the IS-I-T and IS-NMF methods. In other words, although the SMR values of NMF models are higher than the SMR values of the mixture models, since SAR values of mixture models is better than the SAR values of NMF models,

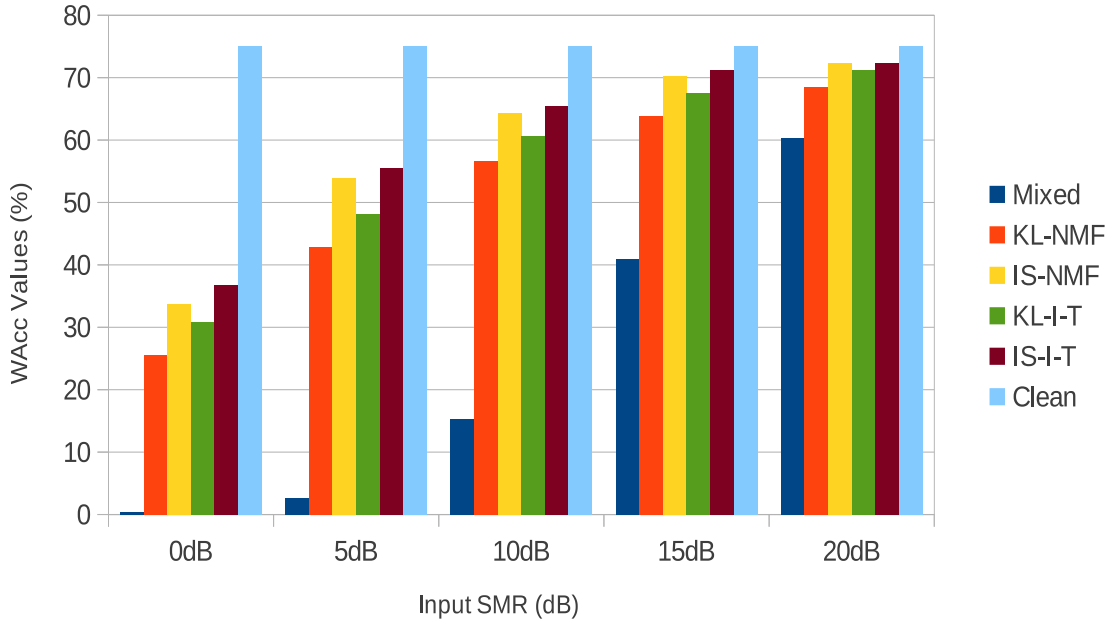

Figure 4.17. ASR performance comparison of mixture and NMF methods.

the speech recognition performance of mixture model outperforms the NMF models (see Figure 4.17). However, it should be noted that the ASR performance difference between the mixture and NMF based approaches with IS divergence is not as high as in KL divergence case.

From these results, it can be concluded that in speech-music separation, preserving the speech signal is more important than suppressing the music signal in speech recognition point of view. With the analysis of the experimental results, using IS divergence (complex Gaussian observation model) in speech-music separation task yields better separation results than KL divergence (Poisson observation model). Using IS divergence in separation decreases the suppression ratio of the music signal. However, since the reconstruction of the speech signal with IS divergence results higher SAR ratios, the speech recognition performances of IS methods (IS-NMF and IS-I-T) are better than KL methods (KL-NMF and KL-I-T) performances. These results suggest that using IS divergence or complex gaussian observation model is more appropriate than KL divergence or poisson observation model for speech-music separation task.

4.5.4. Gamma Markov Chains for Gain Estimation

In [9], the gain estimation problem of the mixture-based approach with KL divergence (KL-I) is pointed out and GMC probabilistic structure was proposed to overcome this problem. In this section, we use both of GMC and IGMC model to enhance the gain estimation performance of the mixture-based approach with KL and IS divergences (KL-I, IS-I). When we compare the gain estimation performances of the divergence measures, it can be concluded that IS divergence method with its original gain estimation update (IS-I-O) has better separation performance than the one with KL divergence (KL-I-O) method.

ASR results with and without GMCs are presented in Figure 4.11 and it can be seen that the proposed gain estimation techniques enhance the speech recognition results. It is very promising that by using the GMC techniques (KL-I-G and IS-I-G) the speech recognition performance can be improved to a level that is very close to the speech recognition performance with the true gain values (KL-I-T and IS-I-T). This can be seen in Figure 4.18 and Table 4.7. Average WAcc of KL-I-O and KL-I-G over all input SMR levels are 35.9% and 53.8% respectively. In other words, the improvement due to the imposing correlation between the gain parameters with GMC is about 50%.

When we make the same analysis for IS case, average WAcc of IS-I-O and IS-I-G over all input SMR levels is 52% and 57.5% respectively. Since baseline performance (IS-I-O) is very high compared to KL-I-O result, the relative improvement of using IGMC for imposing the correlation is about 11% and less than the relative improvement in KL case. With the analysis of SMR and SAR values (See Tables 4.5 and 4.6) the usage of GMC in gain estimation (KL-I-G and IS-I-G) increases both of average SMR and SAR values as compared to the original versions of the methods (KL-I-O and IS-I-O). For example, in KL case, average SMR, SAR values over all input SMR values of KL-I-O and KL-I-G are 22.9, 14.6 and 30.5 and 17.2 dB respectively.

Table 4.5. Output SMR values of mixture based methods with gain estimation strategies.

Output SMR (dB)	Input SMR Values				
Method	0dB	5dB	10dB	15dB	20dB
KL-I-O	4.8	13.7	22.8	32.0	41.5
KL-I-G	**16.6**	**22.9**	29.9	37.4	45.5
IS-I-O	11.5	19.7	28.1	36.7	45.0
IS-I-G	14.3	22.3	**30.3**	**38.3**	**46.4**

Table 4.6. Output SAR values of mixture based methods with gain estimation strategies.

Output SAR (dB)	Input SMR Values				
Method	0dB	5dB	10dB	15dB	20dB
KL-I-O	6.7	11.0	14.9	18.5	22.0
KL-I-G	**11.4**	**14.3**	17.2	20.2	23.1
IS-I-O	9.7	13.2	16.6	20.3	23.3
IS-I-G	10.7	14.3	**17.6**	**20.8**	**23.9**

Table 4.7. Output WAcc values of mixture based methods with gain estimation strategies.

WAcc (%)	Input SMR Values				
Method	0dB	5dB	10dB	15dB	20dB
KL-I-O	3.6	14.8	37.8	56.6	66.8
KL-I-G	27.2	45.9	58.9	66.2	70.7
IS-I-O	19.9	41.0	59.8	68.1	71.2
IS-I-G	**30.5**	**51.1**	**63.8**	**69.6**	**72.3**

4.5.5. Temporal Dependency Experiments

When we analyzed the effect of incorporating temporal continuity information between the jingle frames into the separation framework using Figure 4.19 and Tables 4.8, 4.9, 4.10, it should be noted that, though the temporal continuity information im-

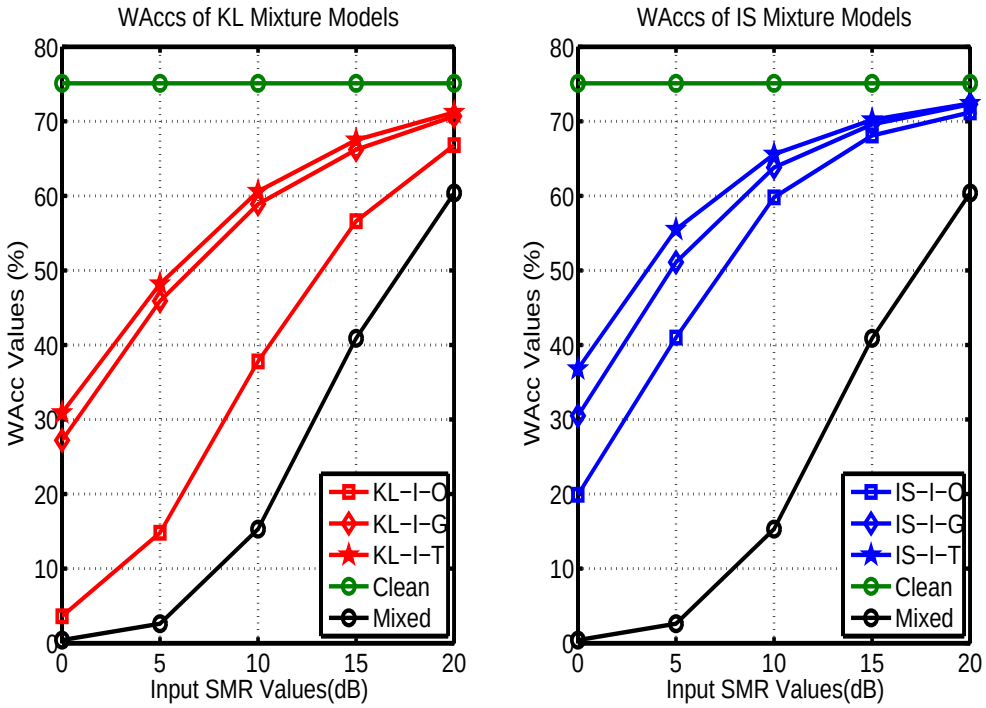

Figure 4.18. Comparison of ASR performances of gain estimation methods.

proves the separation performance for the KL divergence case, its improvement on IS divergence case is almost negligible. Average WAcc of KL-I-O and KL-M-O methods over all input SMR levels are 35.9 and 47.8, the relative improvement of incorporating temporal dependency is about 33% in KL case. In IS case with original updates, the relative improvement of using temporal dependency is 0.5%. The negligible improvement of temporal dependency in IS model can also be observed in SMR and SAR values such that average improvement in SMR and SAR values of temporal dependency (IS-M-O) compared to independent model (IS-I-O) in IS model is 0.3% and 0.4%, respectively. The improvements in KL model with using temporal continuity (KL-M-O) compared to Independent model (KL-I-O) are 20% and 15% in SMR and SAR values.

The effect of temporal dependency with different gain estimation strategies are presented in Figure 4.19. When we analyze the estimated posterior probabilities of the jingle frames for these methods using Figure 4.21, we observed that in IS mixture case, the MPP for each time frame is almost one. MPP for each time frame can be found

as follows:

$$p(r_t)^* = \max_j p(r_t = j|X)$$

Therefore, incorporating temporal dependency does not affect the separation results. However, in KL mixture case, the jingle frames, which have MPP for each time frame, are less crisp and therefore the usage of temporal continuity between the jingle frames improves the separation performance.

Table 4.8. Output SMR values of mixture based methods with temporal dependency and gain estimation strategies.

Separation Method	Input SMR Values				
	0dB	5dB	10dB	15dB	20dB
KL-M-T	**21.7**	**28.6**	**35.5**	**42.6**	**49.8**
KL-M-O	11.2	19.5	27.6	35.7	44.2
KL-M-G	21.1	27.9	34.8	41.9	49.3
IS-M-T	15.9	23.5	31.1	38.9	46.9
IS-M-O	11.6	19.8	28.4	36.7	45.1
IS-M-G	14.6	22.5	30.5	38.4	46.5

Table 4.9. Output SAR values of mixture based methods with temporal dependency and gain estimation strategies.

Separation Method	Input SMR Values				
	0dB	5dB	10dB	15dB	20dB
KL-M-T	11.7	14.5	17.4	20.3	23.2
KL-M-O	9.8	13.8	17.2	20.2	23.1
KL-M-G	11.8	14.6	17.5	20.4	23.3
IS-M-T	**12.2**	**15.2**	**18.2**	**21.1**	**24.1**
IS-M-O	9.8	13.3	16.8	20	23.3
IS-M-G	10.9	14.4	17.7	20.8	23.9

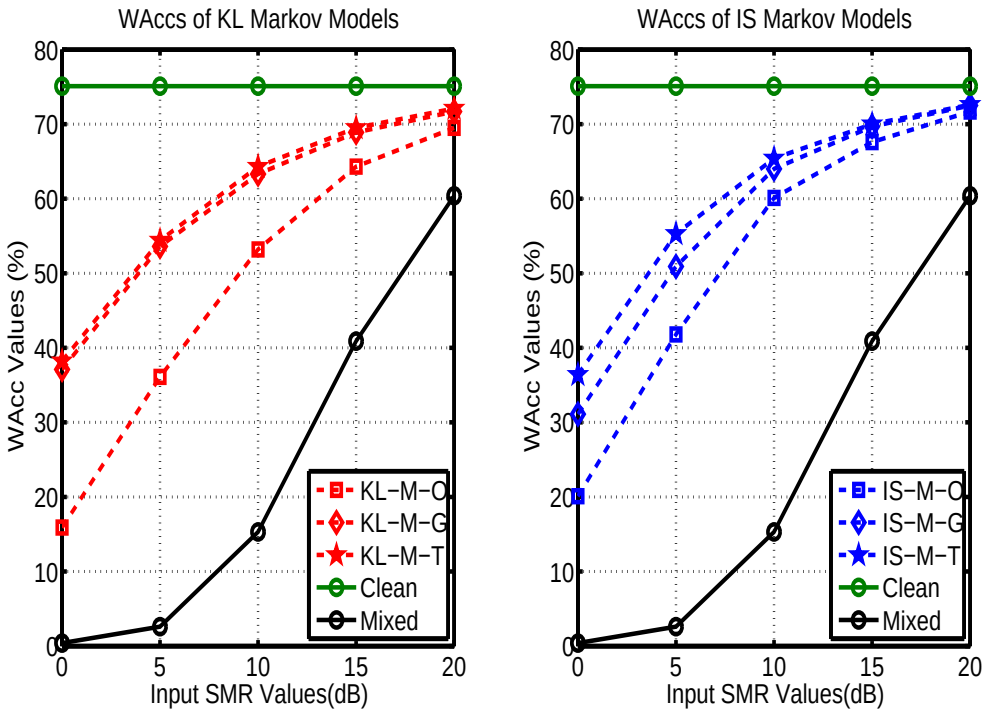

Figure 4.19. Comparison of ASR performances with temporal dependency

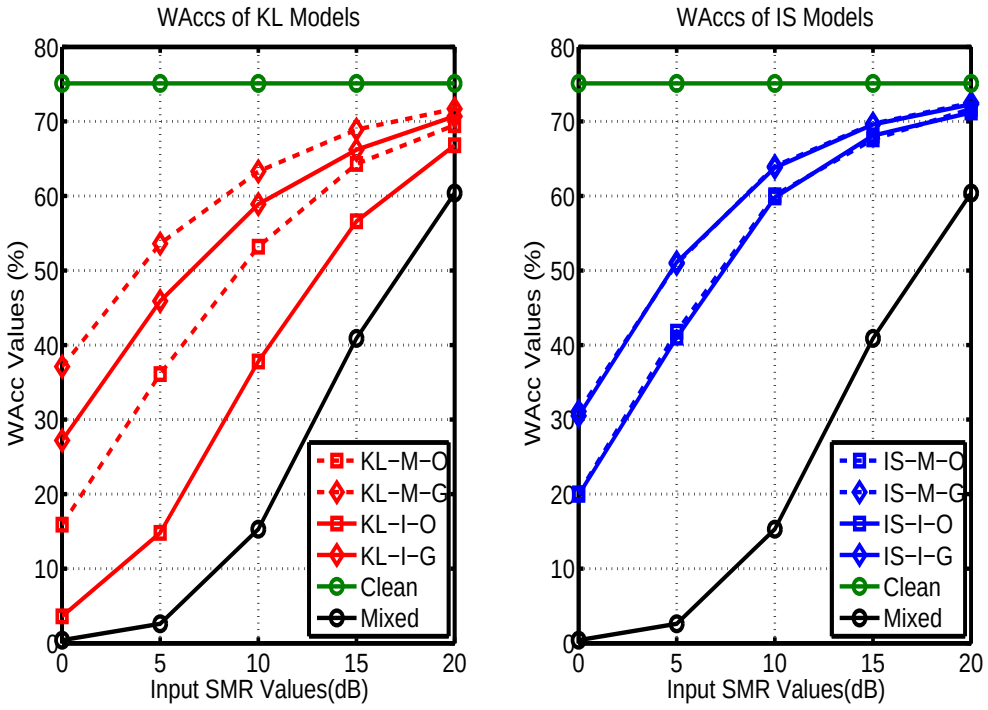

Figure 4.20. Comparison of ASR performances with temporal dependency and gain estimation.

Table 4.10. Output WAcc values of mixture based methods with temporal dependency and gain estimation strategies.

Separation Method	Input SMR Values				
	0dB	5dB	10dB	15dB	20dB
KL-M-T	**38.2**	54.4	64.3	69.5	72.1
KL-M-O	15.9	36.1	53.2	64.3	69.5
KL-M-G	37.1	53.6	63.3	68.9	71.7
IS-M-T	36.4	**55.3**	**65.4**	**70.0**	**72.6**
IS-M-O	20.1	41.8	60.1	67.6	71.7
IS-M-G	31.1	50.9	64.0	69.7	72.5

When we compare the effect of using Markovian structure on the jingle frames and GMC on the gain parameters (see Figure 4.20), it can be noted that imposing the correlation between the gain values with GMC has more impact on the separation performance. For example, average WAcc value over all input SMR values of KL-I-O is 35.9%. When we applied the GMC on the gain values (KL-I-G), the average WAcc value increases to 53.8%. The average WAcc with applying Markovian structure on the jingle frames (KL-M-O) is 47.8%. The relative improvement of using GMC with KL method (KL-I-G) with respect to mixture-based method with original gain update (KL-I-O) is about 50%. However, the relative improvement with temporal dependency constraint is 33.1%. In IS divergence case, as previously noted, temporal dependency has negligible effect on the ASR results as compared to the effect of imposing IGMC on the gain parameters.

4.5.6. Computational Complexity Analysis

The real time (RT) factors of the separation methods are listed in Table 4.11. It is shown that the computational cost of the mixture-based methods are N times higher than the traditional NMF based method, where N is the number of frames in the jingle dictionary. Consequently, in each iteration of the mixture-based method, NMF multiplicative updates are computed for each catalog frame to compute the mixture

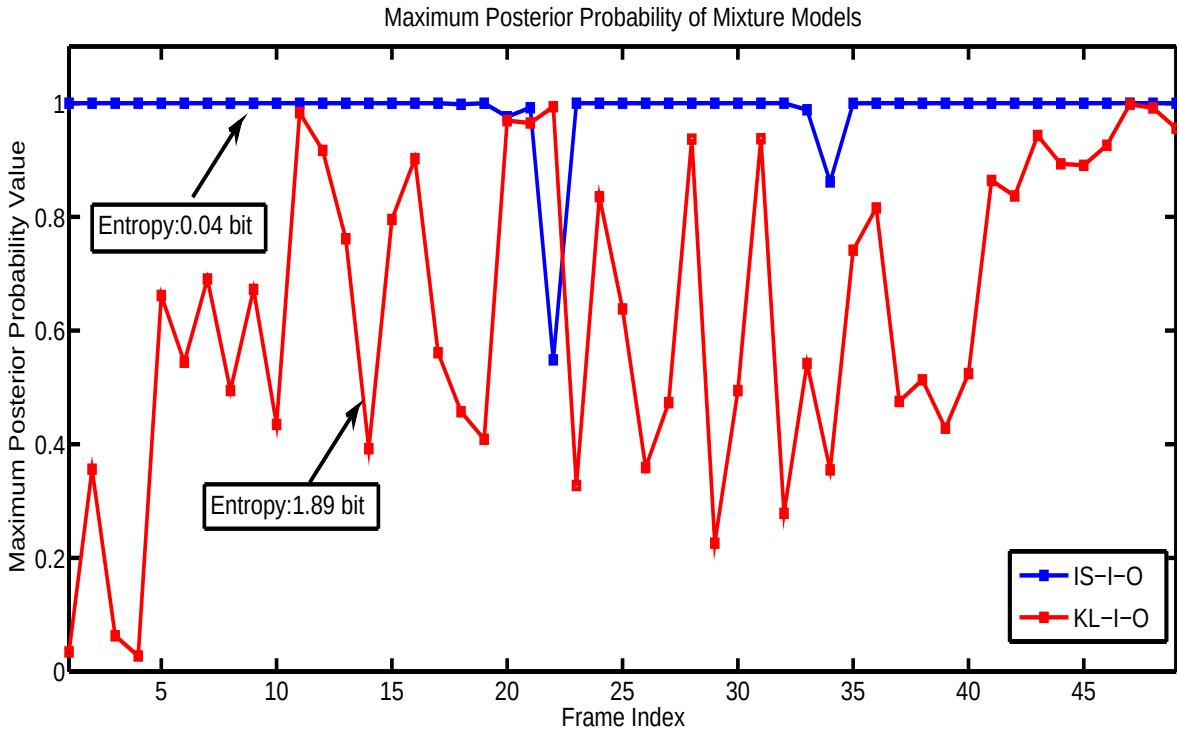

Figure 4.21. Comparison of MPP values for divergence measures.

weights. It should be noted that although the computational cost of the gain estimation strategies are almost the same, Markovian structure on the jingle frames increases the computation time. However, since the proposed mixture-based method can be run for each jingle frame independently, the computation time of the mixture-based methods can be decreased via parallel processing.

Table 4.11. Real time factors of NMF and mixture based methods.

Divergence	Separation Methods				
Measures	NMF	I-O	I-G	M-O	M-G
KL	0.04	2.5	2.5	6.3	6.3
IS	0.08	5.5	5.5	9.5	9.5

4.5.7. More Gain Estimation Strategies for Poisson Observation Model

In this section, the effects of proposed gain estimation techniques to the separation performance are analyzed and compared. As an example for comparing estimation

performances of the methods, the estimated gain parameter values for constant and fading gain cases are shown in Figure 4.22 and 4.23.

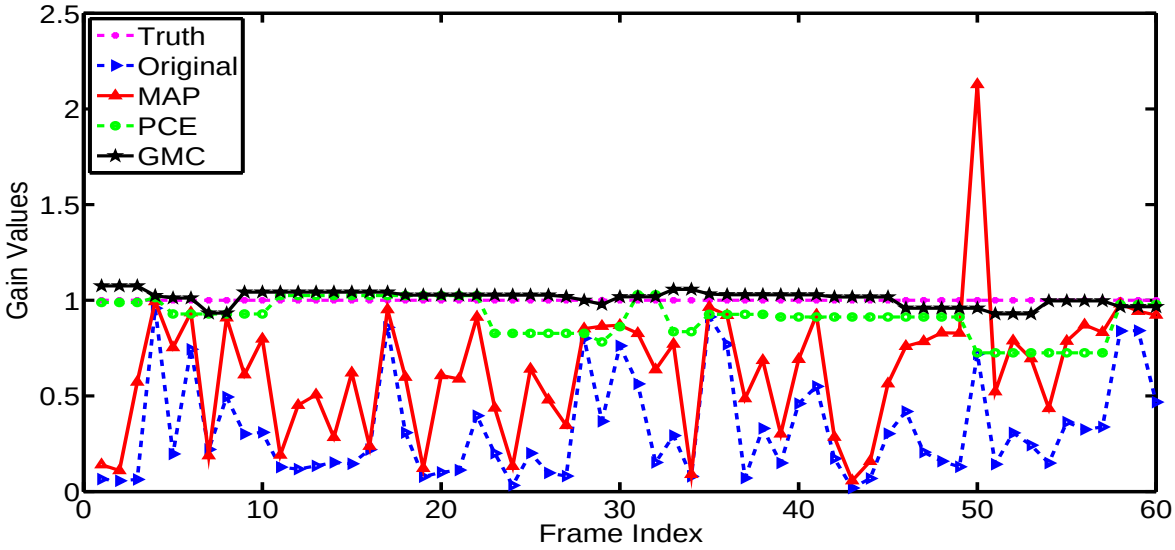

Figure 4.22. Estimation of constant gain parameter.

When we examine the gain estimation results with the original method in Table 4.12, 4.13 and corresponding estimated gain parameter in Figure 4.22, it is observed that when the speech signal suppresses the music signal, the gain parameter is under-estimated. Therefore, music signal is contaminated in speech signal and so SMR value of the original method is very low compared to "Truth" case. In this part, the original method corresponds to estimating the gain parameter by Equation 4.38. When we use MAP, since some of the frames with low MAP are actual active frames, the estimated gain parameters for these frames are increased, so the SMR and SAR values are higher compared to the original case. However, the frames for those active frames are not correctly identified, the gain parameter is over or under estimated. In PCE case, the gain parameter is estimated using the frames which have high MAP, so the gain parameter of the frames are smoothed over these frames. As a result, the gain estimation performance increases as compared to the original case. In GMC method, by imposing correlation between the gain parameter along the frames, it is not allowed to have abrupt changes in the estimated values. This scenario is more realistic because the gain parameter is not changed instantaneously in real life. ASR results with different gain estimation techniques are presented in Table 4.14 and it is experimentally

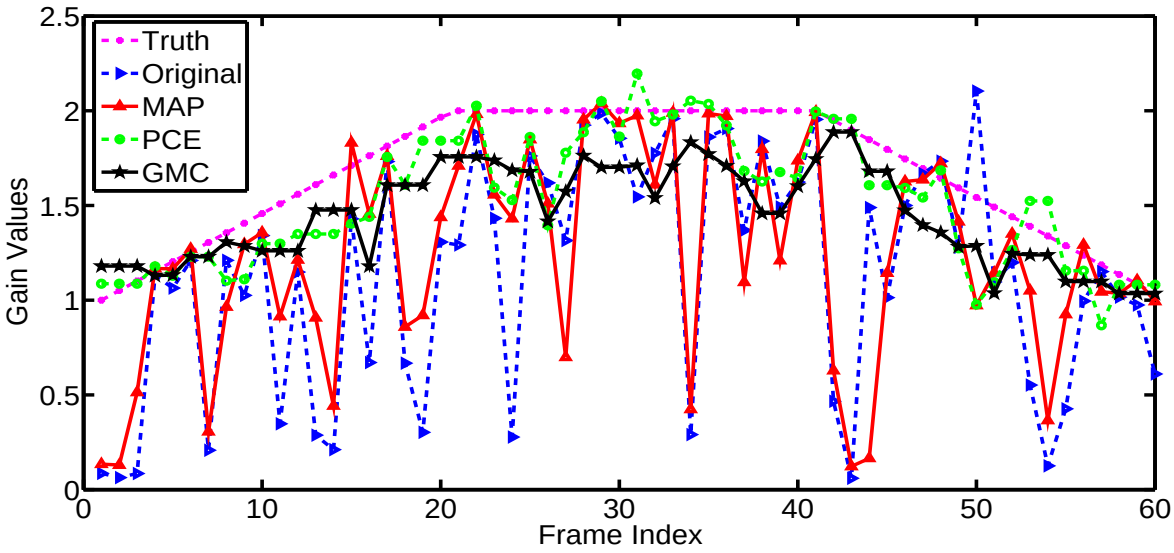

Figure 4.23. Estimation of fading gain parameter.

Table 4.12. Average SMR values.

Output SMR (dB)	Input SMR Values				
Method	0dB	5dB	10dB	15dB	20dB
Truth	17.6	24.2	32.1	38.2	46.2
Original	6.5	15.3	24.3	33.5	42.8
MAP	12.5	20.3	28.4	36.8	45.3
PCE	15.3	22.5	30.1	37.9	46.2
GMC	18.5	24.6	31.4	38.8	46.7

shown that the proposed gain estimation techniques enhance the speech recognition results. It is very promising that by using the proposed techniques the speech recognition performance can be improved to a level that is very close to the speech recognition performance with the true gain values as can be shown in Table 4.14.

As a conclusion, in this section, we address the gain estimation problem of the mixture-based method and propose three different solutions to this problem. MAP and PCE methods are ad-hoc methods which we developed by analyzing the reasons behind estimation errors. Also we applied GMC structure to overcome this gain estimation problem. It is shown that all of these enhancement techniques improves the gain estimation performance of the mixture-based method. Moreover, by using the proposed

Table 4.13. Average SAR values.

Output SAR (dB)	Input SMR Values				
Method	0dB	5dB	10dB	15dB	20dB
Truth	10.9	14.2	17.2	20.2	23.2
Original	7.5	11.4	14.9	18.4	21.9
MAP	10.5	13.8	16.8	19.6	22.3
PCE	11.7	14.5	17.3	20.1	22.9
GMC	11.7	14.4	17.1	20.1	23.1

Table 4.14. Average WAcc values.

WAcc (%)	Input SMR Values				
Method	0dB	5dB	10dB	15dB	20dB
Truth	29.2	46.7	59.4	66.9	70.4
Original	3.5	14.8	41.7	52.4	66.6
MAP	13.7	35.1	48.8	61.2	69.5
PCE	17.6	35.9	55.6	63.6	70.0
GMC	26.1	41.5	57.3	63.4	70.9

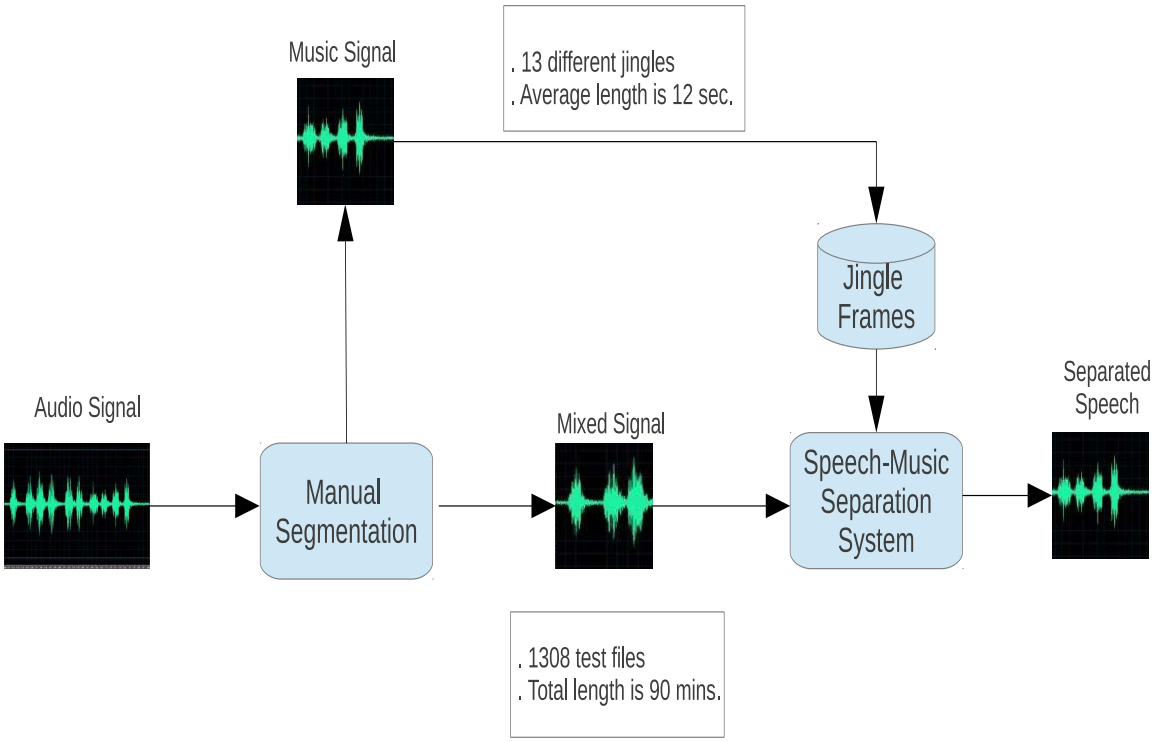

Figure 4.24. Real data experiment setup.

approaches, the separation performance can be improved as if the truth gain values are used in the separation process.

4.5.8. Real World Data Experiments

The separation performance of the proposed method is evaluated with real data samples. In this case, we do not have the speech and music signals separately. Therefore, the separation performance of the methods cannot be measured via SMR and SAR values. We evaluated the separation performances of the proposed methods using speech recognition performance with real data recordings. When working with real data, since we do not have the clean speech data, we cannot calculate the baseline speech recognition results. Therefore, in real world data experiments, the speech recognition result with the mixed speech is used as the baseline result. The real data experiment setup is shown in Figure 4.24.

The real data set contains 1308 sentences which are taken from broadcast news

recordings. The data length is about 90 minutes. The total length of 13 different jingles is about 161 seconds consisting segmented music parts. We labeled the real recordings as speech-music mixture and music parts and then use the music segments as the jingle. In other words, the jingle which will be used in the separation is taken from the audio itself. This approach is very useful and practical because of the fact that all information needed for the separation is obtained from the audio itself.

Table 4.15. Average WAcc values of real data.

Method	WAcc
Mixed	42.8
logMMSE	41.8
logMMSE-SPU	37.6
KL-NMF	37.9
IS-NMF	45.1
KL-M-G	48.2
IS-M-G	47.9

Instead of giving all speech recognition results, the best methods of KL and IS methods are chosen and presented. When we examine the real data results in Table 4.15, it is observed that the baseline WAcc result is very close to 15dB results in Table 4.1. However, the improvement with real data is not as much as in artificially mixed data. The main reason for that in artificially mixed case, the whole of the jingle that generates the background music is known, but in real data case, we do not know how much of the background jingle is segmented as music. Therefore, it is not easy to compare the speech recognition results of real and artificial data sets. However, in the experiments, the speech recognition performance is improved using the catalog based approaches as compared to the mixed signal.

It is quite surprising that though KL-NMF method improves the speech recognition result in artificial data set, the speech recognition result on separated signal with KL-NMF method is worse than the speech recognition performance with the mixed sig-

nal. It is also interesting that the speech recognition performance of IS-NMF method is almost the same as mixture-based methods in the artificial case. However, with the real data, its improvement on the speech recognition performance is 5.3% which is low compared to the speech recognition performance improvement of IS-M-G method. The relative improvement of the IS-M-G method is 11.9% in the real data set. It should be noted that IS-M-G and KL-M-G methods are the two methods which have the best ASR performance in artificial and real data recordings.

Speech-music separation problem can be regarded as a speech enhancement problem. For making a comparison of speech enhancement techniques with the proposed methods, we used the logMMSE [52] and logMMSE with signal-presence uncertainty (SPU) (logMMSE-SPU) [53,54] to recover the speech signal. In Table 4.15, the speech recognition result with the speech enhancement techniques are reported. Both enhancement techniques degrades the speech recognition performance as compared to the mixed signal. In order to make a fair comparison between the speech enhancement techniques and mixture-based methods, each jingle of the mixture is assumed to be known as the noise signal. Prior noise statistics are estimated from these parts.

In order to analyze the improvement in speech recognition performance with the mixture-based speech-music separation method, WAcc value for each sentence in the real data set with and without applying separation method are shown in Figure 4.25. While recognition accuracies in 328 sentences are increased with the proposed method, recognition accuracies in 209 sentences are decreased with the method. The recognition accuracies in the rest of the sentences are unchanged.

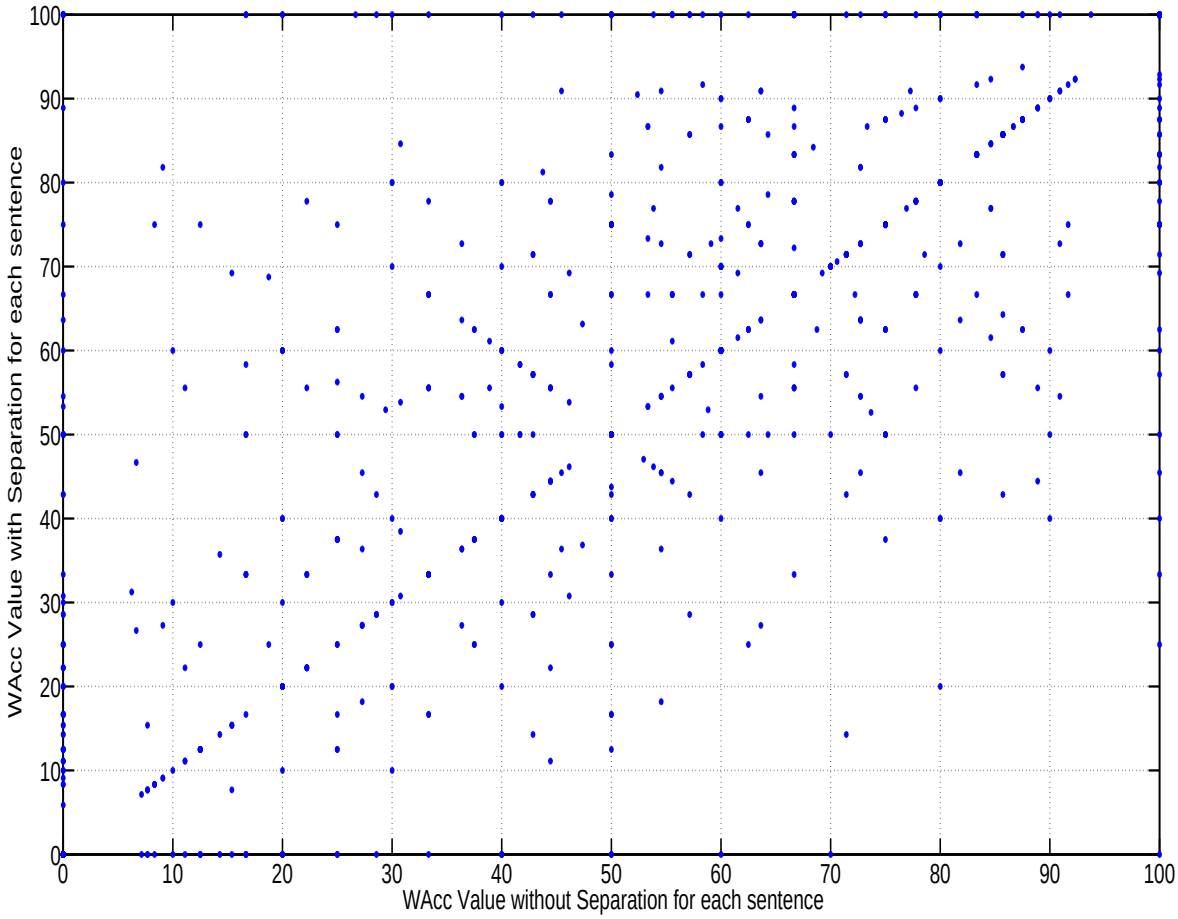

Figure 4.25. Comparison of WAcc values of mixed speech and separated speech with KL-M-G method for each sentence.

5. SPEECH MODELING FOR SPEECH-MUSIC SEPARATION

In Chapter 4, a single-channel speech-music separation algorithm, which uses a mixture model for the music signal and an NMF model for the speech signal, is developed. Although the speech signal is assumed to be generated by an NMF model, the parameters of the model are estimated from the mixture in an unsupervised manner. In other words, no training data is used to estimate the parameters of the speech source signal. From ASR point of view, in training phase of the acoustic model, there are plenty of the speech data which can be used for learning the templates of the speech signals that are used in the separation process. Moreover, pure speech segments in incoming audio signal can also be used for learning the parameters of the speech model. This scenario is shown in Figure 5.1. There are two ways of using prior speech data in the separation process:

- Fixed Template Matrix Case: The prior speech data can be used for learning the parameters of an NMF model for the speech source signal such that the trained parameter, template matrix, is used as a speech model in the separation phase. The disadvantage of this approach is that it does not take speech data in consideration with the mixed signal. This strategy can be applied via fixing the template matrix entries in Section 4.1. A similar strategy which uses NMF modeling approach for both the speech and music signals is applied in Chapter 3.
- Prior Model For Template Matrix Case: Another way of using the prior speech data to train the signal model is to learn the hyper-parameters of the template matrix of the speech model. In the separation process, the posterior distribution of the template matrix entries are calculated using the prior model and the observed signal (mixed signal). The posterior distribution is used for separating the speech from the music. For the Poisson observation model, a Gamma distribution is used as a prior while an Inverse-Gamma distribution is used for the

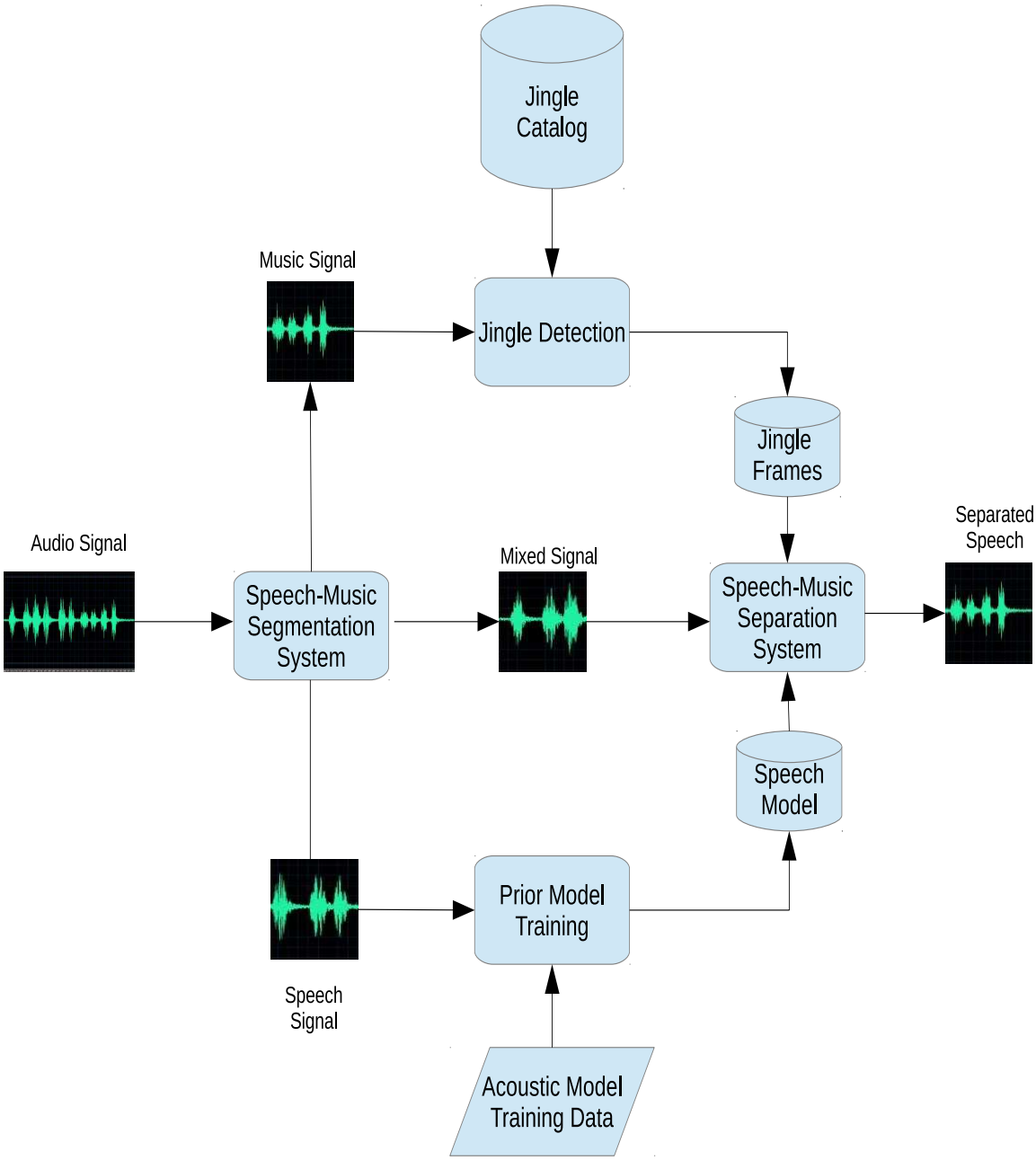

Figure 5.1. Catalog based speech-music separation system framework with prior speech Models.

complex-Gaussian model due to the conjugacy between distributions.

Now, we extend the proposed mixture of NMF model based approaches (KL and IS cases in Chapter 4) such that they can use a prior speech model in the separation process.

5.1. Gamma Priors for the Poisson Model

5.1.1. Model Description

For the Poisson observation model, a prior model for each entry of the template matrix of the speech signal has a Gamma distribution in the form of:

$$D_{fb} \sim \mathcal{G}(D_{fb}; a_{fb}, b_{fb}) \tag{5.1}$$

where a_{fb}, b_{fb} are the hyper-parameters of the template matrix. A Gamma distribution is defined as:

$$\mathcal{G}(x; a, b) = \exp((a-1)\log x - \frac{x}{b} - a\log b - \log\Gamma(a)). \tag{5.2}$$

The hyper-parameters of the template matrix are estimated using the prior speech training data. Although we used a different scale and variance parameter for each entry of the template matrix, they can be coupled for decreasing the number of parameter in the prior model. The rest of the probabilistic model for the speech-music mixture is the same as in Section 4.1. The overall probabilistic model which includes the speech prior is shown in Figure 5.2. Θ_D represents the hyper-parameters of the template matrix, a_{fb} and b_{fb}.

5.1.2. Estimation of Hyper-parameters

In order to estimate the hyper-parameters of the Gamma distribution for the speech templates, the probabilistic interpretation of KL-NMF model similar to Section 3.2.1 is used. The magnitude spectrum of the training data for the speech signal is represented as $\mathbf{S}$ which is equal to the sum of the Poisson sources as follows:

$$S_{ft} = \sum_b s_{fbt} \tag{5.3}$$

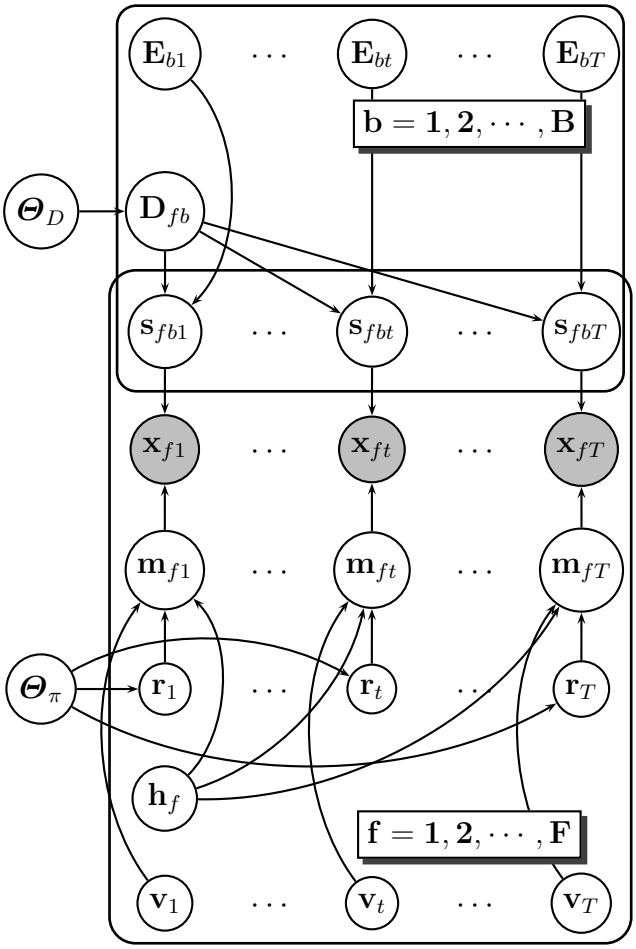

Figure 5.2. Graphical model for speech-music mixture with speech priors.

where s_{fbt} represents the b'th speech source variable at the f'th frequency bin and the t'th time frame.

$$s_{fbt} \sim \mathcal{PO}_c(s_{fbt}; D_{fb}E_{bt}) \tag{5.4}$$

The entry of the template matrix has a Gamma prior distribution which is defined as Equation 5.1. The overall joint log-likelihood of the latent sources, s_{fbt}, random variable, D_{fb}, and the data can be defined as:

$$\phi = p(\mathbf{S}, s, \mathbf{D}|\mathbf{E}, a_{fb}, b_{fb}) = p(\mathbf{S}|s)p(s|\mathbf{D}, \mathbf{E})p(\mathbf{D}|a_{fb}, b_{fb}) \tag{5.5}$$

$$\begin{aligned}\log\phi &= \sum_{f,b,t}\Big[-D_{fb}E_{bt}+s_{fbt}\log(D_{fb}E_{bt})-\log\Gamma(s_{fbt}+1)+(a_{fb}-1)\log D_{fb}\\ &- \frac{1}{b_{fb}}D_{fb}-\log\Gamma(a_{fb})-a_{fb}\log(b_{fb})+a_{fb}\log(a_{fb})+\log\delta(S_{ft}-\sum_{b}s_{fbt})\Big]\end{aligned}$$

Since the posterior distributions of the entries of the template, D_{fb} and the latent speech sources, s_{fbt}, are coupled, we cannot compute the overall joint posterior distribution exactly. In this case, we use the variational technique that factorizes the posterior distribution into the posteriors of the decoupled random variables as follows:

$$q(s) \propto \exp(\langle\log p(\mathbf{S},s,\mathbf{D}|\mathbf{E})\rangle_{q(\mathbf{D})}) \tag{5.6}$$

$$q(\mathbf{D}) \propto \exp(\langle\log p(\mathbf{S},s,\mathbf{D}|\mathbf{E})\rangle_{q(\mathbf{s})}) \tag{5.7}$$

Moreover, by updating the entries of the excitation matrix such that it maximizes the expected value of the joint log-likelihood as follows:

$$\mathbf{E} = \arg\max_{\mathbf{E}}(\langle\log\phi\rangle_{q(s)q(\mathbf{D})}) \tag{5.8}$$

Posterior distribution of the latent speech sources given the training data can be calculated as follows:

$$q(s_{f1t},...,s_{fBt}) \propto \mathcal{M}(s_{f1t},...,s_{fBt};S_{ft},p_{f1t},...,p_{fBt}) \tag{5.9}$$

with posterior cell probabilities of the latent speech sources:

$$p_{fbt} = \frac{\exp(\langle\log D_{fb}\rangle)E_{bt}}{(\sum_b\exp(\langle\log D_{fb}\rangle)E_{bt}} \tag{5.10}$$

Marginal expectation of the latent speech sources can be calculated using the posterior parameters as:

$$\langle s_{fbt}\rangle = S_{ft}p_{fbt}$$

The posterior distributions of the entries of the template matrix are Gamma distribution due to the conjugacy between Gamma and Poisson distributions and can be calculated as follows:

$$\begin{aligned} q(D_{fb}) &\propto \exp((a_{fb} + \sum_t \langle s_{fbt} \rangle - 1) \log D_{fb} - (\frac{1}{b_{fb}} + \sum_t E_{bt}) D_{fb}) \\ &\propto \mathcal{G}(D_{fb}; \alpha_{fb}, \beta_{fb}) \end{aligned}$$

with parameters:

$$\begin{aligned} \alpha_{fb} &= a_{fb} + \sum_t \langle s_{fbt} \rangle \\ \beta_{fb} &= (\frac{1}{b_{fb}} + \sum_t E_{bt})^{-1} \end{aligned}$$

The sufficient statistics of the template matrix entries which are used for calculating the other parameters in the model are:

$$\begin{aligned} \exp(\langle \log D_{fb} \rangle) &= \exp(\Psi(\alpha_{fb})) \beta_{fb} \\ \langle D_{fb} \rangle &= \alpha_{fb} \beta_{fb} \end{aligned}$$

The update equation for the entries of the excitation matrix are found by using the following equation:

$$E_{bt} = \frac{\sum_b \langle s_{fbt} \rangle}{\sum_f \langle D_{fb} \rangle} \tag{5.11}$$

By using the training data, the hyper-parameters of the entries of the template matrix are calculated. Hence, in the separation phase, pre-computed hyper-parameters are used as a prior model for the speech signal.

5.1.3. Separation Method with Gamma Priors

In this section, we describe the inference technique used for deriving the update equations of the posterior distributions of the latent sources and parameters of the speech and music signals in the probabilistic model. Since the posterior distributions of the template matrix, $\mathbf{D}$ and the latent speech, music and active jingle frame index sources, s, m and r are coupled, we cannot compute the overall posterior distribution exactly. Therefore, we use the variational inference technique that factorizes the posterior distribution into the posteriors of the decoupled random variables as follows:

$$\begin{aligned} q(s,m,r) &\propto \exp(\langle \log \phi \rangle_{q(\mathbf{D})}) \\ q(\mathbf{D}) &\propto \exp(\langle \log \phi \rangle_{q(s,m,r)}) \\ \mathbf{E}^* &\propto \arg\max_{\mathbf{E}} (\langle \log \phi \rangle_{q(s,m,r)q(\mathbf{D})}) \end{aligned}$$

where ϕ represents the joint likelihood of the data and the latent sources and Θ represents the parameters in the model such as $a_{fb}, b_{fb}, \pi, h, v$. The joint likelihood can be decomposed as follows:

$$\phi = p(\mathbf{X}, s, m, \mathbf{D}, \mathbf{E}, r, f, v|\Theta) = p(\mathbf{X}|s,m)p(s|\mathbf{D},\mathbf{E})p(\mathbf{D}|a_{fb},b_{fb})p(m|r,h,v)p(r|\pi) \tag{5.12}$$

The joint log-likelihood can be written as:

$$\begin{aligned} \log \phi &= \sum_{t,j} [r_t = j] \Bigg[\sum_{fb} (-D_{fb}E_{bt} + s_{fbt} \log(D_{fb}E_{bt}) - \log \Gamma(s_{fbt}+1) \\ &+ (a_{fb}-1)\log D_{fb} - \frac{1}{b_{fb}} D_{fb} - \log \Gamma(a_{fb}) - a_{fb}\log(b_{fb}) + a_{fb}\log(a_{fb}) \Bigg] \\ &+ \sum_{t,j} [r_t = j] \Bigg[\sum_{f} (-C_{fj}h_f v_t + m_{ft} \log(C_{fj}h_f v_t) - \log \Gamma(m_{ft}+1)) \\ &+ \log \delta(X_{ft} - \sum_b s_{fbt} - m_{ft}) + \log \pi_j \Bigg] \end{aligned}$$

The posterior distribution of the latent sources in the model is:

$$q(s, m, r) = q(s, m|r)q(r) = q(s_{f1t}, ..., s_{fBt}, m_{ft}|r)q(r) \tag{5.13}$$

With prior model for each entry of template matrix of the speech signal, the expectation of log values of the speech template entries are used in the posterior distribution calculation as:

$$\begin{aligned} q(s_{f1t}, ..., s_{fBt}, m_{ft}|r) \quad &\propto \quad \exp\Bigg(\sum_b (s_{fbt}(\langle \log D_{fb}\rangle E_{bt}) - \log\Gamma(s_{fbt}+1)) \\ &+ \quad m_{ft}(\log C_{fj}h_f v_t) - \log\Gamma(m_{ft}+1)\Bigg)\delta(X_{ft} - \sum_b s_{fbt} - m_{ft}) \end{aligned}$$

As in Chapter 4, the posterior distribution is a MMM as shown below:

$$q(s_{f1t}, ..., s_{fBt}, m_{ft}|r) \propto \mathcal{M}(s_{f1t}, ..., s_{fBt}, m_{ft}; X_{ft}, p^j_{f1t}, ..., p^j_{fBt}, p^j_{ft}) \tag{5.14}$$

with posterior cell probabilities of the latent speech sources:

$$p^j_{fbt} = \frac{\exp(\langle \log D_{fb}\rangle)E_{bt}}{(\sum_b \exp(\langle \log D_{fb}\rangle)E_{bt} + C_{fj}h_f v_t} \tag{5.15}$$

and with cell probabilities of the latent music sources:

$$p^j_{ft} = \frac{C_{fj}h_f v_t}{(\sum_b \exp(\langle \log D_{fb}\rangle)E_{bt} + C_{fj}h_f v_t}. \tag{5.16}$$

The posterior distribution on the jingle frame indexes can be written as:

$$q(r_t|\mathbf{X}) = \frac{\mathcal{PO}(X_{ft}; \sum_b \langle D_{fb}\rangle E_{bt} + C_{fj}h_f v_t)\pi_j}{\sum_j \mathcal{PO}(X_{ft}; \sum_b \langle D_{fb}\rangle E_{bt} + C_{fj}h_f v_t)\pi_j} = \langle [r_t = j]\rangle \tag{5.17}$$

The marginal expectation of the latent sources are found as:

$$\langle s_{fbt} \rangle = X_{ft}(\sum_j \langle [r_t = j] \rangle p^j_{fbt}) \tag{5.18}$$

$$\langle m_{ft} \rangle = X_{ft}(\sum_j \langle [r_t = j] \rangle p^j_{ft}). \tag{5.19}$$

Since the entries of the template matrix are random variables with prior distribution, the posterior distribution of the entries must be calculated using the following equations:

$$q(D_{fb}) \propto \exp((a_{fb} + \sum_t \langle s_{fbt} \rangle - 1) \log D_{fb} - (\frac{1}{b_{fb}} + \sum_t E_{bt}) D_{fb}) \tag{5.20}$$

$$\propto \mathcal{G}(D_{fb}; \alpha_{fb}, \beta_{fb}) \tag{5.21}$$

with the posterior parameters are:

$$\alpha_{fb} = a_{fb} + \sum_t \langle s_{fbt} \rangle \tag{5.22}$$

$$\beta_{fb} = (\frac{1}{b_{fb}} + \sum_t E_{bt})^{-1} \tag{5.23}$$

where α_{fb} and β_{fb} are shape and scale parameters of the posterior distribution of the template entries. The expectation of the template entry variable and its log value are needed for the calculation of the other parameters and the posterior distribution. They can be found using the calculated α_{fb} and β_{fb} values as:

$$\exp(\langle \log D_{fb} \rangle) = \exp(\Psi(\alpha_{fb}))\beta_{fb} \tag{5.24}$$

$$\langle D_{fb} \rangle = \alpha_{fb}\beta_{fb} \tag{5.25}$$

The excitation matrix for the speech signal can be found by maximizing the expectation of the joint log-likelihood. The update equation for the excitation matrix entries are

found using the following equation:

$$E_{bt} = \frac{\sum_{f,j} \langle [r_t = j] \rangle \langle s^j_{fbt} \rangle}{\sum_f \langle D_{fb} \rangle} \tag{5.26}$$

The update equations for gain and filtering parameters are the same as in Chapter 4.

5.2. Inverse-Gamma Priors for Complex Gaussian Model

5.2.1. Model Description

For the complex Gaussian observation model, the prior model for each entry of the template matrix of the speech signal is an Inverse-Gamma and described as:

$$D_{fb} \sim \mathcal{IG}(D_{fb}; a_{fb}, b_{fb}) \tag{5.27}$$

where a_{fb}, b_{fb} are the hyper-parameters of the template matrix. Inverse-Gamma distribution is defined as:

$$\mathcal{IG}(x; a, b) = \exp(-(a+1)\log x - \frac{1}{bx} - a \log b - \log \Gamma(a)) \tag{5.28}$$

The rest of the probabilistic model for the speech-music mixture is the same as in Section 4.1. The corresponding graphical model is shown in Figure 5.2.

5.2.2. Estimation of Hyper-parameters

In order to estimate the hyper-parameters of Inverse-Gamma distribution for the speech templates, the probabilistic interpretation of IS-NMF similar to Section 3.2.2 is used. The complex spectrum of the training data for the speech signal is represented as $\mathbf{S}$ which is equal to the sum of the complex Gaussian sources as follows:

$$S_{ft} = \sum_b s_{fbt} \tag{5.29}$$

where s_{fbt} represents the b'th speech source variable at the f'th frequency bin and the t'th time frame.

$$s_{fbt} \sim \mathcal{N}_c(s_{fbt}; D_{fb}E_{bt}) \tag{5.30}$$

The entry of the template matrix has an Inverse-Gamma prior distribution which is defined as Equation 5.27. The overall joint log-likelihood of the latent sources, s_{fbt}, random variable, D_{fb}, and the data can be defined as:

$$\phi = p(\mathbf{S}, s, \mathbf{D}|\mathbf{E}, a_{fb}, b_{fb}) = p(\mathbf{S}|s)p(s|\mathbf{D}, \mathbf{E})p(D|a_{fb}, b_{fb}) \tag{5.31}$$

$$\log \phi \quad = \quad \sum_{f,b,t} \left(- \log D_{fb}E_{bt} - \frac{|s_{fbt}|^2}{D_{fb}E_{bt}}) + (-a_{fb} - 1) \log D_{fb} - \frac{1}{b_{fb}D_{fb}} - \log \Gamma(a_{fb}) - a_{fb} \log(b_{fb} \right)$$

Since the posterior distributions of the entries of the template, D_{fb} and the latent speech sources, s_{fbt}, are coupled, we cannot compute the overall joint posterior distribution exactly. In this case, we use the variational technique that factorizes the posterior distribution into the posteriors of the decoupled random variables as follows:

$$q(s) \quad \propto \quad \exp(\langle \log p(\mathbf{S}, s, \mathbf{D}|\mathbf{E}) \rangle_{q(\mathbf{D})}) \tag{5.32}$$

$$q(\mathbf{D}) \quad \propto \quad \exp(\langle \log p(\mathbf{S}, s, \mathbf{D}|\mathbf{E}) \rangle_{q(\mathbf{s})}) \tag{5.33}$$

Moreover, by updating the entries of the excitation matrix such that it maximizes the expected value of the joint log-likelihood as follows:

$$\mathbf{E} \quad = \quad \arg\max_{\mathbf{E}} (\langle \log \phi \rangle_{q(s)q(\mathbf{D})}) \tag{5.34}$$

Posterior distribution of the latent speech sources given the training data can be calculated as follows:

$$
\begin{aligned}
p(s_{fbt}|S_{ft},\Theta) &= N_c(s_{fbt};\mu_{fbt},\Sigma_{fbt}) && (5.35)\\
\lambda_{fbt} &= (\langle\frac{1}{D_{fb}}\rangle)^{-1}E_{bt} && (5.36)\\
\kappa_{fbt} &= \frac{\lambda_{fbt}}{\sum_b \lambda_{fbt}} && (5.37)\\
\Sigma_{fbt} &= \lambda_{fbt}(1-\kappa_{fbt}) && (5.38)\\
\mu_{fbt} &= \kappa_{fbt}S_{ft} && (5.39)
\end{aligned}
$$

where λ and κ are auxiliary variables to shorten the equations. Marginal expectation of the latent speech sources can be calculated using the posterior parameters as:

$$
\langle|s_{fbt}|^2\rangle = \Sigma_{fbt} + |\mu_{fbt}|^2
$$

The posterior distributions of the entries of the template matrix are Inverse-Gamma distribution due to the conjugacy between Inverse-Gamma and Gaussian distributions and can be calculated as follows:

$$
\begin{aligned}
q(D_{fb}) &\propto \exp((-a_{fb}-T-1)\log D_{fb} - \frac{1}{D_{fb}}(\frac{1}{b_{fb}} + \sum_t \langle|s_{uti}|^2\rangle\langle\frac{1}{E_{bt}}\rangle))\\
&\propto \mathcal{IG}(D_{fb};\alpha_{fb},\beta_{fb})
\end{aligned}
$$

with parameters:

$$
\begin{aligned}
\alpha_{fb} &= a_{fb}+T\\
\beta_{fb} &= (\frac{1}{b_{fb}} + \sum_t \langle|s_{uti}|^2\rangle\langle\frac{1}{E_{bt}}\rangle)^{-1}
\end{aligned}
$$

The sufficient statistics of the entries of the template matrix which are used to calculate the rest of the parameters in the model are:

$$\begin{aligned} \langle D_{fb} \rangle &= \frac{1}{\beta_{fb}(\alpha_{fb}-1)} \\ \langle \frac{1}{D_{fb}} \rangle &= \alpha_{fb}\beta_{fb} \\ \langle \log D_{fb} \rangle &= -\psi(\alpha_{fb}) - \log \beta_{fb} \end{aligned}$$

The entries of the excitation matrix which maximizes the expectation of the joint log-likelihood of the data is calculated using the following equation:

$$E_{bt} = \frac{1}{F} \sum_{f} \langle |s_{fbt}|^2 \rangle \langle \frac{1}{D_{fb}} \rangle.$$

By using the training data, hyper-parameters of the entries of the template matrix are calculated. Hence, in the separation phase, pre-computed hyper-parameters are used as a prior model for the speech signal.

5.2.3. Separation Method with Inverse-Gamma Priors

In this section, the separation method with Inverse-Gamma priors is described. A similar method described in Section 5.1.3 will be used. We use the variational inference technique that factorizes the posterior distribution into the posteriors of the decoupled random variables as follows:

$$\begin{aligned} q(s,m,r) &\propto \exp(\langle \log \phi \rangle_{q(\mathbf{D})}) \\ q(\mathbf{D}) &\propto \exp(\langle \log \phi \rangle_{q(s,m,r)}) \\ \mathbf{E}^* &\propto \arg\max_{\mathbf{E}} (\langle \log \phi \rangle_{q(s,m,r)q(\mathbf{D})}) \end{aligned}$$

where ϕ represents the joint likelihood of the data and the latent sources and Θ represents the parameters in the model such as $a_{fb}, b_{fb}, \pi, h, v$. The joint likelihood can be

decomposed as follows:

$$\phi = p(\mathbf{X}, s, m, \mathbf{D}, \mathbf{E}, r, f, v|\Theta) = p(\mathbf{X}|s, m)p(s|\mathbf{D}, \mathbf{E})p(\mathbf{D}|a_{fb}, b_{fb})p(m|r, h, v)p(r|\pi)$$

$$\begin{aligned}\log\phi &= \sum_{t,j}[r_t = j]\Big[\sum_{fb}\big(-\log D_{fb}E_{bt} - \frac{|s^j_{fbt}|^2}{D_{fb}E_{bt}} + (-a_{fb}-1)\log D_{fb} - \frac{b_{fb}}{D_{fb}} - \log\Gamma(a_{fb}) \\ &+ a_{fb}\log(b_{fb})\big) + (\sum_f(-\log C_{fj}h_f v_t - \frac{|m^j_{ft}|^2}{C_{fj}h_f v_t}) + \log\delta(X_{ft} - \sum_b s_{fbt} - m_{ft}) + \log\pi_j\Big]\end{aligned}$$

The joint posterior distribution of the latent speech and music sources and jingle frame indexes, $q(s, m, r)$, is a CGMM as shown in [7]. The overall joint posterior distribution of the latent speech and music sources can be decomposed conditioned on the jingle frame index, r, as

$$q(s, m, r) = q(s, m|r)q(r).$$

Conditional posterior distribution of the latent speech and music sources are complex Gaussian distributed as:

$$\begin{aligned}p(s_{fbt}|X, r_t = j) &= \mathcal{N}_c(s^j_{fbt}; \mu^j_{fbt}, \Sigma^j_{fbt}) \\ p(m_{ft}|X, r_t = j) &= \mathcal{N}_c(m^j_{ft}; \mu^j_{ft}, \Sigma^j_{ft}).\end{aligned}$$

The parameters of the complex Gaussian distributions of the latent speech and music sources can be computed using:

$$\begin{aligned}
p^j_{fbt} &= \frac{\langle\frac{1}{D_{fb}}\rangle^{-1}E_{bt}}{\sum_b(\langle\frac{1}{D_{fb}}\rangle^{-1}E_{bt}) + C_{fj}h_f v_t} \\
\Sigma^j_{fbt} &= p^j_{fbt}(\sum_{i\neq b}\langle\frac{1}{D_{fi}}\rangle^{-1}E_{it} + C_{fj}h_f v_t) \\
\mu^j_{fbt} &= p^j_{fbt}X_{ft} \\
p^j_{ft} &= \frac{C_{fj}h_f v_t}{\sum_b(\langle\frac{1}{D_{fb}}\rangle^{-1}E_{bt}) + C_{fj}h_f v_t} \\
\Sigma^j_{ft} &= p^j_{ft}(\sum_{b}\langle\frac{1}{D_{fb}}\rangle^{-1}E_{bt} + C_{fj}h_f v_t) \\
\mu^j_{ft} &= p^j_{ft}X_{ft}
\end{aligned}$$

where p^j_{fbt} and p^j_{ft} are the auxiliary variables used to shorten the equations. The conditional posterior distribution of the jingle indexes can be computed as follows:

$$q(r_t|X) = \frac{\prod_{ft}\mathcal{N}_c(X_{ft};0,C_{fj}h_f v_t + \sum_b\langle\frac{1}{D_{fb}}\rangle^{-1}E_{bt})\pi_j}{\sum_j\prod_{ft}\mathcal{N}_c(X_{ft};0,C_{fj}h_f v_t + \sum_b\langle\frac{1}{D_{fb}}\rangle^{-1}E_{bt})\pi_j}$$

The conditional marginal expectation of the latent sources can be calculated using the parameters as:

$$\begin{aligned}
\langle[r_t = j]\rangle &= q(r_t = j|x,\theta) \\
\langle|s^j_{fbt}|^2\rangle &= \Sigma^j_{fbt} + |\mu^j_{fbt}|^2 \\
\langle|m^j_{ft}|^2\rangle &= \Sigma^j_{ft} + |\mu^j_{ft}|^2
\end{aligned}$$

The posterior distribution of the each entry of the template matrix is Inverse-Gamma distribution due to the conjugacy property of the complex Gaussian and Inverse-

Gamma distributions with parameters:

$$\begin{aligned} q(D_{fb}) &\propto \mathcal{IG}(D_{fb}; \alpha_{fb}, \beta_{fb}) \\ \alpha_{fb} &= a_{fb} + T \\ \beta_{fb} &= (\frac{1}{b_{fb}} + \sum_{t,j} \frac{\langle [r_t = j] \rangle \langle |s^j_{uti}|^2 \rangle}{E_{bt}})^{-1} \end{aligned}$$

The sufficient statistics of the Inverse-Gamma distribution can be calculated using the following equations:

$$\langle \frac{1}{D_{fb}} \rangle = \alpha_{fb} \beta_{fb}.$$

The excitation matrix which maximizes the expectation of the joint log-likelihood of the data is calculated using the following equation:

$$E_{bt} = \frac{1}{F} \sum_{f,j} \langle [r_t = j] \rangle \langle |s^j_{fbt}|^2 \rangle \langle \frac{1}{D_{fb}} \rangle.$$

The update equations for gain and filtering parameters are the same as in Chapter 4.

5.3. Experimental Results

5.3.1. Speech Recognition System and Test Set

For speech recognition tests, we used a CMU-Sphinx HMM-based continuous density speech recognizer which is trained to recognize Turkish Broadcast News speech. The gender-dependent acoustic models are trained using MFCCs and their deltas and double-deltas calculated in 25ms frames. The test set contains 240 utterances distributed approximately uniformly across 8 speakers. The total length of the test set is about 70 minutes and the average length of the utterances is about 18 seconds. The test system is summarized in Figure 5.3.

The test utterances are mixed with 10 different jingles at $0, 5, 10, 15$ and 20 dB

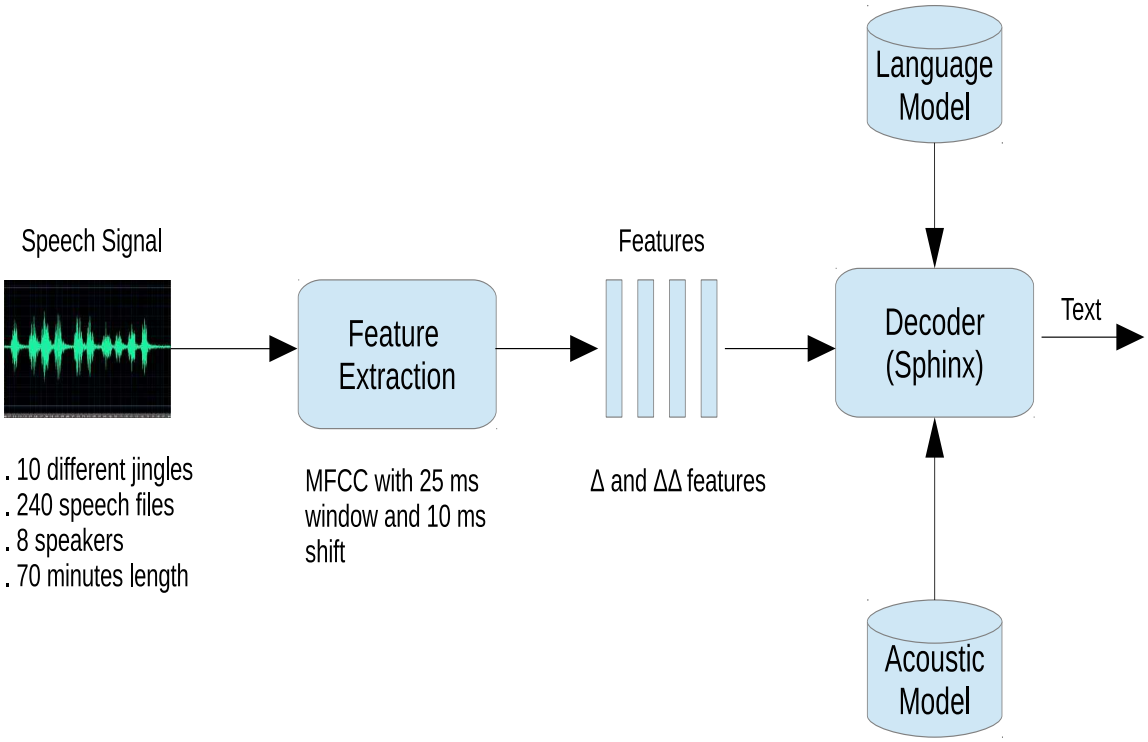

Figure 5.3. ASR system and test set for mixture based separation with speech models.

SMR levels to create the test set. The average length of the jingles is 7 seconds. The background music signal is generated by repeating the jingles up to the length of the speech. The jingles are taken from real broadcast news jingles. In this study, we assume, which jingle is used to generate the background music is known as a prior. While WAcc of the clean speech data is 73.9%, WAcc values of the mixed data without any separation method is shown in Table 5.1.

Table 5.1. Baseline WAcc values.

Baseline	Input SMR Values				
Results	0dB	5dB	10dB	15dB	20dB
Clean	73.9	73.9	73.9	73.9	73.9
Mixed	1.1	4.8	16.8	36.9	54.8

The magnitude or power spectrum are computed using 1024-point length frames and 512 point frame shift is used. In order to train the speech model, four types of speech data set are used and the properties of these sets are listed in Table 5.2 and

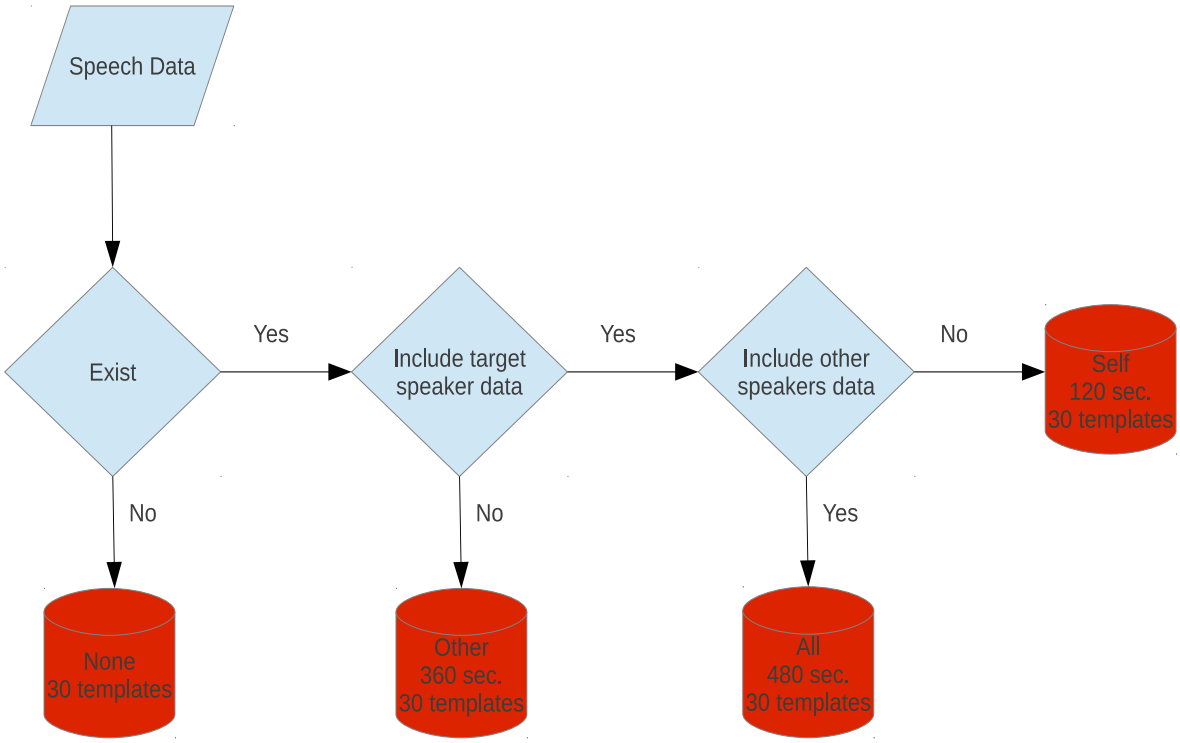

Figure 5.4. Training data types for speech signal.

shown in Figure 5.4. In our approach, the prior speech model contains the hyper-parameters of gamma distributions (Gamma distribution in Poisson model or Inverse-Gamma distribution in complex Gaussian model). It is assumed that each frequency bin of the template vector of the speech signal has a different gamma distribution.

Table 5.2. Speech training data set properties.

Data Set	# of Speakers	Definition of the set	Length (min.)	# of Bases
Self	1	The same speaker	2	30
All	4	Including Speaker	8	30
Other	3	Excluding Speaker	6	30
None	0	No speech data	0	30

5.3.2. Evaluation Plan

In our experimental study, the effects of three major factors on the mixture-based separation method are tested. These factors can be listed as follows.

- Divergence Measure (KL or IS): The aim is to compare the effect of divergence measures on the separation performance.
- Prior Speech Data Type (None (N), Self (S)), All (A) and Other (O): The aim is to analyze the effect of using different types of speech training data on the separation performance. None type refers to the mixture-based method without prior speech model which is described in Chapter 4.
- Gain Estimation Strategy (The original method (O) and Gamma Chains (G))

As a complete example of naming, 'IS-G-A' represents separation with IS divergence with IGMC is used as the gain estimation method and 'All' type speech training data is used to get the prior speech model.

5.3.3. Experimental Analysis

KL SMR Value Analysis (Table 5.3):

(i) As compared to 'None' model, other models ('All','Self' and 'Other') gets higher SMR values.

(ii) As compared to 'KL-O-N', using GMC (KL-G-N) makes slightly better improvement on SMR values than using prior speech models (KL-O-A, KL-O-O and KL-O-S).

(iii) Although with 'Original' gain estimation strategy, 'All' and 'Self' models (KL-O-A and KL-O-S) gives higher SMR values as compared to 'Other' (KL-O-O), with GMC gain estimation strategy they (KL-G-A and KL-G-S) gives similar output SMR values to 'Other' method (KL-G-O).

KL SAR Value Analysis (Table 5.4):

(i) As compared to 'None' model, other models ('All', 'Self' and 'Other') gets higher SAR values with 'Original' gain estimation strategy.

(ii) With GMC on gain values, all speech models yields similar SAR values.

(iii) As compared to 'KL-O-N', using GMC (KL-G-N) makes better improvement on

Table 5.3. Output SMR values of KL mixture based methods with different prior speech data types and gain estimation strategies.

Separation Method	Input SMR Values				
	0dB	5dB	10dB	15dB	20dB
KL-O-N	8.1	17.0	25.9	35.0	44.2
KL-G-N	20.3	26.9	33.5	40.5	47.8
KL-O-O	12.6	20.6	28.5	36.5	44.8
KL-G-O	19.5	26.5	33.4	40.4	47.9
KL-O-A	13.3	21.2	28.9	36.8	44.9
KL-G-A	19.9	26.7	33.6	40.6	48.0
KL-O-S	13.2	21.2	29.5	37.9	45.4
KL-G-S	16.3	24.1	33.2	40.8	48.0

SAR values than using prior speech models (KL-O-A,KL-O-O and KL-O-S).

Table 5.4. Output SAR values of KL mixture based methods with different prior speech data types and gain estimation strategies.

Separation Method	Input SMR Values				
	0dB	5dB	10dB	15dB	20dB
KL-O-N	8.9	12.4	15.6	18.9	21.8
KL-G-N	12.2	14.9	17.6	20.5	23.4
KL-O-O	9.9	13.3	16.3	19.4	22.3
KL-G-O	12.1	14.8	17.5	20.3	23.2
KL-O-A	10.3	13.6	16.6	19.6	22.5
KL-G-A	12.3	15.0	17.7	20.5	23.4
KL-O-S	10.2	13.5	16.1	17.8	20.9
KL-G-S	12.2	14.9	17.6	19.1	22.0

KL WAcc Value Analysis (Table 5.5 and Figures 5.5, 5.6, 5.7):

(i) As compared to 'None' model, other models ('All', 'Self' and 'Other') gets higher WAcc values with 'Original' and 'GMC' gain estimation strategies.

(ii) With GMC on gain values, all speech models yields similar WAcc values.

(iii) As compared to 'KL-O-N', using GMC (KL-G-N) makes better improvement on WAcc values than using prior speech models (KL-O-A,KL-O-O and KL-O-S).

Table 5.5. WAcc values of KL mixture based methods with different prior speech data types and gain estimation strategies.

Separation Method	Input SMR Values				
	0dB	5dB	10dB	15dB	20dB
KL-O-N	6.3	17.6	36.6	54.0	63.9
KL-G-N	29.4	44.0	57.3	64.5	67.9
KL-O-O	16.9	32.8	49.7	61.2	66.6
KL-G-O	40.3	55.3	62.9	67.9	69.4
KL-O-A	18.4	34.3	50.4	61.2	66.2
KL-G-A	41.0	55.4	63.2	67.8	69.5
KL-O-S	17.1	33.4	49.8	61.2	65.7
KL-G-S	36.4	51.3	60.7	66.8	67.8

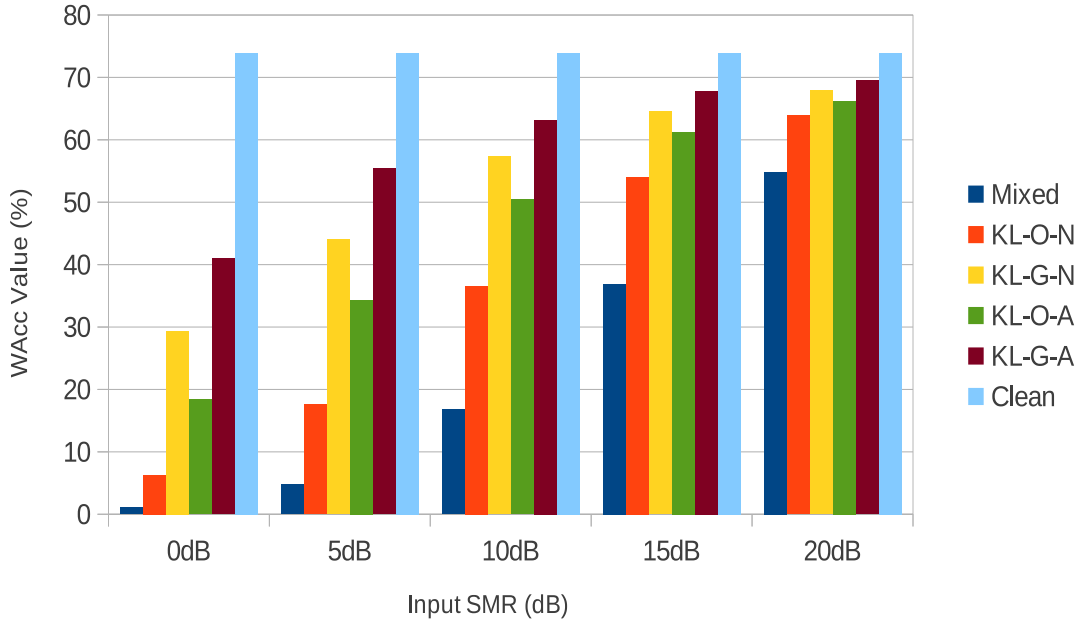

Figure 5.5. ASR result with 'None' speech model.

IS SMR Value Analysis (Table 5.6):

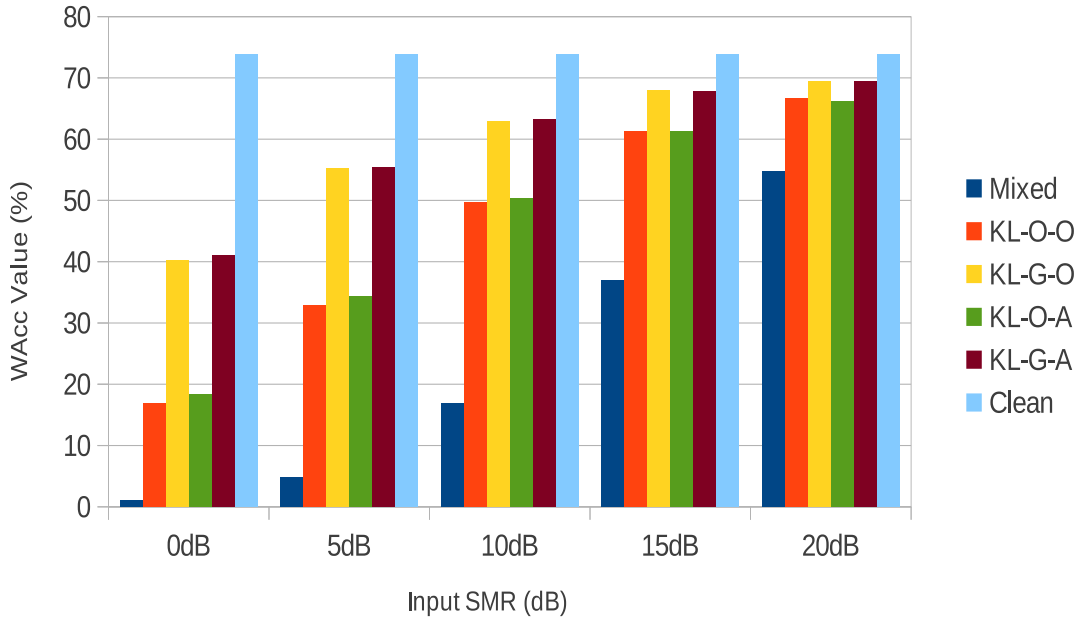

Figure 5.6. ASR result with 'All' and 'Other' speech models.

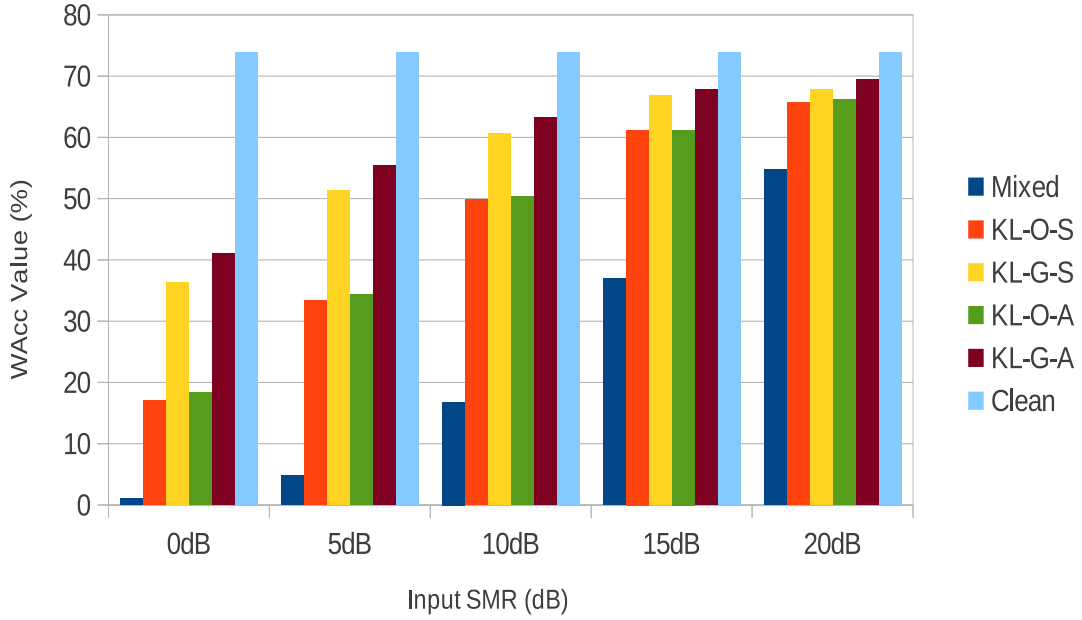

Figure 5.7. ASR result with 'All' and 'Self' speech models.

(i) As compared to 'None' model, other models ('All','Self' and 'Other') gets higher SMR values.

(ii) As compared to 'IS-O-N', using GMC (IS-G-N) makes slightly better improvement on SMR values than using prior speech models (IS-O-A,IS-O-O and IS-O-S).

(iii) Although with 'Original' gain estimation strategy, 'All' and 'Self' models (IS-O-A and IS-O-S) gives higher SMR values as compared to 'Other' (IS-O-O), with GMC gain estimation strategy they (IS-G-A and IS-G-S) gives similar output SMR values to 'Other' method (IS-G-O).

Table 5.6. Output SMR values of IS mixture based methods with different prior speech data types and gain estimation strategies.

Separation Method	Input SMR Values				
	0dB	5dB	10dB	15dB	20dB
IS-O-N	12.3	20.9	29.6	37.8	46.0
IS-G-N	16.4	23.9	31.5	39.2	47.0
IS-O-O	15.4	23.6	31.2	38.9	46.9
IS-G-O	17.1	24.5	31.9	39.6	47.4
IS-O-A	16.1	23.7	31.3	39.1	47.0
IS-G-A	17.1	24.5	32.0	39.7	47.5
IS-O-S	16.2	23.8	31.5	39.2	47.0
IS-G-S	17.2	24.7	32.2	39.8	47.4

IS SAR Value Analysis (Table 5.7):

(i) As compared to 'None' model, other models ('All', 'Self' and 'Other') gets higher SAR values.

(ii) As compared to 'IS-O-N', using GMC (IS-G-N) makes slightly better improvement on SAR values than using prior speech models (IS-O-A,IS-O-O and IS-O-S).

(iii) Although with 'Original' gain estimation strategy, 'All' and 'Self' models (IS-O-A and IS-O-S) gives higher SAR values as compared to 'Other' (IS-O-O), with GMC gain estimation strategy they (IS-G-A and IS-G-S) gives similar output SAR values to 'Other' method (IS-G-O).

Table 5.7. Output SAR values of IS mixture based methods with different prior speech data types and gain estimation strategies.

Separation Method	Input SMR Values				
	0dB	5dB	10dB	15dB	20dB
IS-O-N	9.9	13.8	17.5	21.0	24.3
IS-G-N	12.0	15.3	18.5	21.6	24.9
IS-O-O	11.5	15.0	18.2	21.4	24.7
IS-G-O	12.3	15.5	18.6	21.7	25.0
IS-O-A	11.7	15.0	18.2	21.4	24.7
IS-G-A	12.3	15.5	18.6	21.7	25.0
IS-O-S	11.7	15.1	18.3	21.5	24.8
IS-G-S	12.1	15.3	18.6	21.8	25.0

IS WAcc Value Analysis (Table 5.8):

(i) As compared to 'None' model, other models ('All','Self' and 'Other') gets higher WAcc values.

(ii) As compared to 'IS-O-N', using GMC (IS-G-N) makes worse improvement on WAcc values than using prior speech models (IS-O-A,IS-O-O and IS-O-S).

Table 5.8. WAcc values of IS mixture based methods with different prior speech data types and gain estimation strategies.

Separation Method	Input SMR Values				
	0dB	5dB	10dB	15dB	20dB
IS-O-N	24	43.2	60.2	66.7	70.1
IS-G-N	38.3	54.9	64.5	69.5	70.5
IS-O-O	48.0	59.8	66.6	69.9	71.6
IS-G-O	52.8	63.1	68.3	70.7	72.1
IS-O-A	47.1	60.1	66.7	69.2	71.6
IS-G-A	52.1	63.0	66.4	70.5	72.1
IS-O-S	48.1	59.2	66.1	69.1	71.3
IS-G-S	51.9	62.4	67.7	69.8	71.5

Overall Experimental Analysis:

When we analyze the SMR results in Tables 5.3 and 5.6, it is shown that, for both of divergence measures (KL or IS), prior speech model (All, Self and Other) increases the SMR values with original gain estimation strategy. However, in GMC case, the SMR values are almost the same as with no speech prior model (None). For SAR values in Tables 5.4 and 5.7, it is observed that for both of observation models (KL or IS) and Gain estimation strategies (Original or GMC), incorporating prior speech information increases the SAR values. The speech recognition performance of a separation method is affected by both of the SMR and the SAR values. Therefore, using prior speech model in the separation for all conditions (KL or IS, Original or GMC) improves the speech recognition performance. This fact can be seen in Tables 5.5 and 5.8 and Figure 5.8.

When we compare the effects of prior speech models and the gain estimation strategies for divergence measures, it is really interesting that using prior speech models in IS case improves the separation performance more than using GMC in gain estimation. However, for KL case, using GMC in gain estimation makes more improvement than using prior speech models in the separation. The reason why using the prior speech model or GMC in gain estimation strategy makes more relative improvement in KL case than IS case is that the baseline separation performance of IS (IS-O-N) is better than KL (KL-O-N) case. As a result, incorporating prior speech information in the separation method enhances the separation performance of both of divergence measures with both of gain estimation strategies (Original or GMC).

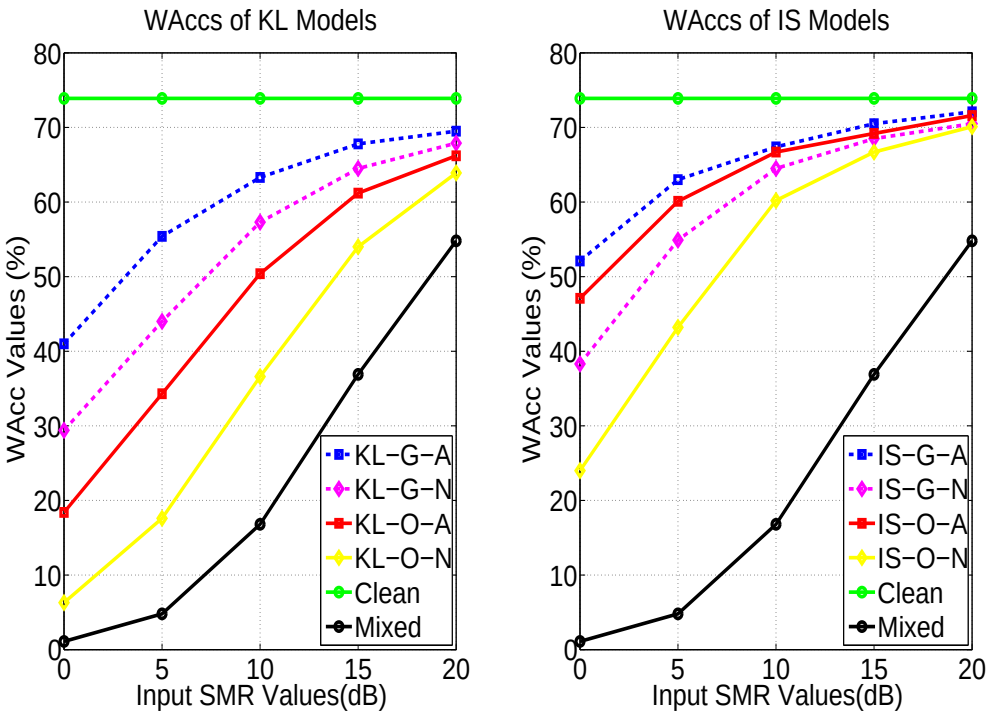

Figure 5.8. Comparison of ASR performances with prior speech models.

6. SUB-WORD SPECIFIC SPEECH MODELS FOR SPEECH-MUSIC SEPARATION

In Chapter 5, the single-channel speech-music separation algorithm, which uses a mixture model for the music signal and an NMF model for the speech signal, is developed. In this chapter, an NMF-based single-channel source separation method is described. The sources, speech and music signals, are modeled using the NMF method. In NMF modeling approach, the source model refers to the template matrix of the resultant NMF of the training data matrix. The NMF model for the music signal is trained using each jingle data by itself. The focus of this study will be on the analysis of the modeling of the speech signal. Although in the previous chapters, a general NMF model is used for representing the speech source signal, more detailed model type information can be obtained using the output of an ASR system. For example, for each time frame, word, phone or state identity information can be obtained using the ASR output. Since the speech signal has a non-stationary spectral structure, it is more advantageous to represent the speech signal using more specific models for each time frame. However, there are two issues related with the use of sub-word speech models for the source separation task. They are:

(i) *Training of the word or sub-word speech models:* For representing the speech signals using a sub-word model, it is necessary to provide a training corpus which is labeled according to sub-word units. Fortunately, forced-alignment methods in ASR technology can provide sub-word labeled data using trained acoustic models. In other words, we train an acoustic model using the transcribed speech data and then we force training data frames to align with preferred sub-word units. As a result of the forced-alignment procedure, for each time frame of the clean speech data, a sub-word model unit identity is assigned. By collecting all frames assigned to each sub-word model unit, an NMF model for each unit can be trained.

(ii) *Using word or sub-word speech models in the separation:* In order to use a specific sub-word model in the speech-music separation process, the identity of the sub-

word model for each time frame of the mixed signal has to be known. Since the content of the mixed signal is unknown, it is necessary to use the ASR system to transcribe the mixed signal. With the forced-alignment of the ASR system output with the mixed signal, sub-word unit identity is estimated for each time-frame. Hence, the separation process can be carried out using sub-word models for each time frame.

6.1. Phone Model Training

As an example of sub-word model training, phone-model training procedure is described in detail in this section. In order to train a different model for each phone, the alignment of the clean speech data to each phone is necessary. The phone alignment of the speech data can be obtained using a previously trained acoustic model. The alignment of the acoustic model and the phone models is obtained using Viterbi alignment strategy (A forced-alignment method). In this strategy, each frame is only assigned to a phone model. The phone model training procedure is shown in Figure 6.1. Since the acoustic model is trained using the clean speech data and is also used in

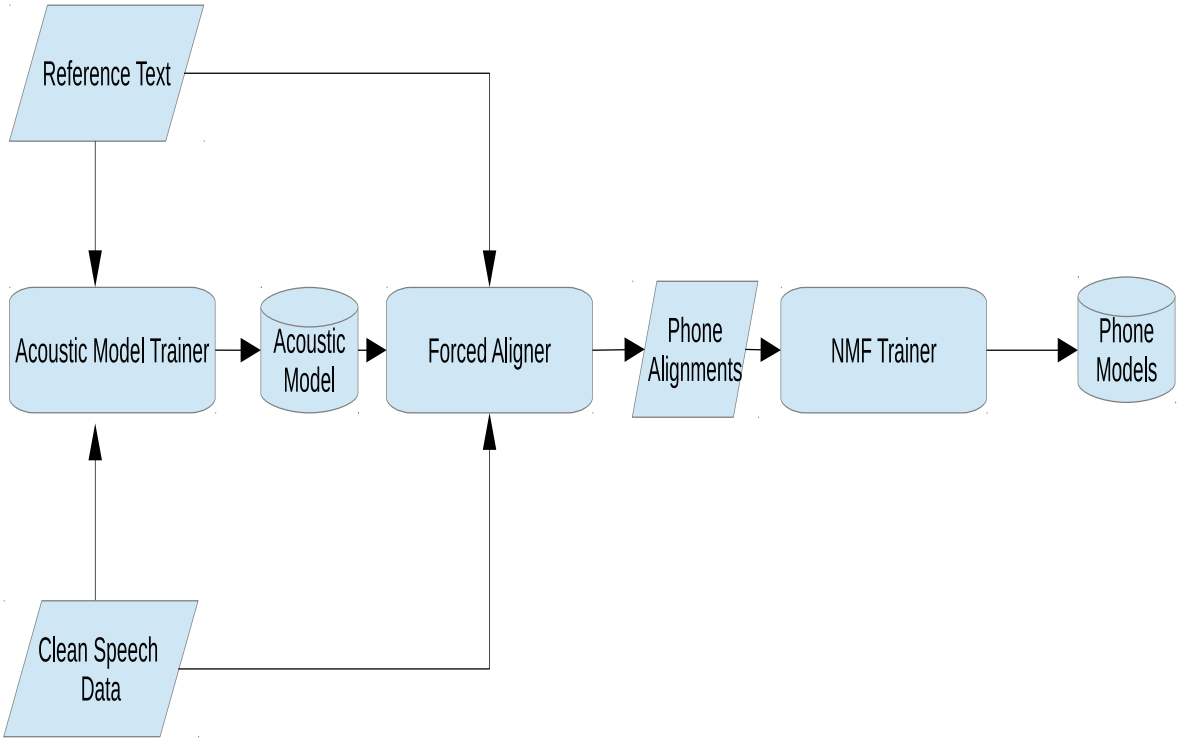

Figure 6.1. Phone model training procedure.

the speech recognition process, in order to get the phone alignments, no extra model or data is required.

It should be noted that the phone alignments obtained using the acoustic model and the clean speech data do not have to be correct phone alignments because the alignment process is a statistical method and it is prone to errors. However, since manual phone-based alignment process of the speech signal is very costly, in speech processing applications, automatic alignment systems are preferred to obtain the phone alignment for each time frame.

6.2. Separation without a Speech Model

As a baseline system, the separation task is performed using only a music model similar to Chapter 4. In this case, no speech model is used in the separation phase. In other words, the template and excitation matrices of the speech signal are estimated from the mixed signal simultaneously. The difference from the method in Chapter 4 is that the music signal is represented using an NMF model in this part. The reason for choosing an NMF model for the music signal is to focus on the sub-word modeling approach for the speech signal rather than the music representation techniques.

As in the previous chapters, it is assumed that the identity of the jingle is known as a prior. For each jingle, an NMF model is trained using the frames of the jingle itself. For each mixed signal, the NMF model which generates the background music is used as a music model. The separation task is carried out using NMF update equations via fixing the template matrix of the music signal. The excitation matrix of the music signal is estimated from the mixture. For the speech signal, the template and the excitation matrices are estimated from the mixed signal simultaneously. This process is shown in Figure 6.2. This approach is called as 'No Speech Model'. In this strategy, the speech model is trained during the separation process and it is a specific model for each test signal.

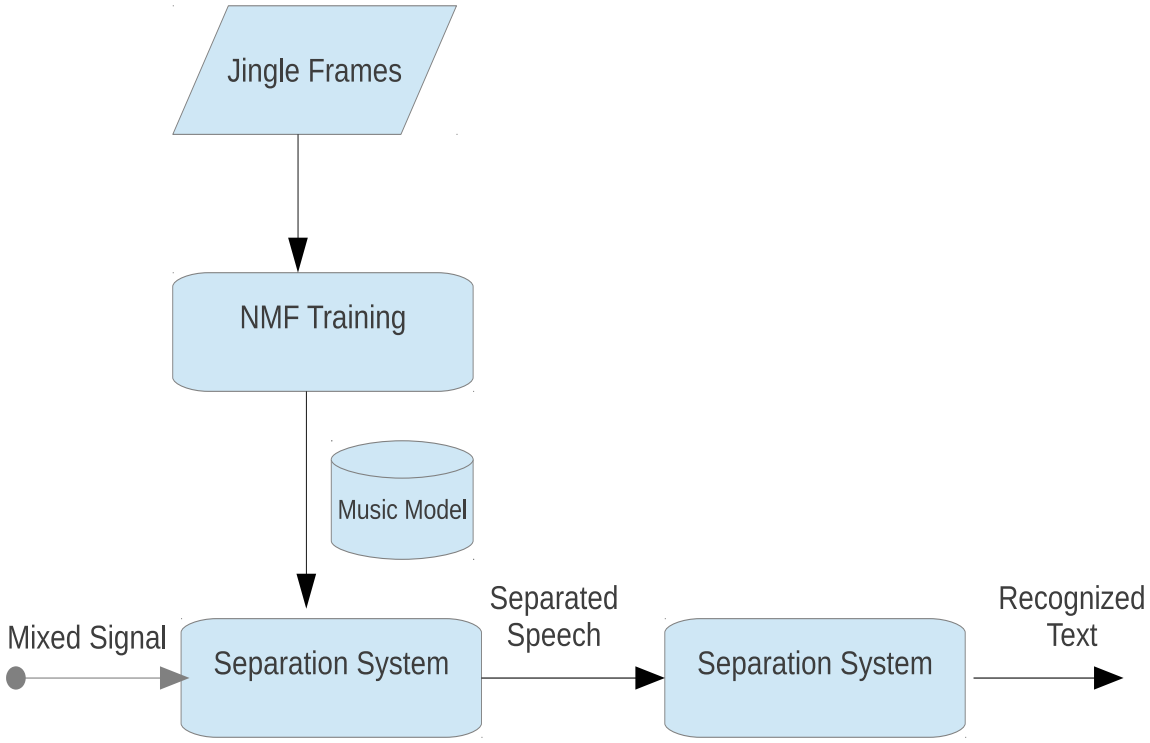

Figure 6.2. Separation without speech model (No Speech).

6.3. Separation with a General Speech Model

As another baseline system, the separation can be performed using a jingle-specific music model and a general speech model which is trained using all the clean speech data frames. Different from the Chapter 5, in this case, a fixed general speech model is used during the separation phase. In other words, the template for the speech signal is estimated from clean speech data and these fixed templates are used in the separation process as a speech model. The difference from the Chapter 4 is that the music model is represented using an NMF model in this part. The music model in this scenario is the same as in Section 6.2.

In the separation phase, the templates of the music and speech signals are fixed for the simplicity. The separation is also performed using NMF update equations. The excitations for the music and speech signals are estimated from the mixed signal. The process is shown in Figure 6.3. This approach is called as 'Speech'. In this strategy, the speech model is trained using the clean speech database and it is a general model

for all test signals.

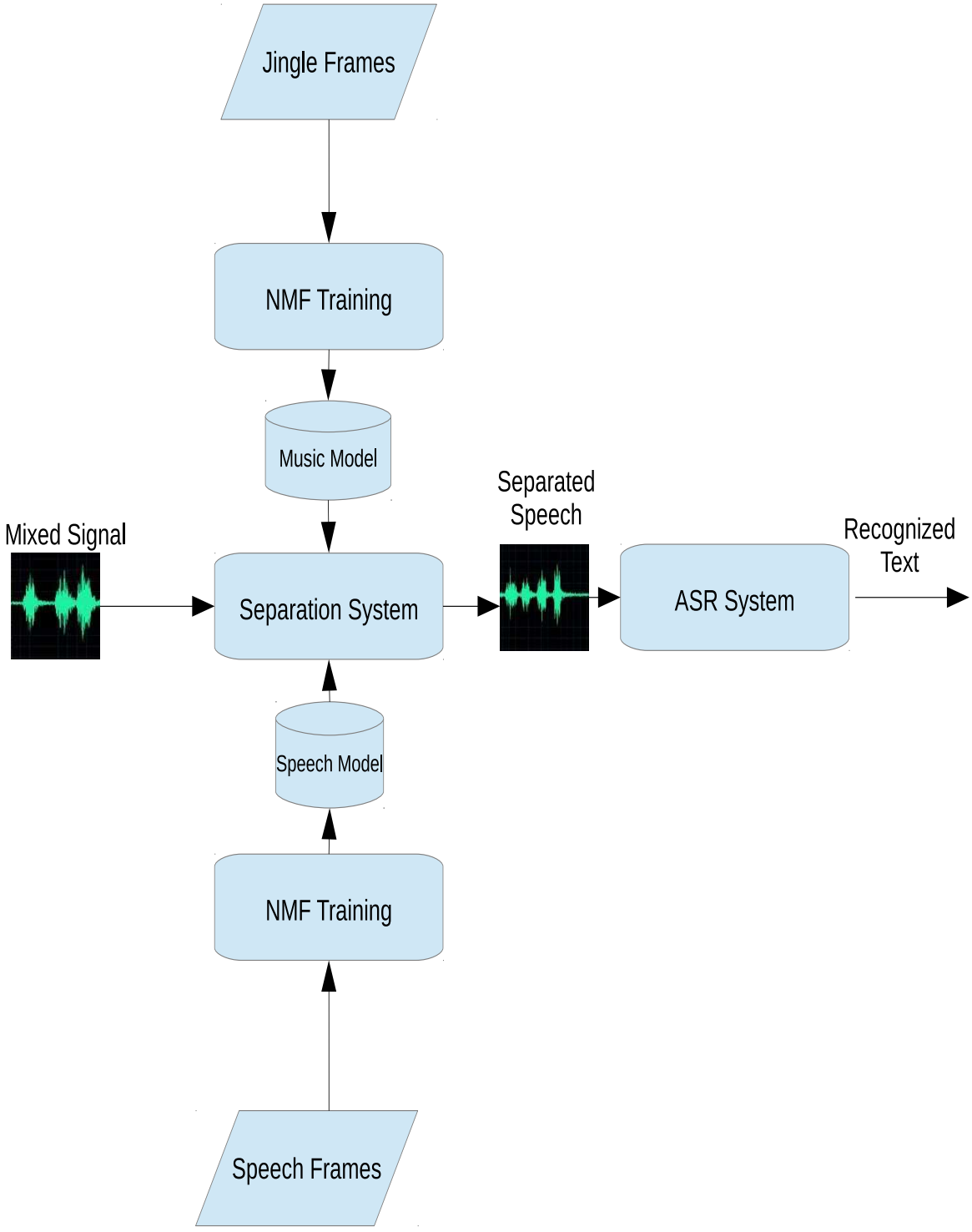

Figure 6.3. Separation with a general speech model (General Speech).

6.4. Separation with Known References

In this case, it is assumed that there are some clean speech data available for training the phone models as in the previous section. The clean speech data is aligned with the phones and an NMF model for each phone is trained using this aligned data. The training procedure is shown in Figure 6.1. In order to show the best performance, which can be achieved using Phone/State models in the speech-music separation, it is assumed that the content of the utterance is given as a prior (The reference text is in our hand or there is a speech recognizer with 0% WER.).

The phone alignments for the test data are obtained using the clean test speech and their references. During the separation phase, the model of the aligned phone is used as the speech model. Although it is assumed that aligning the clean speech data with the acoustic model gives the true phone alignments, it should be considered that this assumption is not necessarily correct. In other words, it is not guaranteed that the true phone alignments are the ones which are found using the acoustic model and the clean speech data. However, since there is no way to get better alignments for the time frames, the result of the alignment process with the acoustic model is considered as the best ones in this situation. Phone/State alignments are obtained as shown in Figure 6.4. This approach is called as 'Phone/State-Oracle'. The approach is called as 'Oracle' because of the fact that, for the speech signal point of view, it is impossible to obtain more information about the source signal.

6.5. Separation with Recognized Clean Speech

In this case, it is assumed that there are some clean speech data without transcriptions (reference texts) to train the phone models. The clean speech data is automatically transcribed using the speech recognition system and the recognized texts are used as reference to obtain the phone alignments. With clean speech case, it is assumed that the unmixed (clean) version of the test utterance is given but the content of the utterance is not known. This is a hypothetical case because, when clean speech of the mixed signal is available, there is no need to perform the separation process. However,

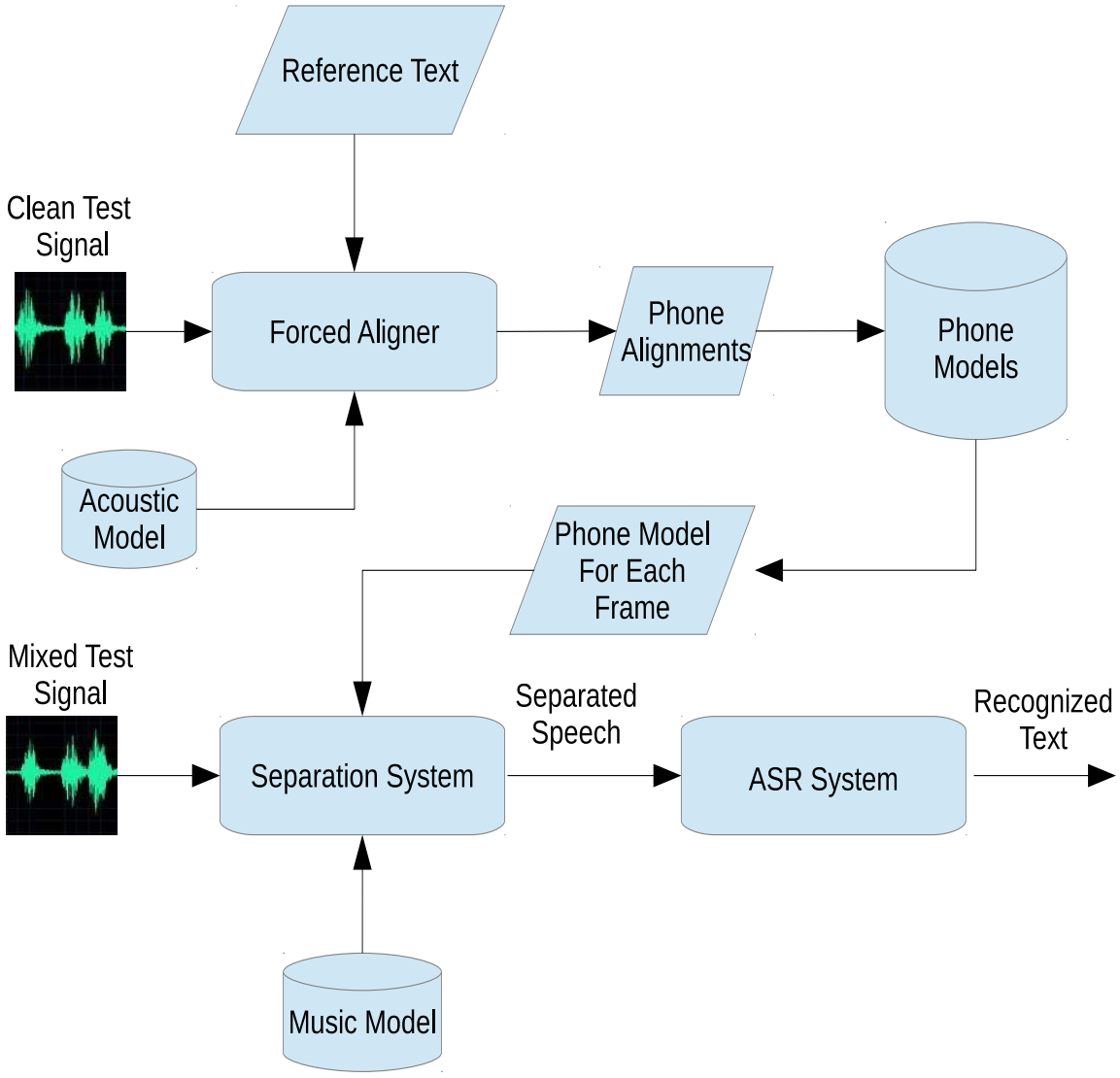

Figure 6.4. Separation with known references (Phone/State Oracle).

it is useful to show the effect of using erroneous transcriptions in phone alignment process.

The speech recognition system is used for getting the content of the utterance. The phone alignments for the test data are obtained using the clean test samples and their recognition outputs. In the separation phase, the aligned phone model is used for each frame as a speech model. The process of obtaining phone alignments is shown in Figure 6.5. This approach is called as 'Phone-Clean'.

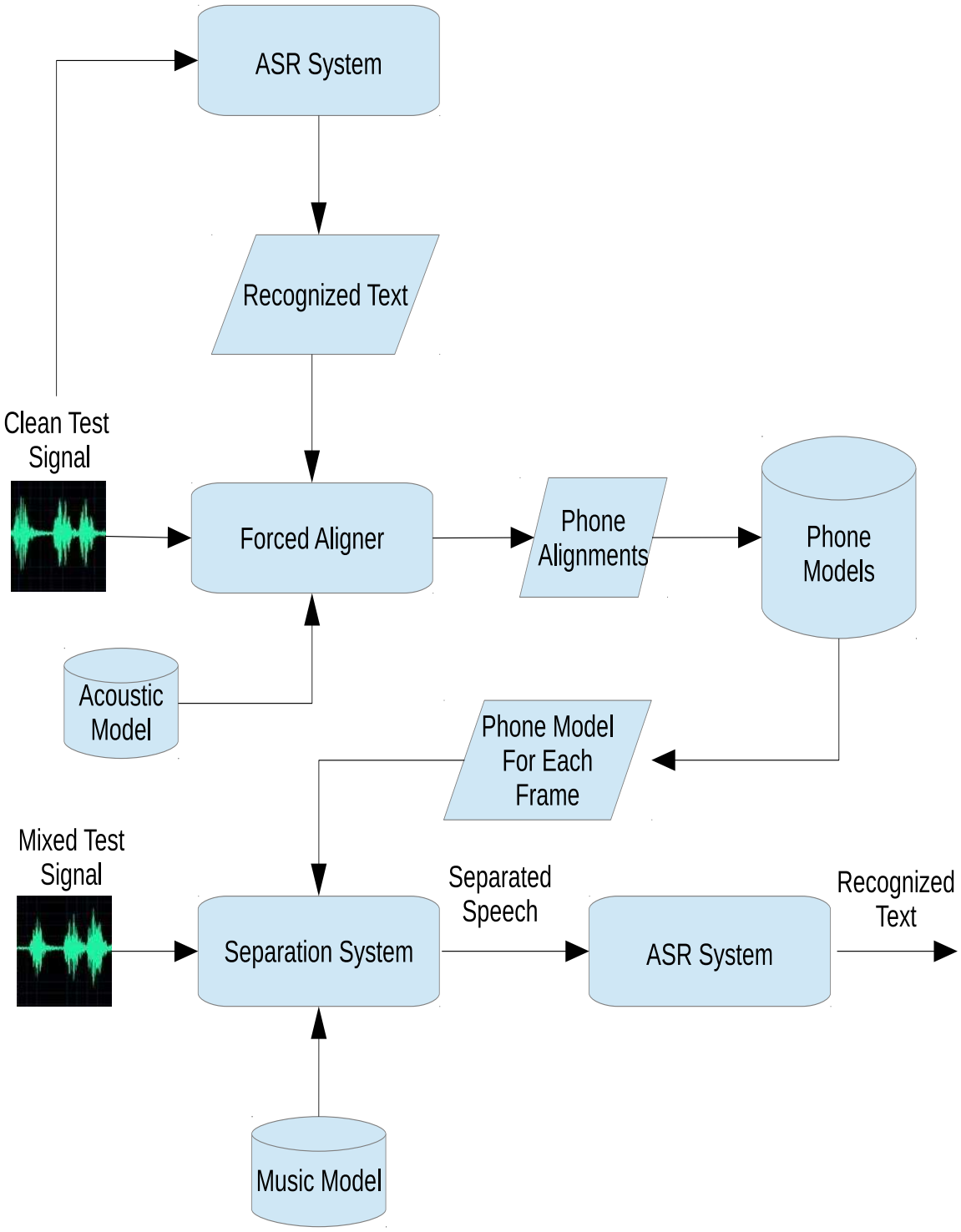

Figure 6.5. Separation with recognized clean speech (Phone/State Clean).

6.6. Multi-Pass Separation

In a real speech-music separation system, the phone alignment of the target signal has to be achieved using a speech recognition system because the content of the

target signal is unknown. Since the alignment accuracy is important for a phone-based separation system and the accuracy depends on the recognition accuracy, it is necessary to use a separated speech signal as an input to the speech recognition system. The 'None' and 'Speech' methods described in Sections 6.2 and 6.3 respectively can be used as the first separation system. In this study, 'Speech' method is preferred to the 'None' method due to the higher recognition accuracy and hence better phone alignment performance. As an output of the first pass, the recognized text is used for obtaining the phone alignment for each time frame.

Phone/State identity for each time frame is found via aligning the separated speech with the acoustic model. After obtaining the phone/state identity for each time frame, in the second pass separation, a specific phone/state model is used in the separation process with a jingle specific music model. In this approach, the separation of the speech signal is carried out using a two-pass process. Two-pass separation process is shown in Figure 6.6. Since 2nd pass of the method can be applied multiple times, the proposed scheme is called 'Multi-pass' separation method.

6.7. Experimental Results

6.7.1. Speech Recognition System and Test Set

For speech recognition tests, we used a CMU-Sphinx HMM-based continuous density speech recognizer which is trained to recognize Turkish Broadcast News speech. The gender-independent acoustic models are trained using MFCCs and their deltas and double-deltas calculated in 25ms frames. The test set contains 1200 utterances distributed approximately uniformly across 8 speakers. The total length of the test set is about 120 minutes and the average length of the utterances is about 6 seconds.

The test utterances are mixed synthetically with 10 different jingles at 10dB SMR level to create the test set. The average length of the jingles is 7 seconds. The background music signal is generated by repeating the jingles up to the length of the speech. The jingles are taken from real broadcast news jingles. In this study, we

assume, the which jingle is used to generate the background music is known as a prior. Since in this chapter, KL divergence based NMF method is used for modeling the source signals, the magnitude spectrograms of the source signals are used. The spectrograms are computed using 1024-point length frames and 512 point frame shift is used.

The NMF model for the speech signal is trained using the data which is used for acoustic modeling. The training data contains about 110 hours of transcribed utterances and does not contain the speech data of the target speakers. The number of template vectors for the speech signal is fixed at 120 for all experiments. The phone or state alignments are obtained using the ASR system described in Section 6.7.1. In sub-word experiments, 30 different phones are modeled using 30 template vectors. In state-based separation experiments, 3000 tied states are modeled using 30 template vectors.

In all scenarios, the NMF model of the music signal, which is trained using the jingle frames, is used for the separation. The number of template vectors for the music signal is fixed at 30 for all the experiments. For the speech signal part of the mixture, 4 different modeling approaches are used.

6.7.2. Experimental Analysis

In this section, ASR results obtained from the proposed separation methods are summarized. As a reference, ASR results with clean and mixed signals are also presented in Table 6.1. ASR results are measured using WER and Phone Error Rate (PER). In Table 6.1, 'No Speech' refers to the ASR result with the separated speech in which the separation is performed without a speech model (The method is described in Section 6.2). In Table 6.1, 'General Speech' refers to the ASR result with the separated speech in which the separation is performed with a general speech model (The method is described in Section 6.3).

'Phone-Oracle' case, the reference text of the test utterances are used to obtain the phone identity for each time frame. Although the known reference case is called

Table 6.1. Baseline ASR results.

Speech Model	WER	PER
Clean Speech	23.6	6.3
Mixed Speech	74.0	51.7
No Speech	55.6	29.6
General Speech	49.1	27.4

as 'Oracle', it is not guaranteed that the actual and the aligned phone identities are the same. However, since it is not possible to obtain a better alignment for the test utterances, this situation is the called as the oracle one for our purposes. 'Phone-Oracle' methodology is described in Section 6.4. 'State-Oracle' case is the same as 'Phone-Oracle' except that state models are used instead of phone models.

In 'Phone-Clean' case, which is described in Section 6.5, the recognized text of the unmixed (clean) test utterances are used to obtain the phone identity for each frame. This case represents another oracle situation such that the unmixed (clean) versions of the test utterances are available but the reference texts (transcriptions) must be obtained automatically from the speech recognition system. The expected separation performance of 'Phone-Clean' models is in between the 'Phone-Oracle' and 'Phone' models.

In 'Phone-Multi-pass' case, phone-based speech models are used in the second separation pass of multi-pass separation system which is described in Section 6.6. In the first pass, a general speech model is used in the separation. The phone-based speech models, which are obtained using alignment of the separated speech of the first pass of multi-pass system and the recognized text of the separated speech, are used in the second pass of the multi-pass system. In the second-pass of the method, for each frame of the mixed signal, a different phone-based speech model is used.

In 'Phone-Multi-pass-Nbest' case, instead of using the best hypothesized transcription for obtaining the phone-identity for each frame, 10 best hypothesized transcriptions are used to provide the all possible phone-identities which are used as the

speech model in the separation. In other words, for each time frame, it is possible to use more than one phone model in the separation process. However, in this approach, all hypothesized transcriptions are used by ignoring using their order in N-best list. In other words, the order or the scores in N-best list are not used in the separation process.

Table 6.2. ASR results with different strategies.

Speech Model	WER	PER
Phone-Oracle	32.1	13.8
Phone-Clean	32.9	14.2
Phone-Multi-pass	46.3	25.2
Phone-Multi-pass-Nbest	46.1	24.9
State-Oracle	22.1	6.4
State-Multi-pass	47.2	25.9

When the ASR results in Tables 6.1 and 6.2 are analyzed, it can be concluded that using 'Phone' models can improve the separation performance if the phone-identities are estimated using the recognized text of the separated speech with 'General Speech' model. The relative improvement in WER with 'Phone-Multi-pass' model as compared to 'General Speech' model is 5.7%. As compared to the 'Phone-Oracle' situation, the relative improvement on ASR performance can be considered as a small amount. However, when the phone accuracy of 'General Speech' model is considered (Which is 72.6%), this small improvement is expected. When the phone accuracy is 100% which corresponds to the 'Phone-Oracle', the WER is 32.1% and in Mixed case the WER is 74.0%. Therefore, with 72.6% phone accuracy performance, the separation method can improve the ASR performance up to the about 43.6%. When the effect of incorrect phone identities to the separation performance is considered, the 46.3% WER performance with 'Phone-Multi-pass' model is not surprising.

Instead of using the best hypothesis to obtain the phone alignments, 10 best hypothesis are used in 'Phone-Multi-pass-Nbest' case. However, the improvement using the 10-best hypothesis is almost negligible. In order to make a deep analysis, the phone

error rate of n-best approach must be calculated. Moreover, using all members of the n-best list equally is not fair because of the fact that best hypothesis is more likely and in the separation process its phone alignment result must have more weight as compared to the other hypothesis. Actually n-best hypothesis have to be used with the confidence scores in the separation process.

The surprising result is obtained using the 'State-Multi-pass' model due to the fact that though 'State-Oracle' speech recognition result is very close to the clean speech data result, the improvement with 'State-Multi-pass' model is not as much as with 'Phone-Multi-pass' model. The reason for this circumstance could be understood with more deep analysis.

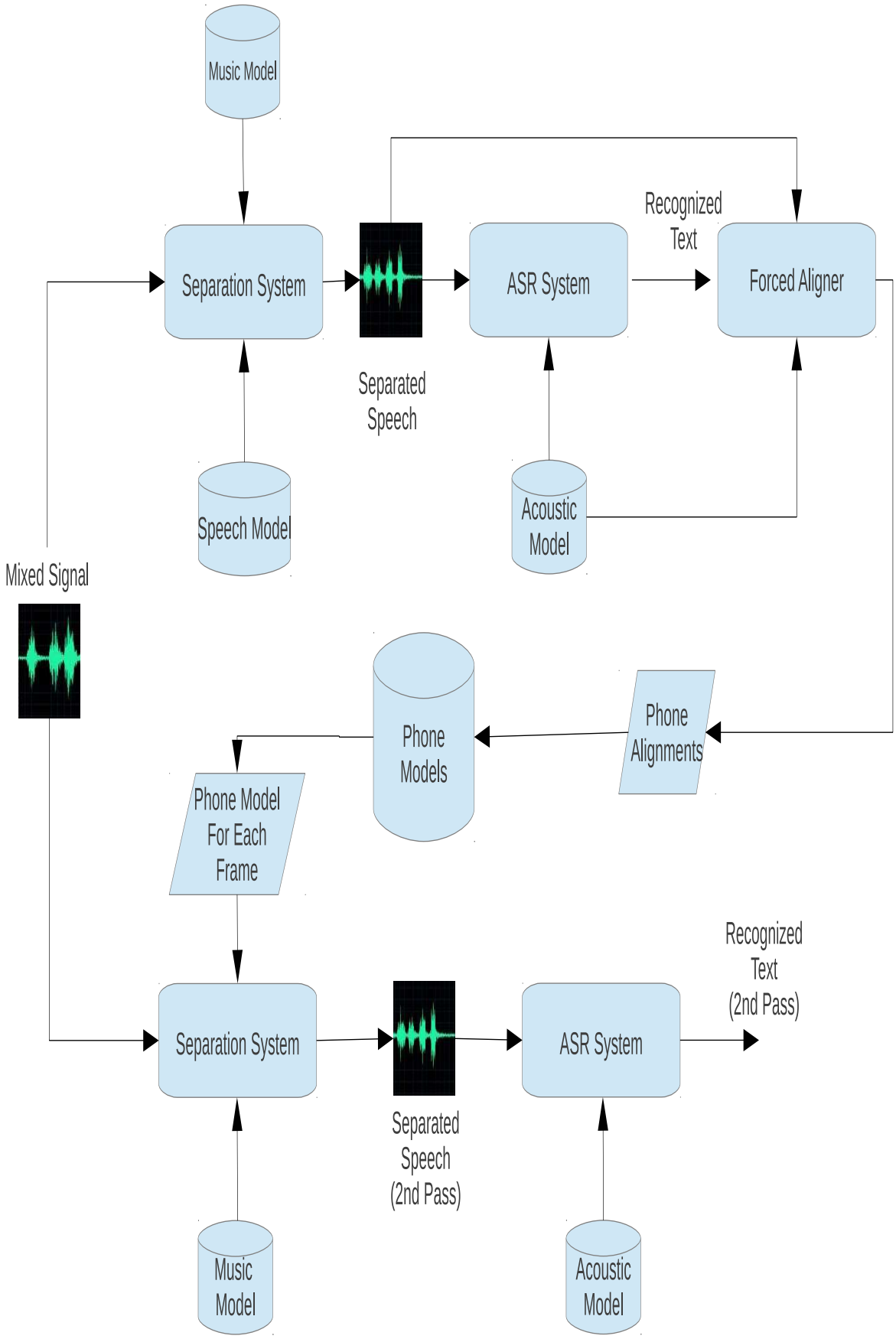

Figure 6.6. Multi-pass separation strategy.

7. CONCLUSION

In this dissertation, we investigate single-channel speech-music separation problem for ASR task in a probabilistic perspective. We proposed a mixture of NMF model for representing spectrum of speech-music mixture signal. As a baseline method, we used traditional NMF based separation method for single-channel source separation. Following sections summarize the main conclusions for each previous chapter.

7.1. NMF Based Single-Channel Source Separation

In Chapter 3, we described an NMF-based single-channel speech-music separation method. In Chapter 3, we did not only describe the basics of NMF as a matrix factorization method, but also probabilistic interpretations corresponds to both NMF model with KL and IS divergences were used in development of speech-music separation task. In Chapter 3, we analyzed the effect of training data type to speech-music separation performance. For modeling speech signal, we used following 4 training data types in the separation:

- 'Self' case refers to the training data of the target speaker which is the same as the mixed signal which has to be separated.
- 'Other' case refers to the training data from 3 different people who are from the same gender as the target speaker.
- 'All' case refers to the training data from 4 people which includes the target speaker.
- 'None' case refers to no training data is used for modeling the speech signal.

Similarly, for modeling music signal, we also use 4 training data types as:

- 'Original' case refers to the jingle itself which is used to create the background music signal of the mixed signal. In 'Original' case, the frames of the jingle are used as the template vectors of NMF model.

- 'Self' case refers to the jingle itself which is used to create the background music signal of the mixed signal. The templates are trained from the jingle frames.
- 'Other' case refers to the training data from 9 different jingle which are not used in background music generation.
- 'All' case refers to the jingles which includes the jingle that is used for the background music generation.

For both of divergence measures and all input SMR values, using the frames of the jingle as the template vectors ('Original') model outperformed the other music models. Therefore, in case of the known jingle, instead of using a trained NMF model, the frames of the jingle must be used as the template vectors of NMF model in the separation phase. In case of 'Original' model for the music signal, using a pre-trained speech model outperformed the 'None' model which corresponds to estimating the speech signal from the mixture. However, the type of the training data for the speech signal is not important for separation performance point of view. It is a good news for us due to the fact that for all experimental conditions it cannot be ensured that the target speaker data can be included in the speech model training.

There is a different approach in 'None' and other modeling techniques for the speech signal. In 'None' modeling approach, the templates and the corresponding excitations are estimated from the mixed signal simultaneously. In other words, the templates are trained from the mixed signal. The advantage of this approach is to consider the structure of the mixed signal. However, the disadvantage is that the estimation is done from a noisy signal. For pre-trained speech modeling ('Self', 'All' and 'Other') approach for the speech signal, the advantage is to train the templates form the clean speech. The disadvantage in this case is not to consider the mixed signal to be separated.

7.2. Music Modeling for Speech-Music Separation

In Chapter 4, we developed a probabilistic approach to the single-channel speech-music separation problem with the assumption that music part of the mixture is gen-

erated from a known catalog of the jingles. For each mixed signal, we assumed, one of the jingles in the catalog is used for creating the background music signal by choosing some random parts of the jingle via applying a gain or frequency filtering to the selected part of the jingle.

First, we used three different approaches, a mixture model, Markovian model and NMF model with two different divergence measures, KL and IS, to model the music signal and compare the performances of these techniques. The novelty of the proposed approach was due to the fact that we used different models for each source signal. As a result, we combined an NMF model for the speech signal and a mixture model for the music signal for representing the spectrum of speech-music mixture signal.

Moreover, we addressed the gain estimation problem of the mixture-based methods and proposed GMC for KL divergence and IGMC for IS divergence structures to overcome the gain estimation issue. It was shown that for both divergence measures imposing the correlation between the gain parameters with GMC or IGMC techniques improved the gain estimation performance of the mixture-based method. In order to impose the temporal dependency on the jingle frames, we proposed applying a Markovian structure on the jingle frames and it was experimentally shown that though in KL case Markovian structure improved the separation performance significantly, the effect of using Markovian structure in IS case was almost negligible. Furthermore, we used the developed methods as a front-end in an ASR system and showed that the ASR performance could be improved using such systems.

Furthermore, we tested the proposed systems with the real data recordings and showed that catalog based methods could improve the speech recognition performance and outperformed the traditional NMF based methods. As a reference separation method, speech enhancement method, logMMSE, was used to reduce the effect of background music. However, the usage of the speech enhancement technique did not improve the speech recognition performance as compared to the mixed signals. As a result, this work was an initial attempt to develop probabilistic models to solve background music removal problem and the experimental results were promising for

the future studies.

7.3. Speech Modeling for Speech-Music Separation

In Chapter 5, we focused on modeling approaches for speech signal for speech-music separation task. We proposed using pre-trained speech models as a prior in the separation phase. We extended the proposed mixture of NMF based method using prior models for the speech signal. For both divergence measures, using prior speech model strategy was applied. While Gamma prior was applied for Poisson model, Inverse-Gamma prior was applied for complex Gaussian model.

We developed a variational inference method for Poisson (KL) and complex Gaussian model (IS) which uses the prior speech model in the separation method. We evaluated the proposed separation method in speech recognition test and show the advantage of using prior speech model in the separation method. Moreover, we compared the effect of prior speech model type in Poisson (KL) and complex Gaussian (IS) models with different gain estimation strategies. As a result we showed that for both observation models incorporating prior speech information improved the separation performance of the mixture of NMF based separation method.

Furthermore, it was shown that the separation performance of 'Other' type model was as good as the 'Self' and 'All' type models. This was a good result for the speech-music separation systems due to the fact that it is not always possible to make sure that the speaker in the mixed segment of the audio are in the training data of the speech model. It was surprising that the separation results obtained using the 'Self' model were not better than 'All' and 'Other' models.

7.4. Sub-word Specific Speech Models for Speech-Music Separation

In Chapter 6, we proposed using sub-word models for representing the speech signal in speech-music separation task. We used NMF models for both source signals and focus on modeling speech signal in detailed. The main motivation in this study is

to increase performance of an ASR system by background music removal. Therefore, acoustic model training data of the ASR system can be used for learning the parameters of sub-word model to be used during the separation of the background music. In this thesis, we applied forced-alignment strategy to provide the training data for the sub-word units.

We analyzed the performance improvement of the separation system with an assumption that for each time frame, the sub-word unit identity is known as a prior. It is shown that with known sub-word unit identities for each time frame, the recognition accuracy of the separated speech signal can reach the accuracy with clean speech data. We prose a multi-pass separation strategy in this study and show that using the sub-word units instead of a general speech improves the recognition accuracy.

7.5. Future Work

We developed a probabilistic approach to single-channel speech-music separation problem for improving an ASR system performance by background music removal. We assume that jingle that generates the background music is known as a prior. Although the identity of the jingle can be detected using the music part of the mixed signal, it cannot guaranteed that whole background music signal is generated by this jingle. Therefore, as a future work, it should be assumed that in addition to known jingle frames, there can be some jingle components which cannot be observed in music part of the audio signal.

In Chapter 4, real time factors of traditional NMF based and proposed mixture of NMF based methods are compared. The computational cost of the proposed method is higher than the traditional NMF based approach. However, the proposed method can be processed in parallel for each frame of the jingle due to the fact that an NMF update is required in the proposed method. In order to use the proposed method in a practical system, it should be parallelized as a future work.

The linear spectrum of the signals is used as a feature in this study. However,

Mel-scaled spectrum can also be used in the proposed method as a feature. Using Mel-scaled spectrum not only enables to model the source signal in an effective way, but also decreases the computational cost of the method due to decreased number of frequency bins as compared to linearly spaced spectrum.

In Chapter 6, although sub-word units are used for representing the speech signal, for both speech and music signals are modeled using NMF method in contrast to representing the music signal with a mixture model. As a future work, NMF models of sub-word modeling units should be used with a mixture model for music signal. Moreover, a general training data is used for learning sub-word based NMF models but as similar to Chapter 5, the effect of training data type must be investigated for making a deep analysis of separation performance of the sub-word based method.

APPENDIX A: DISTRIBUTION PROPERTIES

- Gamma Distribution is defined as:

$$\mathcal{G}(x;a,b) = \exp((a-1)\log x - \frac{x}{b} - a\log b - \log\Gamma(a)) \tag{A.1}$$

- Mean and variance of Gamma distribution can be calculated using scale and shape parameters as:

$$E[x] = ab \tag{A.2}$$
$$Var[x] = ab^2 \tag{A.3}$$

- Inverse-Gamma Distribution is defined as:

$$\mathcal{IG}(x;a,b) = \exp(-(a+1)\log x - \frac{1}{bx} - a\log b - \log\Gamma(a)) \tag{A.4}$$

- Mean and variance of Inverse-Gamma distribution can be calculated using scale and shape parameters as:

$$E[x] = \frac{b}{a-1} \tag{A.5}$$
$$Var[x] = \frac{b^2}{(a-1)^2(a-2)} \tag{A.6}$$

- Distribution of a random variable which is inverse of Gamma Distributed random variable is defined as:

$$x \sim \mathcal{G}(x;a,b) \rightarrow \frac{1}{x} \sim \mathcal{IG}(\frac{1}{x};a,\frac{1}{b}) \tag{A.7}$$

- Distribution of a random variable which is inverse of Inverse-Gamma Distributed random variable is defined as:

$$x \sim \mathcal{IG}(x;a,b) \rightarrow \frac{1}{x} \sim \mathcal{G}(\frac{1}{x};a,\frac{1}{b}) \tag{A.8}$$

- Poisson Distribution is defined as:

$$\mathcal{PO}(s;\lambda) = \exp(s\log\lambda - \lambda - \log\Gamma(s+1)) \tag{A.9}$$

- Complex Gaussian Distribution is defined as:

$$N_c(x|\mu,\lambda) = |\pi\Sigma|^{-1}\exp(-(x-\mu)^H\Sigma^{-1}(x-\mu)) \tag{A.10}$$

APPENDIX B: DERIVATIONS OF MIXTURE OF NMF MODEL UPDATE EQUATIONS

B.1. Update Equations for Poisson Case

After calculating the expectations, we can find the model parameters that maximize the likelihood of the data. In M-Step, the expected value of the joint log-likelihood of the data and the latent sources under the posterior distribution of the latent sources, which is represented as Q, is calculated and used for finding the maximizing model parameters. Q value can be computed as follows:

$$\begin{aligned} Q &= \sum_{f,t,j} \langle [r_t = j] \rangle \Big\{ \big[\sum_b -D_{fb}E_{bt} + \langle s^j_{fbt} \rangle \log(D_{fb}E_{bt}) - \langle \log \Gamma(s^j_{fbt}+1) \rangle \big] \\ &- C_{fj}h_f v_t + \langle m^j_{ft} \rangle \log(C_{fj}h_f v_t) - \langle \log \Gamma(m^j_{ft}+1) \rangle + \langle \log \delta(X_{ft} - \sum_b s^j_{fbt} - m^j_{ft}) \rangle \Big\} \\ &=^c \sum_{f,t,j} \langle [r_t = j] \rangle \Big\{ \big[\sum_b -D_{fb}E_{bt} + \langle s^j_{fbt} \rangle \log(D_{fb}E_{bt}) \big] - C_{fj}h_f v_t + \langle m^j_{ft} \rangle \log(C_{fj}h_f v_t) \Big\} \end{aligned}$$

We compute the parameters of the speech spectrum, D and E matrices. Each entry of the template matrix, D, can be calculated as

$$\frac{\partial Q}{\partial D_{fb}} = \sum_{t,j} \langle [r_t = j] \rangle \left[\frac{\langle s^j_{fbt} \rangle}{D_{fb}} - E_{bt} \right] = 0 \tag{B.1}$$

$$D_{fb} = \frac{\sum_{t,j} \langle [r_t = j] \rangle \langle s^j_{fbt} \rangle}{\sum_t E_{bt}} \tag{B.2}$$

Now, we find the each entry of the excitation matrix of the speech spectrogram, **E**, using the following equation

$$\frac{\partial Q}{\partial E_{bt}} = \sum_{f,j} \langle [r_t = j] \rangle \left[\frac{\langle s^j_{fbt} \rangle}{E_{bt}} - D_{fb} \right] = 0 \tag{B.3}$$

$$E_{bt} = \frac{\sum_{f,j} \langle [r_t = j] \rangle \langle s^j_{fbt} \rangle}{\sum_f D_{fb}} \tag{B.4}$$

$$\tag{B.5}$$

We want to find the filtering parameter for each frequency bin, h_f, and gain parameter for each time frame, v_t. The filtering parameter for each frequency bin can be found using

$$\frac{\partial Q}{\partial h_f} = \sum_{t,j} \langle [r_t = j] \rangle \left[\frac{\langle m^j_{ft} \rangle}{h_f} - C_{fj} v_t \right] = 0 \tag{B.6}$$

$$h_f = \frac{\sum_{t,j} \langle [r_t = j] \rangle \langle m^j_{ft} \rangle}{\sum_{t,j} \langle [r_t = j] \rangle C_{fj} v_t} \tag{B.7}$$

The gain parameter for time t can be found using

$$\frac{\partial Q}{\partial v_t} = \sum_{f,j} \langle [r_t = j] \rangle \left[\frac{\langle m^j_{ft} \rangle}{v_t} - C_{fj} h_f \right] = 0 \tag{B.8}$$

$$v_t = \frac{\sum_{f,j} \langle [r_t = j] \rangle \langle m^j_{ft} \rangle}{\sum_{f,j} \langle [r_t = j] \rangle C_{fj} h_f} \tag{B.9}$$

B.2. Update Equations for complex Gaussian Case

The parameters of the posterior distribution for the latent speech source can be calculated using the following equations:

$$
\begin{aligned}
p(s^j_{fbt}|X, r_t\Theta) &= \frac{p(s^j_{fbt}|\Theta)p(X|s^j_{fbt}, r_t, \Theta)}{p(X|r_t, \Theta)} = N_c(s^j_{fbt}; \mu^j_{fbt}, \Sigma^j_{fbt}) \\
p(X|s^j_{fbt}, r_t, \Theta) &= N_c(X - s^j_{fbt}; 0, \sum_{i\neq b} D_{fi}E_{it} + C_{fj}h_f v_t) \\
\log p(s^j_{fbt}|X, r_t\Theta) &= {}^c - \frac{(s^j_{fbt})^2}{D_{fb}E_{bt}} - \frac{(X - s^j_{fbt})^2}{\sum_{i\neq b} D_{fi}E_{it} + C_{fj}h_f v_t} = -\frac{(s^j_{fbt} - \mu^j_{fbt})^2}{\Sigma^j_{fbt}}
\end{aligned}
$$

The conditional posterior mean and variance of b-th speech source in frequency bin f and time frame t conditioned on j-th jingle frame are:

$$
\begin{aligned}
-\frac{(s^j_{fbt})^2}{\Sigma^j_{fbt}} &= -\frac{(s^j_{fbt})^2}{D_{fb}E_{bt}} - \frac{(X_{ft} - s^j_{fbt})^2}{\sum_{i\neq b} D_{fi}E_{it} + C_{fj}h_f v_t} && \text{(B.10)} \\
\frac{1}{\Sigma^j_{fbt}} &= \frac{1}{D_{fb}E_{bt}} + \frac{1}{\sum_{i\neq b} D_{fi}E_{it} + C_{fj}h_f v_t} && \text{(B.11)} \\
\Sigma^j_{fbt} &= \frac{(D_{fb}E_{bt})(\sum_{i\neq b} D_{fi}E_{it} + C_{fj}h_f v_t)}{D_{fb}E_{bt} + \sum_{i\neq b} D_{fi}E_{it} + C_{fj}h_f v_t} && \text{(B.12)} \\
&= \frac{D_{fb}E_{bt}}{\sum_b D_{fb}E_{bt} + C_{fj}h_f v_t}(\sum_{i\neq b} D_{fi}E_{it} + C_{fj}h_f v_t) && \text{(B.13)} \\
-\frac{(-2s^j_{fbt}\mu^j_{fbt})}{\Sigma^j_{fbt}} &= -\frac{(-2X_{ft}s^j_{fbt})}{\sum_{i\neq b} D_{fi}E_{it} + C_{fj}h_f v_t} && \text{(B.14)} \\
\mu^j_{fbt} &= \frac{\Sigma^j_{fbt}}{\sum_{i\neq b} D_{fi}E_{it} + C_{fj}h_f v_t}X_{ft} && \text{(B.15)} \\
&= \frac{D_{fb}E_{bt}}{\sum_b D_{fb}E_{bt} + C_{fj}h_f v_t}X_{ft} && \text{(B.16)}
\end{aligned}
$$

The parameters of the posterior distribution for the latent music source can be calculated using the following equations:

$$
\begin{aligned}
p(m_{ft}^j|X, r_t\Theta) &= \frac{p(m_{ft}^j|\Theta)p(X|m_{ft}^j, r_t, \Theta)}{p(X|r_t, \Theta)} = N_c(m_{ft}^j; \mu_{ft}^j, \Sigma_{ft}^j) \\
p(X|m_{ft}^j, r_t, \Theta) &= N_c(X - m_{ft}^j; 0, \sum_b D_{fb}E_{bt}) \\
\log p(s_{ft}^j|X, r_t\Theta) &= {}^c - \frac{(m_{ft}^j)^2}{C_{fj}h_f v_t} - \frac{(X - m_{ft}^j)^2}{\sum_b D_{fb}E_{bt}} = -\frac{(m_{ft}^j - \mu_{ft}^j)^2}{\Sigma_{ft}^j}
\end{aligned}
$$

The conditional posterior mean and variance of the latent music source in frequency bin f and time frame t conditioned on j-th jingle frame are:

$$
\begin{aligned}
-\frac{(m_{ft}^j)^2}{\Sigma_{ft}^j} &= -\frac{(m_{ft}^j)^2}{C_{fj}h_f v_t} - \frac{(X - m_{ft}^j)^2}{\sum_b D_{fb}E_{bt}} && \text{(B.17)} \\
\frac{1}{\Sigma_{ft}^j} &= \frac{1}{C_{fj}h_f v_t} + \frac{1}{\sum_b D_{fb}E_{bt}} && \text{(B.18)} \\
\Sigma_{ft}^j &= \frac{(C_{fj}h_f v_t)(\sum_b D_{fb}E_{bt})}{\sum_b D_{fb}E_{bt} + C_{fj}h_f v_t} && \text{(B.19)} \\
\Sigma_{ft}^j &= \frac{C_{fj}h_f v_t}{\sum_b D_{fb}E_{bt} + C_{fj}h_f v_t}(\sum_b D_{fb}E_{bt}) && \text{(B.20)} \\
-\frac{(-2m_{ft}^j\mu_{ft}^j)}{\Sigma_{ft}^j} &= -\frac{(-2X_{ft}m_{ft}^j)}{\sum_b D_{fb}E_{bt}} && \text{(B.21)} \\
\mu_{ft}^j &= \frac{\Sigma_{ft}^j}{\sum_b D_{fb}E_{bt}}X_{ft} && \text{(B.22)} \\
&= \frac{C_{fj}h_f v_t}{\sum_b D_{fb}E_{bt} + C_{fj}h_f v_t}X_{ft} && \text{(B.23)}
\end{aligned}
$$

After calculating the expectations, we can find the model parameters that maximize the likelihood of the data. In M-Step, the expected value of the joint log-likelihood of the data and the latent sources under the posterior distribution of the latent sources, which is represented as Q, is calculated and used to find the maximizing model parameters. Q value can be computed as follows:

$$
Q = \sum_{f,t,j}\langle[r_t = j]\rangle\Big\{\big[\sum_b - \log(D_{fb}E_{bt}) - \frac{\langle|s_{fbt}|^2\rangle}{D_{fb}E_{bt}}\big] - \log(C_{fj}h_f v_t) - \frac{\langle|m_{ft}|^2\rangle}{C_{fj}h_f v_t}\Big\}
$$

$$\frac{\partial Q}{\partial D_{fb}} = \sum_{t,j} \langle [r_t = j] \rangle \left[-\frac{1}{D_{fb}} + \frac{\langle |s_{fbt}|^2 \rangle}{D_{fb}^2 E_{bt}} \right] = 0 \tag{B.24}$$

$$D_{fb} = \frac{1}{T} \sum_{t,j} \langle [r_t = j] \rangle \frac{\langle |s_{fbt}^j|^2 \rangle}{E_{bt}} \tag{B.25}$$

Now, we find each entry of the excitation matrix of the speech spectrogram, E, using the following equation

$$\frac{\partial Q}{\partial E_{bt}} = \sum_{f,j} \langle [r_t = j] \rangle \left[-\frac{1}{E_{bt}} + \frac{\langle |s_{fbt}|^2 \rangle}{E_{bt}^2 D_{fb}} \right] = 0 \tag{B.26}$$

$$E_{bt} = \frac{1}{F} \sum_{f,j} \langle [r_t = j] \rangle \frac{\langle |s_{fbt}^j|^2 \rangle}{D_{fb}} \tag{B.27}$$

We want to find the filtering parameter for each frequency bin, h_f, and gain parameter for each time frame, v_t. The filtering parameter for each frequency bin can be found using

$$\frac{\partial Q}{\partial h_f} = \sum_{f,j} \langle [r_t = j] \rangle \left[-\frac{1}{h_f} + \frac{\langle |m_{ft}|^2 \rangle}{C_{fj} h_f^2 v_t} \right] = 0 \tag{B.28}$$

$$h_f = \frac{1}{T} \sum_{t,j} \langle [r_t = j] \rangle \frac{\langle |m_{ft}^j|^2 \rangle}{C_{fj} v_t}. \tag{B.29}$$

The gain parameter for time t can be found using

$$\frac{\partial Q}{\partial v_t} = \sum_{t,j} \langle [r_t = j] \rangle \left[-\frac{1}{v_t} + \frac{\langle |m_{ft}|^2 \rangle}{C_{fj} h_f v_t^2} \right] = 0 \tag{B.30}$$

$$v_t = \frac{1}{F} \sum_{b,j} \langle [r_t = j] \rangle \frac{\langle |m_{ft}^j|^2 \rangle}{C_{fj} h_f} \tag{B.31}$$

REFERENCES

1. Arisoy, E., D. Can, S. Parlak, H. Sak and M. Saraclar, "Turkish Broadcast News Transcription and Retrieval", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 17, No. 5, pp. 874–883, 2009.

2. Raj, B., V. Parikh and R. Stern, "The Effects of Background Music on Speech Recognition Accuracy", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 851–854, 1997.

3. Jang, G. and T. Lee, "A Maximum Likelihood Approach to Single-Channel Source Separation", *The Journal of Machine Learning Research*, Vol. 4, pp. 1365–1392, 2003.

4. Demir, C., A. T. Cemgil and M. Saraçlar, "Catalog-Based Single-Channel Speech-Music Separation", *11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, pp. 2782–2785, 2010.

5. Cemgil, A. T., "Bayesian Inference in Non-Negative Matrix Factorisation Models", *Computational Intelligence and Neuroscience*, p. 17, 2009.

6. Févotte, C., N. Bertin and J. Durrieu, "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis", *Neural Computation*, Vol. 21, No. 3, pp. 793–830, 2009.

7. Demir, C., A. T. Cemgil and M. Saraçlar, "Catalog-Based Single-Channel Speech-Music Separation For Automatic Speech Recognition", *19th European Signal Processing Conference (EUSIPCO 2011)*, pp. 2133–2137, 2011.

8. Demir, C., A. T. Cemgil and M. Saraçlar, "Catalog-based Single-Channel Speech-Music Separation with the Itakura-Saito Divergence", *20th European Signal Processing Conference (EUSIPCO 2012)*, pp. 2812–2816, IEEE, 2012.

9. Demir, C., A. T. Cemgil and M. Saraçlar, "Gain Estimation Approaches in Catalog-Based Single-Channel Speech-Music Separation", *Workshop on Automatic Speech Recognition and Understanding (ASRU 2011)*, pp. 185–190, 2011.

10. Demir, C., M. Saraçlar and A. T. Cemgil, "Single-Channel Speech-Music Separation for Robust ASR With Mixture Models", *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 21, No. 4, pp. 725–736, 2013.

11. Demir, C., A. T. Cemgil and M. Saraçlar, "Semi-supervised Single-Channel Speech-Music Separation For Automatic Speech Recognition", *12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, pp. 681–684, 2011.

12. Demir, C., A. T. Cemgil and M. Saraçlar, "Effect of Speech Priors in Single-Channel Speech-Music Separation for ASR.", *13th Annual Conference of the International Speech Communication Association (Interspeech 2012)*, pp. 1235–1238, 2012.

13. Raj, B., R. Singh and T. Virtanen, "Phoneme-Dependent NMF for Speech Enhancement in Monoural Mixtures", *12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, pp. 1217–1220, 2011.

14. Schmidt, M. and R. Olsson, "Single-Channel Speech Separation Using Sparse Non-Negative Matrix Factorization", *International Conference on Spoken Language Processing (ICSLP 2006)*, pp. 2614–2617, 2006.

15. Kristjansson, T., J. Hershey, P. Olsen, S. Rennie and R. Gopinath, "Super-Human Multi-Talker Speech Recognition: The IBM 2006 Speech Separation Challenge System", *International Conference on Spoken Language Processing (ICSLP 2006)*, pp. 97–100, 2006.

16. Vanroose, P., "Blind Source Separation of Speech and Background Music for Improved Speech Recognition", *The 24th Symposium on Information Theory*, pp.

103–108, 2003.

17. Blouet, R., G. Rapaport and C. Fevotte, "Evaluation of Several Strategies for Single Sensor Speech/Music Separation", *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, pp. 37–40, 2008.

18. Raj, B., T. Virtanen, S. Chaudhuri and R. Singh, "Non-Negative Matrix Factorization Based Compensation of Music for Automatic Speech Recognition", *11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, pp. 717–720, 2010.

19. Grais, E. M. and H. Erdoğan, "Single Channel Speech Music Separation Using Nonnegative Matrix Factorization and Spectral Masks", *17th International Conference on Digital Signal Processing (DSP), 2011*, pp. 1–6, IEEE, 2011.

20. Choi, S., A. Cichocki, H. Park and S. Lee, "Blind Source Separation and Independent Component Analysis: A Review", *Neural Information Processing-Letters and Reviews*, Vol. 6, No. 1, pp. 1–57, 2005.

21. Weiss, R. and D. Ellis, "Speech Separation Using Speaker-Adapted Eigenvoice Speech Models", *Computer Speech & Language*, Vol. 24, pp. 16–29, 2010.

22. Smaragdis, P., M. Shashanka, M. Inc and B. Raj, "A Sparse Non-Parametric Approach for Single Channel Separation of Known Sounds", *Proc. of Neural Information Processing Systems (NIPS)*, pp. 1705–1713, 2009.

23. Benaroya, L., F. Bimbot and R. Gribonval, "Audio Source Separation with a Single Sensor", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 1, pp. 191–199, 2006.

24. Virtanen, T., "Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 3, pp. 1066–1074, 2007.

25. Roweis, S., "One Microphone Source Separation", *Advances in neural information processing systems*, pp. 793–799, 2001.

26. Ghahramani, Z. and M. Jordan, "Factorial Hidden Markov Models", *Machine learning*, Vol. 29, No. 2, pp. 245–273, 1997.

27. Benaroya, L. and F. Bimbot, "Wiener Based Source Separation with HMM/GMM Using a Single Sensor", *International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003)*, pp. 957–961, 2003.

28. Srinivasan, S., J. Samuelsson and W. Kleijn, "Codebook-Based Bayesian Speech Enhancement", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 1, pp. 1077–1080, 2005.

29. Benaroya, L., L. Donagh, F. Bimbot and R. Gribonval, "Non Negative Sparse Representation for Wiener Based Source Separation with a Single Sensor", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 6, 2003.

30. Tsai, W., D. Rodgers and H. Wang, "Blind Clustering of Popular Music recordings Based on Singer Voice Characteristics", *Computer Music Journal*, Vol. 28, No. 3, p. 78, 2004.

31. Ozerov, A., P. Philippe, R. Gribonval and F. Bimbot, "One Microphone Singing Voice Separation Using Source-Adapted Models", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 90–93, 2005.

32. Gales, M., D. Pye and P. C. Woodland, "Variance Compensation within the MLLR Framework for Robust Speech recognition and Speaker Adaptation", *International Conference on Spoken Language Processing (ICSLP 1996)*, Vol. 3, pp. 1832–1835, IEEE, 1996.

33. Ellis, D. and R. Weiss, "Model-Based Monaural Source Separation Using a Vector-

Quantized Phase-Vocoder Representation", *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 957–960, 2006.

34. Eggert, J. and E. Korner, "Sparse Coding and NMF", *2004 IEEE International Joint Conference on Neural Networks Proceedings*, Vol. 4, pp. 2529–2533, 2004.

35. Grais, E. M. and H. Erdoğan, "Single Channel Speech Music Separation Using Nonnegative Matrix Factorization with Sliding Windows and Spectral Masks", *12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, 2011.

36. Grais, E. M. and H. Erdoğan, "Hidden Markov Models as Priors for Regularized Nonnegative Matrix Factorization in Single-Channel Source Separation", *13th Annual Conference of the International Speech Communication Association (Interspeech 2012)*, 2012.

37. Hoyer, P. O., "Non-Negative Sparse Coding", *12th IEEE Workshop on Neural Networks for Signal Processing*, pp. 557–565, IEEE, 2002.

38. Hoyer, P., "Non-Negative Matrix Factorization with Sparseness Constraints", *The Journal of Machine Learning Research*, Vol. 5, p. 1469, 2004.

39. Virtanen, T., "Separation of Sound Sources by Convolutive Sparse Coding", *ISCA Tutorial and Research Workshop (ITRW) on Statistical and Perceptual Audio Processing*, 2004.

40. Smaragdis, P., "Non-Negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs", *International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2004)*, pp. 494–499, 2004.

41. O'Grady, P. and B. Pearlmutter, "Convolutive Non-Negative Matrix Factorisation with a Sparseness Constraint", *Proceedings of the 2006 16th IEEE Signal Pro-*

cessing Society Workshop on Machine Learning for Signal Processing, 2006, pp. 427–432, 2006.

42. Lee, D. D. and H. S. Seung, "Learning the Parts of Objects by Non-Negative Matrix Factorization", *Nature*, Vol. 401, No. 6755, pp. 788–791, 1999.

43. Lee, D. D. and H. S. Seung, "Algorithms for Non-Negative Matrix Factorization", *Advances in neural information processing systems*, pp. 556–562, 2000.

44. Virtanen, T., A. Cemgil and S. Godsill, "Bayesian Extensions to Non-Negative Matrix Factorisation for Audio Signal Modelling", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1825–1828, 2008.

45. Cemgil, A. and O. Dikmen, "Conjugate Gamma Markov Random Fields for Modelling Nonstationary Sources", *Independent Component Analysis and Signal Separation (ICA 2007)*, pp. 697–705, 2007.

46. Dikmen, O. and A. T. Cemgil, "Unsupervised Single-Channel Source Separation Using Bayesian NMF", *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*, pp. 93–96, IEEE, 2009.

47. Virtanen, T. and A. T. Cemgil, "Mixtures of Gamma Priors for Non-Negative Matrix Factorization Based Speech Separation", *Independent Component Analysis and Signal Separation (ICA 2009)*, pp. 646–653, Springer, 2009.

48. Vincent, E., R. Gribonval and C. Févotte, "Performance Measurement in Blind Audio Source Separation", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, pp. 1462–1469, 2006.

49. Ozerov, A., C. Févotte and M. Charbit, "Factorial Scaled Hidden Markov Model for Polyphonic Audio Representation and Source Separation", *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*, pp. 121–124, IEEE, 2009.

50. Mysore, G., P. Smaragdis and B. Raj, "Non-Negative Hidden Markov Modeling of Audio with Application to Source Separation", *Latent Variable Analysis and Signal Separation*, pp. 140–148, 2010.

51. Nakano, M., J. Le Roux, H. Kameoka, Y. Kitano, N. Ono and S. Sagayama, "Non-negative Matrix Factorization with Markov-Chained Bases for Modeling Time-Varying Patterns in Music Spectrograms", *Latent Variable Analysis and Signal Separation*, pp. 149–156, 2010.

52. Ephraim, Y. and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-spectral Amplitude Estimator", *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 33, No. 2, pp. 443–445, 1985.

53. Cohen, I., "Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator", *Signal Processing Letters*, Vol. 9, No. 4, pp. 113–116, 2002.

54. Malah, D., R. Cox and A. Accardi, "Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-stationary Noise Environments", *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 2, pp. 789–792, IEEE, 1999.