Speech Enhancement using Training-based Non-negative Matrix Factorization Techniques

Hanwook Chung

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

July 2018

Abstract

In this thesis, we develop novel training-based non-negative matrix factorization (NMF) algorithms for single and multi-channel speech enhancement.

After introducing the problem and reviewing background material, we first present a regularized NMF algorithm with Gaussian mixtures and masking model for single-channel speech enhancement. The proposed framework seeks to exploit the statistical properties of the clean speech and noise. This is accomplished by including the log-likelihood functions (LLF) of the clean speech and noise magnitude spectra, based on Gaussian mixture models (GMM), as the regularization terms in the NMF cost function. Moreover, we incorporate the masking effects of the human auditory system to further improve the enhanced speech quality.

Second, we introduce a training and compensation algorithm of the class-conditioned NMF model for single-channel speech enhancement. The main goal is to reduce the residual noise components that have features similar to the speech. To this end, during the training stage, the basis vectors of different sources are obtained in a way that prevents them from representing each other, based on the concept of classification. Another goal is to handle the mismatch between the characteristics of the training and test data. This is accomplished by employing extra free basis vectors during the enhancement stage to capture the features which are not included in the training data.

Finally, we present a novel multi-channel speech enhancement algorithm based on a Bayesian NMF model. Essentially, we consider the Poisson-distributed latent variable model for multi-channel NMF. During the training stage, the NMF parameters are estimated from the tensor-based training data. During the enhancement stage, the clean speech signal is estimated via the NMF-based minimum variance distortionless response (MVDR) beamforming technique. To this end, the source locations are estimated by observing the spatial output power of a delay-and-sum (DS) beamformer applied to the NMF-based preprocessed noisy speech signal.

For each one of the above algorithms, objective experiments are carried out for different combinations of speaker, noise types and signal-to-noise ratio. The results show that the proposed methods provide better speech enhancement performance than the selected benchmark algorithms under considered test conditions.

Sommaire

Dans cette thèse, nous développons des algorithmes novateurs de rehaussement de la parole à un ou plusieurs canaux faisant appel à la factorisation matricielle non négative (*non-negative matrix factorization* - NMF) avec entrainement.

Après avoir introduit la problématique et passé en revue les connaissances de base pertinentes, nous présentons tout d'abord un algorithme de rehaussement à un canal qui utilise une approche NMF régularisée comportant un mélange de gaussiennes ainsi qu'un modèle de masquage. Le cadre proposé vise à exploiter les propriétés statistiques de la parole non-bruitée et du bruit. Ceci est accompli en incluant les fonctions log-vraisemblance des spectres d'amplitude de parole et de bruit, modélisées à l'aide de mélanges gaussiens, comme étant les termes de régularisation dans la fonction de coût de la NMF. De plus, nous intégrons les effets de masquage du système auditif humain afin d'améliorer davantage la qualité de la parole rehaussée.

Deuxièmement, nous introduisons un algorithme d'entrainement et de compensation d'un modèle NMF conditionné par la classe pour le problème du rehaussement de la parole à un canal. L'objectif principal est de réduire les composantes de bruit résiduel qui ont des caractéristiques similaires à la parole. A cette fin, les vecteurs de base des différentes sources sont obtenus au cours de la phase d'entrainement sur la base de la notion de classification qui les empêche de se représenter les uns les autres. Un autre objectif consiste à gérer les différences entre les caractéristiques des données d'entrainement et de test. Ceci est accompli en incluant des vecteurs de base supplémentaires pendant l'étape de rehaussement afin de capter les caractéristiques qui ne sont pas incluses dans les données d'apprentissage.

Finalement, nous présentons un nouvel algorithme de rehaussement de la parole à plusieurs canaux basé sur un modèle NMF bayésien. Essentiellement, nous considérons le modèle de variable latente avec une distribution de Poisson dans une version de l'algorithme NMF à plusieurs canaux. Pendant la phase d'entraînement, les paramètres NMF sont estimés à partir de données formées de tenseurs. À l'étape du rehaussement, le signal de parole non bruité est estimé à l'aide de la technique de formation de faisceau à variance minimale sans distorsion (*minimum variance distortionless response* - MVDR) et d'une NMF. Spécifiquement, l'emplacement des sources est estimé en observant la puissance de sortie spatiale d'un dispositif de formation de faisceau par retard et addition (*delay-and-sum*) appliqué au signal de parole bruité prétraité par une NMF.

Pour chacun des algorithmes ci-dessus, des expériences objectives sont effectuées pour différentes combinaisons de types de locuteurs et de bruits. Les résultats montrent que les méthodes proposées offrent de meilleures performances de rehaussement de la parole que les algorithmes de référence sélectionnés, et ce pour plusieurs conditions.

Acknowledgments

First and foremost, I would like to thank Professor Benoit Champagne for his guidance, based on strong scientific rigor, and for the many great opportunities he gave, including attending international conferences, that have all enhanced my theoretical knowledge and research experience. I also appreciate my co-supervisor Professor Eric Plourde for providing guidance and validating theoretical development while preparing the manuscripts.

I am also grateful for the financial support provided by a scholarship from Natural Sciences and Engineering Research Council of Canada (NSERC) and by Microsemi Corporation, without which this thesis would not have been possible.

I would also like to acknowledge the role of Mr. Dean Morgan and Patrick Lionais at Microsemi as well as Professor Wei-Ping Zhu at Concordia University for many interesting discussions and advises while conducting the research project.

This journey in Montreal would not have been the same without the encouragement from Dr. Jaeok Park, Dr. Dusik Kim and Dr. Kuwook Cha. I would also like to appreciate my companions in Korea, especially Dr. Sang Bae Chon for always giving many advises not only regarding the research but also the philosophical perspectives about life, and my best friends Mr. Yongjin Shin, Taejung Kwon and Chanwoong Chung for their heartfelt cheers. Finally, I would like to express my deepest gratitude to my beloved parents and my younger brothers.

Contents

1 Introduction 1

1.1 The Speech Enhancement Problem 1

1.2 Speech Enhancement Algorithms 2

1.3 Research Motivations 6

1.4 Research Objectives and Contributions 8

1.5 Thesis Organization 11

2 NMF-based Speech Enhancement 12

2.1 Background on NMF 12

2.2 Application of NMF to Speech Enhancement 20

2.2.1 Single-channel application 20

2.2.2 Multi-channel application 24

3 Regularized NMF with Gaussian Mixtures and Masking Model 31

3.1 Research Motivations and Contributions 32

3.2 Proposed Training Stage 34

3.3 Proposed Enhancement Stage 38

3.3.1 Regularized NMF with Gaussian mixtures 39

3.3.2 Weighted Wiener filtering based on masking threshold 42

3.4 Experiments 50

3.4.1 Methodology 51

3.4.2 Benchmark algorithms 53

3.4.3 Results 55

3.5 Summary 61

4 Training and Compensation of Class-conditioned NMF Bases 63
4.1 Research Motivations and Contributions . . . 64
4.2 Probabilistic Generative Models . . . 67
4.2.1 NMF model . . . 67
4.2.2 Classification model . . . 68
4.3 Proposed Training Stage . . . 69
4.3.1 Prior structures . . . 70
4.3.2 VBEM algorithm . . . 71
4.3.3 Parameter initialization and normalization . . . 74
4.4 Proposed Enhancement Stage . . . 75
4.4.1 Noise classification . . . 75
4.4.2 Basis compensation . . . 77
4.5 Experiments . . . 80
4.5.1 Methodology . . . 81
4.5.2 Benchmark algorithms . . . 84
4.5.3 Results . . . 86
4.6 Summary . . . 93

5 Multi-channel Extension of Bayesian NMF Model 94
5.1 Research Motivations and Contributions . . . 95
5.2 Proposed Training Stage . . . 97
5.2.1 PGM of multi-channel NMF with KL-divergence . . . 97
5.2.2 Prior structures . . . 98
5.2.3 VBEM algorithm . . . 99
5.2.4 Parameter initialization and normalization . . . 103
5.3 Proposed Enhancement Stage . . . 104
5.3.1 MNMF parameter estimation . . . 104
5.3.2 Source localization and noise correlation matrix computation . . . 105
5.3.3 NMF-based MVDR beamforming . . . 108
5.4 Experiments . . . 110
5.4.1 Methodology . . . 111
5.4.2 Benchmark algorithms . . . 113
5.4.3 Results . . . 114

5.5 Summary 115

6 Conclusion and Future Works 119

A Variational Lower Bound 122

A.1 Single-channel Class-conditioned NMF Model 122

A.2 Multi-channel Bayesian NMF Model 124

B A Brief Summary of NCP Method 126

References 128

List of Figures

2.1 Graphical illustration of the concept of the auxiliary function. 14

2.2 Geometric illustration of the propagation delay between a source and microphone array. 28

3.1 Example of masking threshold (dotted: normalized power spectrum of a female speaker, solid: masking threshold, dashed: absolute hearing threshold). 46

3.2 SDR, SIR and SAR values for different weighting factors in WWF. 47

3.3 Proposed mapping function from NMR, Φ_{kl}, to weighting factor, η_{kl}, based on a sigmoid function. 49

3.4 Simplified block diagrams of RNG and RNG-WWF methods. 51

3.5 Examples of proposed weighting factor. Each column from left to right respectively correspond to input SNR of 0, 5 and 10 dB. Each row from top to bottom shows the noisy speech magnitude spectrum, time-frequency representation of the proposed weighting factor and the weighting factor at the time frame of 1.3s and 2.1s. 56

3.6 SDR and SSNR comparisons for Factory 1 (top) and Hfchannel (bottom) noises. 58

3.7 Example of magnitude spectra of the clean, noisy and estimated clean speech for the benchmark and proposed algorithms. A female speech is degraded with Buccaneer 1 noise at 0 dB input SNR 61

4.1 A simplified block diagram of the proposed VNCP-BC method. 80

4.2 The posterior class probabilities $p(d_i = 1|\, \mathbf{w}_m)$. 87

4.3 An example of noise classification. Top shows the true noise type and bottom shows the estimated noise type using the proposed method. 88

4.4 Examples of magnitude spectrograms of the clean, noisy and estimated clean speech using the VNCP-BC method. A male speech is degraded by a noise consisting of different types as shown in Figure 4.3, at 0 dB input SNR. . . 89

5.1 A simplified block diagram of the proposed enhancement stage. 109

5.2 Geometric illustration of the spherical coordinates. 113

5.3 Examples of magnitude spectrograms of the clean speech, noisy speech ($j =$ 1) and estimated clean speech using the VMNMF method. A female speech is degraded by the Street noise at 0 dB input SNR. 118

List of Tables

3.1 A comparison between different perceptually-motivated and/or weighting methods . . . 54
3.2 Average results for Buccaneer 1 noise . . . 59
3.3 Average results for Hfchannel noise . . . 59
3.4 Average results for Babble noise . . . 60
3.5 Average results for Factory 1 noise . . . 60

4.1 Algorithm summary of the proposed enhancement stage . . . 81
4.2 Summary of the test noise types . . . 83
4.3 Average results for additive Bus noise (matched) . . . 90
4.4 Average results for additive Pedestrian noise (matched) . . . 90
4.5 Average results for additive Street noise (matched) . . . 90
4.6 Average results for additive Cafe noise (mismatched) . . . 91
4.7 Average results for additive Factory 1 noise (mismatched) . . . 91
4.8 Average results for additive Babble noise (mismatched) . . . 91
4.9 Average results for filtered noisy speech . . . 92
4.10 Comparison of computational times (in seconds) . . . 92

5.1 Algorithm summary of the proposed enhancement stage . . . 110
5.2 Average results for Bus noise . . . 116
5.3 Average results for Cafe noise . . . 116
5.4 Average results for Pedestrian noise . . . 117
5.5 Average results for Street noise . . . 117

List of Acronyms

AHTH	absolute hearing threshold
ARMA	auto-regressive moving average
ATF	acoustic transfer function
DFT	discrete Fourier transform
DNN	deep neural network
DoA	direction-of-arrival
EM	expectation-maximization
FWSSNR	frequency-weighted segmental signal-to-noise ratio
GMM	Gaussian mixture model
GSC	generalized sidelobe canceler
HMM	hidden Markov model
ICA	independent component analysis
IDFT	inverse discrete Fourier transform
IS	Itakura-Saito
KL	Kullback-Leibler
LCMV	linearly constrained minimum variance
LLF	log-likelihood function
MAP	maximum *a posteriori*
ML	maximum likelihood
MMSE	minimum mean-square error
MNMF	multi-channel non-negative matrix factorization
MOS	mean opinion score
MU	multiplicative update
MVDR	minimum variance distortionless response

NMF	non-negative matrix factorization
NMR	noise-to-mask ratio
NTF	non-negative tensor factorization
PCA	principal component analysis
PCP	posterior class probability
PDF	probability density function
PESQ	perceptual evaluation of speech quality
PGM	probabilistic generative model
PSD	power spectral density
RMM	Rayleigh mixture model
RNN	recurrent neural network
SAR	source-to-artifact ratio
SCM	spatial covariance matrix
SDR	source-to-distortion ratio
SIR	source-to-interference ratio
SNR	signal-to-noise ratio
SRP	steered response power
SSNR	segmental signal-to-noise ratio
STFT	short-time Fourier transform
STSA	short-time spectral amplitude
TDoA	time difference-of-arrival
VBEM	variational Bayesian expectation-maximization
VQ	vector quantization
WGM	weighted geometric mean
WWF	weighted Wiener filter

Chapter 1

Introduction

In this chapter, we first introduce the speech enhancement problem and present survey of several representative algorithms developed in the past to address this problem. This is followed by a statement of the research objectives and contributions of the thesis.

1.1 The Speech Enhancement Problem

Speech is one of the predominant means for communications between humans. In the context of electronic speech communication, the speech signal is generated by the speech production system of a human speaker, captured by a single or multiple microphones and transmitted through a certain medium such as optical fibers, copper wires or simply the air. Once the transmitted speech signal is received, it is reproduced through an electro-acoustic transducer such as a loudspeaker or earphone, and finally reaches the auditory system of a human listener. During the transmission, the speech signal is usually corrupted by various types of noises with acoustic or electromagnetic origin, yielding a noisy speech signal. The general objective of speech enhancement algorithms is to remove the additive background noise from a noisy speech signal in order to improve its quality (naturalness and freedom from distortion) and/or intelligibility (the likelihood of being correctly understood). These algorithms have been an attractive research area for decades and find diverse applications, including mobile telephony, hearing aids, speech coding and automatic speech recognition systems, to name a few. In the following subsection, we briefly introduce representative speech enhancement algorithms.

1.2 Speech Enhancement Algorithms

Numerous algorithms for speech enhancement have been proposed in the past decades. Depending on the number of microphones used for acquiring the noisy speech signal, the algorithms can be divided into two main groups, namely: single-channel and multi-channel.

One of the main advantages of the single-channel speech enhancement algorithms compared to the multi-channel ones, is their low computational complexity in general. Over the years, a considerable amount of research effort has been made on single-channel algorithms, leading to various approaches, such as: Wiener filtering [1, 2], spectral subtraction [3, 4], Bayesian minimum mean-square error (MMSE) estimators [5–7], subspace decomposition [8–10] and state-space [11–13] methods. The Wiener filtering and spectral subtraction methods are the most popular choices mainly due to the simplicity of their implementation. The Wiener filter is derived based on the minimum mean-square error (MMSE) criterion, whereas the spectral subtraction is performed based on the variance of the spectral coefficients, as estimated in the maximum likelihood (ML) sense. The Bayesian MMSE estimators for the clean speech spectral amplitudes employ an explicit prior structure for the statistics of the spectral coefficients. The subspace approach is based on the decomposition of the vector space of the observed noisy speech into two orthogonal subspaces, namely the signal subspace and noise subspace.

Regarding the multi-channel speech enhancement algorithms, their main advantage is that they can exploit the spatial features of the acoustic field through a spatio-temporal filter, also known as a beamformer, where coefficients can be designed to extract the clean speech from a given direction in an optimal way. Several methods have been proposed to design optimum beamformers, including: delay-and-sum (DS) [14], minimum variance distortionless response (MVDR) or linearly constrained minimum variance (LCMV) [15, 16], generalized sidelobe cancellation (GSC) [17], and eigen-space beamforming [18]. The main advantage of the DS beamformer is its computational efficiency, since it avoids matrix inversion computation. The LCMV and MVDR beamformers are designed to minimize the output power subject to a linear constraint and hence to preserve the target signal from a given direction while attenuating the ambient noise and interference. The GSC beamformer, which can handle multiple constraints, consists of a constrained and an unconstrained weight vectors. The unconstrained weight vector is used to cancel interference that leaks through the sidelobes of the beamformer specified by the constrained weight vectors. To further reduce the residual noise components in the enhanced speech obtained via a beamforming technique, the authors in [19, 20] apply a single-channel enhancement algorithm to the beamformer output as a post-processor.

These single and multi-channel algorithms were originally introduced by using a minimal amount of *a priori* information about the speech and noise. Consequently, they tend to provide limited performance, especially when the speech is contaminated by adverse noise, such as under low input signal-to-noise ratio (SNR) or non-stationary noise conditions. To overcome these limitations, machine learning (i.e., training-based) techniques have been applied to the speech enhancement task and have shown remarkable performance improvement in recent years. In a machine learning framework, the features of the clean speech and/or noise are obtained using the training data *a priori*, and subsequently used during the enhancement stage. For example, MMSE and maximum *a posteriori* (MAP)-based estimators for single-channel speech enhancement have been proposed by modeling the clean speech spectrum via a Rayleigh mixture model (RMM) [21] or Gaussian mixture model (GMM) [22–24], which provides a more detailed description of the speech distribution. In contrast to the clean speech model, where the parameters are derived from the training data, the noise spectrum is modeled by a single distribution where its model parameters are estimated directly from the noisy speech, which tend to limit the enhancement performance. Further improvements to the MAP estimator have been introduced lately [25], where the authors model both the clean speech and noise spectra by the GMMs. In order to better consider the time-varying spectral characteristics of the noisy speech, algorithms based on a hidden Markov model (HMM), as specified by GMM-based state-conditional densities, have been proposed [26–29]. The noisy speech signal is then modeled by combining the clean speech and noise HMMs, where the model parameters are obtained *a priori* using the training data. The clean speech can be estimated either via the approximated MAP estimator [26] or MMSE estimator [28,29]. The former case can be interpreted as using a single Wiener filter based on the dominant state and its corresponding mixture model parameters, whereas the latter case can be considered as a weighted sum of the state-dependent MMSE estimators where the weights are given by the posterior state probabilities. However, one of the main issues when implementing the HMM-based algorithms is the computational complexity, which grows rapidly as the model size (i.e., the number of states) increases.

Recently, the non-negative matrix factorization (NMF) approach has been successfully applied to various problems such as image representation [30], music transcription [31], single and multi-channel audio source separation [32,33] as well as single and multi-channel speech enhancement [34,35], as an alternative approach to the above mentioned methods[1]. In general, NMF is a dimensionality reduction tool, which decomposes a given observation matrix into basis and activation matrices with a non-negative element constraint [36, 37]. In audio and speech applications, the short-time magnitude or power spectrum can be interpreted as a linear combination of the basis vectors. Based on this representational aspect, the basis vectors are also referred to as a codebook or dictionary, that can be obtained *a priori* using training data. To this end, in a supervised NMF-based framework, the basis vectors are obtained using training data for each source during the training stage, and used subsequently during the test (i.e., separation or enhancement) stage. A number of variants of the NMF algorithms have been proposed, such as by considering various cost functions [38, 39], introducing necessary regularization terms (which corresponds to the prior structures within a statistical framework) [31, 32, 40] or developing more efficient algorithms for the parameter estimation [41–43]. Several representative methods will be discussed along with their limitations in the following subsection.

Deep neural network (DNN) algorithms have also gained enormous interest lately [44], and find diverse applications such as image classification [45] and automatic speech recognition [46]. The DNN training aims at estimating the nonlinear mapping function, specified by the weights and biases of the hidden layers of a processing network, that relates the input features to the output target features. The feed-forward DNN has been applied to single-channel speech enhancement [47, 48] as well as to multi-channel audio source separation [49]. To better capture the temporal dynamics, application of the recurrent neural network (RNN) to single-channel speech enhancement has been introduced in [50, 51]. A

[1] We note that the NMF algorithms (e.g., update rules or parameter estimation scheme) developed for a given application can be often used in other applications. That is, for example, the parameter estimation algorithm originally proposed for unsupervised audio source separation in [32] can be applied to other context, such as image processing or speech enhancement.

combination of NMF and DNN has also been proposed in [52, 53]. The NMF and DNN algorithms differ significantly in terms of underlying modeling structure and training requirements; in this thesis, we focus on a NMF model.

1.3 Research Motivations

To further improve the speech enhancement performance, several modified NMF algorithms have been introduced in recent years. In this subsection, we briefly review some of these contributions and comment on their limitations.

One main issue in a supervised framework is the existence of a mismatch between the characteristics of the training and test data, which in turn leads to a decreased quality of the enhanced speech signal; in particular, the enhanced speech may contain some residual noise components. A possible remedy to this problem is to add explicit regularization terms to the NMF cost function that incorporate some prior knowledge [40, 54]. In these algorithms, however, the basis vectors are fixed during the separation or enhancement stage, which limits the performance when there is a large mismatch between the training and test data. One alternative approach is to use a basis adaptation scheme during the enhancement stage, e.g., [55, 56]. In most basis adaptation algorithms, the basis vectors are adapted from the mixtures of multiple sources, e.g., noisy speech, such that the resulting basis vectors may still exhibit features of different sources. Consequently, adapting the complete set of basis vectors may limit the enhanced speech quality.

Another main problem in the NMF-based framework is that the basis vectors of the different signal sources may share similar characteristics. For example, the basis vectors of the speech spectrum can represent the noise spectrum and hence, the enhanced speech may contain noise components that have similar features to the speech. One possible remedy to this problem is to train the basis vectors of each source in a way that prevents them from representing each other [57–59]. However, in most algorithms aiming at training distinct basis vectors, the latter are derived based on heuristic rules which do not guarantee the convergence of the NMF in general [38,60]. Moreover, the distinct basis vectors are obtained indirectly by means of the activation matrix estimated from the mixed training data, which are generated by adding or concatenating the source signal samples. Hence, they lack an explicit interpretation or characterization in terms of their discriminating ability.

Numerous NMF-based multi-channel speech enhancement algorithms have been also introduced. The authors in [33] developed both the multiplicative update (MU) and expectation-maximization (EM) algorithms for estimating the NMF parameters, based on the Itakura-Saito (IS) divergence. To better exploit the spatial properties of the sources, the authors in [61] aimed at factorizing the spatial covariance matrix (SCM) of the observation in each frequency bin, which is specified by the channel covariance matrices of the individual sources. An extended SCM, formulated as a weighted superposition of multiple direction-of-arrival (DoA) kernels (i.e., differential steering matrices), was proposed in [62]. A joint localization and enhancement method, based on the probabilistic steered response power (SRP) model, was presented in [35]. Besides the need to improve the enhancement performance, computational complexity remains one of the main issues when implementing the multi-channel NMF algorithms. That is, the computational cost increases rapidly as the number of NMF basis vectors, the number of microphones or the dimension of the search grid for the speaker location increase.

1.4 Research Objectives and Contributions

Considering the above limitations of existing NMF-based algorithms for speech enhancement, the main objectives of this thesis were formulated as follows:

1. To exploit the statistical properties of the clean speech and noise signals in order to further improve the perceptual quality of the enhanced speech signal
2. To reduce the residual noise components that have features similar to the speech signals and to better handle a mismatch between the characteristics of the training and test data
3. To improve the performance of the NMF algorithm in the multi-channel speech enhancement task

The main contributions of this thesis toward the above objectives are summarized below.

Regarding the first objective, the log-likelihood functions (LLF) of the magnitude spectra for both the clean speech and noise, based on the GMM, are included as regularization terms in the NMF cost function. By using this proposed regularization as *a priori* information, we can exploit the statistical properties of both the clean speech and noise signals during the enhancement stage. For further improvement of the enhanced speech quality, we employ a weighted Wiener filter (WWF) by incorporating the masking effects of the human auditory system. Specifically, we select the weighting factor in the WWF based on the auditory masking threshold.

Towards the second objective, we consider the probabilistic generative model (PGM) of classification, specified by class-conditional densities, along with the Poisson-distributed PGM of NMF. During the training stage, the basis vectors of different sources are trained by constraining them to belong to different classes, where we use the PGM of classification as an *a priori* distribution for the basis vectors. The NMF and PGM parameters of classification are jointly obtained by using the variational Bayesian expectation-maximization (VBEM) algorithm, which guarantees convergence to a stationary point. During the enhancement stage, to better handle a mismatch between the training and test data, extra free basis vectors are employed to capture the features which are not included in the training data.

The last objective is attained by extending the Bayesian NMF model to a multi-channel framework. An important advantage of the proposed framework is its efficiency in estimating the NMF parameters via the VBEM algorithm, which is facilitated by using the Poisson-distributed PGM of NMF. During the enhancement stage, the clean speech point source signal is estimated via the NMF-based MVDR beamforming technique, whose realization involves two main steps. First, the speech source location is determined by observing the spatial output power of the DS beamformer applied to the NMF-based pre-processed noisy speech signal. Second, the noise correlation matrix is computed using the NMF parameters for the magnitude components, and a combination of the noisy speech phase and steering vector for the phase components.

These contributions have led to publications in peer-reviewed journals and refereed conferences, as listed below:

Journal papers

- H. Chung, R. Badeau, E. Plourde and B. Champagne, “Training and compensation of class-conditioned NMF bases for speech enhancement,” *Neurocomputing*, vol. 284, pp. 107-118, Apr. 2018.

- H. Chung, E. Plourde and B. Champagne, “Regularized non-negative matrix factorization with Gaussian mixtures and masking model for speech enhancement,” *Speech*

Communication, vol. 87, pp. 18-30, Mar. 2017.

- H. Chung, E. Plourde and B. Champagne, "Discriminative training of NMF model based on class probabilities for speech enhancement," *IEEE Signal Processing Letters,* vol. 23, no. 4, pp. 502-506, Feb. 2016.

Conference papers

- H. Chung, E. Plourde and B. Champagne, "Single-channel enhancement of convolutive noisy speech based on a discriminative NMF algorithm," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2302-2306, Mar. 2017.
- H. Chung, E. Plourde and B. Champagne, "Basis compensation in non-negative matrix factorization model for speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2249-2253, Mar. 2016.
- H. Chung, E. Plourde and B. Champagne, "Regularized NMF-based speech enhancement with spectral components modeled by Gaussian mixtures," in *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, six pages, Sep. 2014.

Regarding the contributions of the authors in all papers above, the first author, Mr. Hanwook Chung, developed the idea, derived and implemented the algorithms, conducted the experiments and wrote a first draft of the manuscripts. The co-authors, Professor Eric Plourde and Benoit Champagne, supervised the work by providing guidance, validating theoretical development, and contributing to the editing and writing of the final manuscript. In the journal paper published in Neurocomputing, the second author, Professor Roland Badeau, provided useful inputs in the mathematical derivation and in the preparation of the final manuscript.

1.5 Thesis Organization

The thesis is organized as follows. In Chapter 2, we review the basic principles of the NMF and its application to single and multi-channel speech enhancement. Regularized NMF algorithm with Gaussian mixtures and masking model is presented in Chapter 3. In Chapter 4, we explain a training and compensation algorithm of the class-conditioned NMF bases. The extension of the Bayesian NMF model to multi-channel application is presented in Chapter 5. Conclusion and future works are discussed in Chapter 6.

Throughout the thesis, we use the subscripts or superscripts Y, S and N to indicate the noisy speech, clean speech and noise, respectively. The superscripts T, H and $*$ respectively denote matrix transpose, Hermitian transpose and complex conjugate operation. We use the bold upper case letter to denote matrices, e.g., $\mathbf{W}$, and bold lower case letter for the column vectors, e.g., $\mathbf{w}$. The symbols $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{C}$ denote the sets of real numbers, non-negative real numbers and complex numbers, respectively. The symbol $!$ denotes factorial, $*$ indicates the convolution operation and $||\cdot||_2$ denotes the l_2-norm. The imaginary unit is expressed by $\jmath = \sqrt{-1}$, while $\angle Y$ represents the phase of a complex number Y.

Chapter 2

NMF-based Speech Enhancement

In this chapter, we introduce the fundamental concepts at the basis of the NMF model, with special emphasis on the derivation of the update rules from different points of views. The application of the NMF framework to supervised single and multi-channel speech enhancement is described subsequently.

2.1 Background on NMF

For a given matrix $\mathbf{V} = [v_{kl}] \in \mathbb{R}_+^{K\times L}$, NMF finds a local[1] decomposition $\mathbf{V} \approx \mathbf{W}\,\mathbf{H}$, where $\mathbf{W} = [w_{km}] \in \mathbb{R}_+^{K\times M}$ is a basis matrix, $\mathbf{H} = [h_{ml}] \in \mathbb{R}_+^{M\times L}$ is an activation matrix, and M is the number of basis vectors, typically chosen such that $M < \min(K, L)$ [36, 37, 60]. In contrast to other methods such as principal components analysis (PCA), independent component analysis (ICA) and vector quantization (VQ) which train holistic features, the NMF framework allows only additive and not subtractive combinations of the basis vectors. Therefore, it is shown to be useful and effective to train localized features which correspond to the so-called parts-based representation [36].

[1] The term “local” refers to a local minimum of the NMF cost function.

The factorization is obtained by minimizing a suitable cost function $\mathcal{D}(\mathbf{V}, \mathbf{W}\,\mathbf{H})$, such as the Euclidean (EUC) distance [37], the Kullback-Leibler (KL) divergence [37] or the Itakura-Saito (IS) divergence [38], respectively:

$$\mathcal{D}_{EUC}(\mathbf{V}, \mathbf{W}\,\mathbf{H}) = \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{L}\left(v_{kl} - [\mathbf{W}\,\mathbf{H}]_{kl}\right)^2 \tag{2.1}$$

$$\mathcal{D}_{KL}(\mathbf{V}, \mathbf{W}\,\mathbf{H}) = \sum_{k=1}^{K}\sum_{l=1}^{L}\left(v_{kl}\ln\frac{v_{kl}}{[\mathbf{W}\,\mathbf{H}]_{kl}} - v_{kl} + [\mathbf{W}\,\mathbf{H}]_{kl}\right) \tag{2.2}$$

$$\mathcal{D}_{IS}(\mathbf{V}, \mathbf{W}\,\mathbf{H}) = \sum_{k=1}^{K}\sum_{l=1}^{L}\left(\frac{v_{kl}}{[\mathbf{W}\,\mathbf{H}]_{kl}} - \ln\frac{v_{kl}}{[\mathbf{W}\,\mathbf{H}]_{kl}} - 1\right) \tag{2.3}$$

where $[\cdot]_{kl}$ denotes the (k,l)-th entry of its matrix argument. The NMF solutions can be found iteratively using the corresponding multiplicative update (MU) rules [37, 63]:

$$\text{EUC}: \mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\mathbf{V}\,\mathbf{H}^T}{\mathbf{W}\,\mathbf{H}\,\mathbf{H}^T}, \qquad \mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T\,\mathbf{V}}{\mathbf{W}^T\,\mathbf{W}\,\mathbf{H}} \tag{2.4}$$

$$\text{KL}: \mathbf{W} \leftarrow \mathbf{W} \otimes \frac{(\mathbf{V}\,/(\mathbf{W}\,\mathbf{H}))\,\mathbf{H}^T}{\mathbf{1}_{KL}\,\mathbf{H}^T}, \quad \mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T(\mathbf{V}\,/(\mathbf{W}\,\mathbf{H}))}{\mathbf{W}^T\,\mathbf{1}_{KL}} \tag{2.5}$$

$$\text{IS}: \mathbf{W} \leftarrow \mathbf{W} \otimes \left(\frac{(\mathbf{V}\,/(\mathbf{W}\,\mathbf{H})^2)\,\mathbf{H}^T}{(\mathbf{W}\,\mathbf{H})^{-1}\,\mathbf{H}^T}\right)^{1/2}, \quad \mathbf{H} \leftarrow \mathbf{H} \otimes \left(\frac{\mathbf{W}^T(\mathbf{V}\,/(\mathbf{W}\,\mathbf{H})^2)}{\mathbf{W}^T(\mathbf{W}\,\mathbf{H})^{-1}}\right)^{1/2} \tag{2.6}$$

where the operation $\otimes$ denotes element-wise multiplication, the quotient line and / are element-wise division, $\mathbf{1}_{KL}$ is a $K \times L$ matrix with all entries equal to one, $\leftarrow$ refers to an iterative overwrite, and the exponents in (2.6) are computed element-wisely. The scale indeterminacies in $\mathbf{W}$ and $\mathbf{H}$, which appear as a product in $\mathbf{V}$, can be prevented by normalizing $\mathbf{W}$ using the l_1 or l_2-norm after estimating $\mathbf{W}$, and subsequently compute $\mathbf{H}$, for each iteration [64].

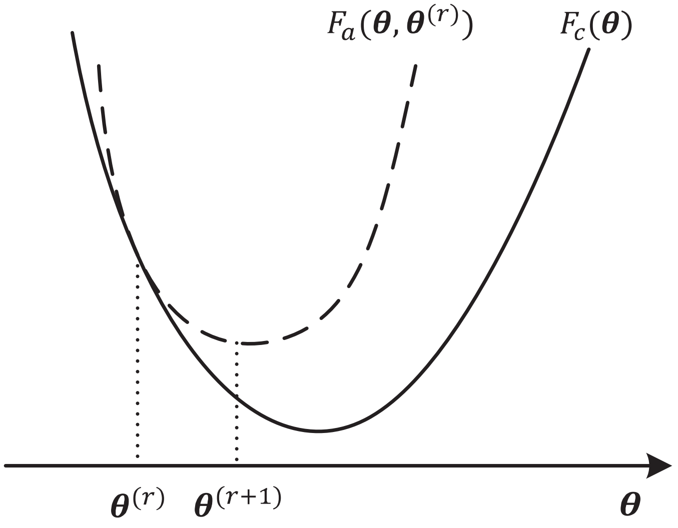
$F_a(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)})$
$F_c(\boldsymbol{\theta})$
$\boldsymbol{\theta}^{(r)}$
$\boldsymbol{\theta}^{(r+1)}$
$\boldsymbol{\theta}$

Fig. 2.1 Graphical illustration of the concept of the auxiliary function.

The MU rules in (2.4)-(2.6) are derived based on the concept of using the auxiliary function [37, 63]. Specifically, let $F_c(\theta) \geq 0$ denote a non-negative cost function to be minimized with respect to a multivariate parameter $\boldsymbol{\theta}$. A function $F_a(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ which satisfies

$$F_a(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \geq F_c(\boldsymbol{\theta}), \qquad F_a(\boldsymbol{\theta}, \boldsymbol{\theta}) = F_c(\boldsymbol{\theta}) \tag{2.7}$$

is called an auxiliary function for $F_c(\boldsymbol{\theta})$, where $\tilde{\boldsymbol{\theta}}$ is an auxiliary variable. It is obvious that $F_c(\boldsymbol{\theta})$ is non-increasing under the following iterative update

$$\boldsymbol{\theta}^{(r+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ F_a(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)}) \tag{2.8}$$

since $F_c(\boldsymbol{\theta}^{(r+1)}) = F_a(\boldsymbol{\theta}^{(r+1)}, \boldsymbol{\theta}^{(r+1)}) \leq F_a(\boldsymbol{\theta}^{(r+1)}, \boldsymbol{\theta}^{(r)}) \leq F_a(\boldsymbol{\theta}^{(r)}, \boldsymbol{\theta}^{(r)}) = F_c(\boldsymbol{\theta}^{(r)})$, where the superscript (r) denotes the r-th iteration. Hence, the iterative update of $\boldsymbol{\theta}$ guarantees the convergence to a stationary point of $F_c(\boldsymbol{\theta})$. The concept of using the auxiliary function is graphically illustrated in Figure 2.1.

For example, the application of the auxiliary function approach to the NMF problem with the KL-divergence is summarized as follows (see [37] for a more detailed discussion). Let the cost function $F_c(\boldsymbol{\theta})$ denote the KL-divergence given by (2.2) with a fixed activation $\mathbf{H}^{(r)}$. We can construct the auxiliary function for $F_c(\boldsymbol{\theta})$ as

$$F_a(\mathbf{W}, \mathbf{W}^{(r)}) = \sum_{k=1}^{K} \sum_{l=1}^{L} \left[v_{kl} \ln v_{kl} - v_{kl} \sum_{m=1}^{M} w_{km} (h_{ml})^{(r)} - v_{kl} + \sum_{m=1}^{M} \gamma_{kl}^{m} \ln \frac{w_{km} (h_{ml})^{(r)}}{\gamma_{kl}^{m}} \right] \tag{2.9}$$

where γ_{kl}^{m} is given by

$$\gamma_{kl}^{m} = \frac{(w_{km})^{(r)} (h_{ml})^{(r)}}{\sum_{m'} (w_{km'})^{(r)} (h_{m'l})^{(r)}}. \tag{2.10}$$

It is straightforward to verify that $F_a(\mathbf{W}, \mathbf{W}) = F_c(\mathbf{W})$. The inequality $F_a(\mathbf{W}, \mathbf{W}^{(r)}) \geq F_c(\mathbf{W})$ can be shown by using Jensen's inequality based on the convexity of the log function as

$$-\ln \sum_{m=1}^{M} w_{km} (h_{ml})^{(r)} \leq -\ln \sum_{m=1}^{M} \gamma_{kl}^{m} \ln \frac{w_{km} (h_{ml})^{(r)}}{\gamma_{kl}^{m}} \tag{2.11}$$

which holds for a non-negative value γ_{kl}^{m} such that $\sum_m \gamma_{kl}^{m}$. By setting the partial derivative of (2.9) with respect to w_{km} to zero, the basis element is found to be

$$(w_{km})^{(r+1)} = \frac{\sum_{l=1}^{L} v_{kl} \gamma_{kl}^{m}}{\sum_{l=1}^{L} (h_{ml})^{(r)}} = \frac{(w_{km})^{(r)}}{\sum_{l=1}^{L} (h_{ml})^{(r)}} \sum_{l=1}^{L} \frac{v_{kl} (h_{ml})^{(r)}}{\sum_{m'}^{M} (w_{km'})^{(r)} (h_{m'l})^{(r)}}. \tag{2.12}$$

The basis element update in (2.12) can be rearranged in a matrix form, which leads to (2.5). The update rule of the activation matrix can be derived by following a similar approach.

There are two main alternative points of views of deriving the update rules for the basis and activation matrices. The first one is interpreting the NMF model within a statistical framework, and the second one is based on a heuristic observation. These are explained below.

1) Statistical interpretation of NMF: The NMF problem also can be interpreted within a statistical framework. That is, it has been shown that the cost functions in (2.1)-(2.3) have corresponding PGMs. For a given matrix $\mathbf{X} = [x_{kl}]$, each entry is assumed to be a sum of M latent variables as

$$x_{kl} = \sum_{m=1}^{M} c_{kl}^{m}. \tag{2.13}$$

where x_{kl} is a non-negative real value for the PGMs corresponding to either the Euclidean distance or the KL-divergence, and a complex value for the PGM corresponding to the IS-divergence. The m-th latent variable c_{kl}^{m} is assumed to be drawn from one of the following distributions [38,41,65]:

$$c_{kl}^{m} \sim p(c_{kl}^{m}|w_{km}, h_{ml}) = \begin{cases} \mathcal{N}(c_{kl}^{m}|w_{km}h_{ml}, 1) & : \text{EUC} \\ \mathcal{P}(c_{kl}^{m}|w_{km}h_{ml}) & : \text{KL} \\ \mathcal{N}_c(c_{kl}^{m}|0, w_{km}h_{ml}) & : \text{IS} \end{cases} \tag{2.14}$$

where $\mathcal{N}(c|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp((c-\mu)^2/(2\sigma^2))$ is the univariate Gaussian distribution with mean μ and variance σ^2, $\mathcal{P}(c|u) = u^c \exp(-u)/(c!)$ is the Poisson distribution with mean u and $\mathcal{N}_c(c|\mu, \sigma^2) = (\pi\sigma^2)^{-1} \exp(-|c-\mu|^2/\sigma^2)$ is the complex-valued univariate Gaussian distribution with $\mu \in \mathbb{C}$ (complex-valued) and variance σ^2. Assuming that the random variables x_{kl} are drawn independently, the logarithm of the distribution of $\mathbf{X}$ is obtained, for each case, as

$$\ln p(\mathbf{X} \,|\, \mathbf{W}, \mathbf{H}) \;=\; \ln \prod_{k=1}^{K} \prod_{l=1}^{L} p(x_{kl}|w_{kl}, h_{ml}) \tag{2.15}$$

$$\overset{c}{=} \begin{cases} \sum_{k=1}^{K}\sum_{l=1}^{L}\left[-\frac{1}{2}\left(x_{kl}-\sum_{m=1}^{M} w_{km}h_{ml}\right)^{2}\right] & : \text{EUC} \\ \sum_{k=1}^{K}\sum_{l=1}^{L}\left[x_{kl}\ln\left(\sum_{m=1}^{M} w_{km}h_{ml}\right)-\sum_{m=1}^{M} w_{km}h_{ml}\right] & : \text{KL} \\ \sum_{k=1}^{K}\sum_{l=1}^{L}\left[-\ln\left(\sum_{m=1}^{M} w_{km}h_{ml}\right)-\frac{|x_{kl}|^{2}}{\sum_{m=1}^{M} w_{km}h_{ml}}\right] & : \text{IS} \end{cases}$$

where $\overset{c}{=}$ denotes equality up to a constant term. By adjusting the notations as $\mathbf{V}=\mathbf{X}$ for the Euclidean distance and the KL-divergence and $\mathbf{V}=[|x_{kl}|^2]$ for the IS-divergence, we can see that the maximization of the LLFs given by (2.15) with respect to w_{km} and h_{ml} are equivalent to the minimization of the cost functions in (2.1)-(2.3), respectively.

The ML estimates of the parameters w_{km} and h_{ml}, given the observations v_{kl}, are obtained via the iterative EM algorithm [66,67]. During the expectation step (E-step), the posterior distribution of the latent variable given the observation is calculated. During the maximization step (M-step), the parameters are estimated by maximizing the expectation of the complete-data LLF with respect to the posterior distribution. For example, when considering the Poisson-distributed PGM of NMF in (2.14), which corresponds to the KL-divergence, the application of the EM algorithm is summarized as follows. During the E-step, the posterior distribution $p(\mathbf{c}_{kl}^{m}|v_{kl})$ where $\mathbf{c}_{kl}=[c_{kl}^{1},...,c_{kl}^{M}]$ is computed, which is shown to be a multinomial distribution [41]:

$$\mathcal{M}(\mathbf{c}_{kl};v_{kl},\bar{\mathbf{p}}_{kl})=\delta\left(v_{kl}-\sum_{m=1}^{M}c_{kl}^{m}\right)v_{kl}!\prod_{m=1}^{M}\frac{(\bar{p}_{kl}^{m})^{c_{kl}^{m}}}{c_{kl}^{m}!} \tag{2.16}$$

where $\delta(x)$ is the Kronecker delta function defined by $\delta(x)=1$ for $x=0$ and $\delta(x)=0$ otherwise. The entries of $\bar{\mathbf{p}}_{kl}=[\bar{p}_{kl}^{m}]$ (also referred to as cell probabilities such that

$\sum_m \bar{p}_{kl}^m = 1$) are given by

$$\bar{p}_{kl}^m = \frac{w_{km} h_{ml}}{\sum_{m'} w_{km'} h_{m'l}}. \tag{2.17}$$

During the M-step, the basis and activation elements are estimated by maximizing

$$\mathcal{L}_C(\mathbf{V} \,|\, \mathbf{W}, \mathbf{H}) \overset{c}{=} \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{m=1}^{M} (-w_{km} h_{ml} + \bar{c}_{kl}^m \ln(w_{km} h_{ml})) \tag{2.18}$$

where $\bar{c}_{kl}^m$ is the conditional expectation of the latent variable c_{kl}^m with respect to the posterior distribtion $p(c_{kl}^m | v_{kl})$, i.e., the mean value of the multinomial distribution in (2.16), given by

$$\bar{c}_{kl}^m = \mathbb{E}[c_{kl}^m | v_{kl}] = \bar{p}_{kl}^m v_{kl} = \frac{w_{km} h_{ml}}{\sum_{m'=1}^{M} w_{km'} h_{m'l}} v_{kl}. \tag{2.19}$$

The iterative NMF solutions obtained through the EM algorithm are shown as [41]

$$(w_{km})^{(r+1)} = \frac{\sum_{l=1}^{L} (\bar{c}_{kl}^m)^{(r)}}{\sum_{l=1}^{L} (h_{ml})^{(r)}} \tag{2.20}$$

$$(h_{ml})^{(r+1)} = \frac{\sum_{k=1}^{K} (\bar{c}_{kl}^m)^{(r)}}{\sum_{k=1}^{K} (w_{km})^{(r+1)}} \tag{2.21}$$

where again, the superscript (r) refers to the r-th iteration.

2) Heuristic MU rules: The NMF solutions can be found by using the so-called heuristic MU rules, which can be considered as a generalized version of the MU rules given by (2.4)-(2.6). To this end, the gradient of the cost function is expressed as the difference of two non-negative terms such that $\nabla \mathcal{D}(\mathbf{V}, \mathbf{W}\,\mathbf{H}) = \nabla^{+} \mathcal{D}(\mathbf{V}, \mathbf{W}\,\mathbf{H}) - \nabla^{-} \mathcal{D}(\mathbf{V}, \mathbf{W}\,\mathbf{H})$, i.e., where $\nabla^{+} \mathcal{D}(\mathbf{V}, \mathbf{W}\,\mathbf{H}) \geq 0$ and $\nabla^{-} \mathcal{D}(\mathbf{V}, \mathbf{W}\,\mathbf{H}) \geq 0$. By taking advantage of this representation, the heuristic MU rules are shown to be [38]:

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\nabla^{-}_{\mathbf{W}} \mathcal{D}(\mathbf{V}, \mathbf{W}\,\mathbf{H})}{\nabla^{+}_{\mathbf{W}} \mathcal{D}(\mathbf{V}, \mathbf{W}\,\mathbf{H})}, \quad \mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\nabla^{-}_{\mathbf{H}} \mathcal{D}(\mathbf{V}, \mathbf{W}\,\mathbf{H})}{\nabla^{+}_{\mathbf{H}} \mathcal{D}(\mathbf{V}, \mathbf{W}\,\mathbf{H})}. \tag{2.22}$$

In general, the heuristic MU rules do not guarantee the convergence to a stationary point [38]. Nevertheless, they are widely used due to the simplicity of their derivation and implementation, especially in diverse regularized algorithms, e.g., [32, 40, 68]. To this end, a regularized cost function can be written as

$$\mathcal{J}(\mathbf{W}, \mathbf{H}) = \mathcal{D}(\mathbf{V}, \mathbf{W}\,\mathbf{H}) + \lambda\,\mathcal{R}(\mathbf{W}, \mathbf{H}) \tag{2.23}$$

where $\lambda > 0$ is a regularization coefficient and $\mathcal{R}(\mathbf{W}, \mathbf{H})$ is a regularization term. The convergence behavior as well as the performance of the target application generally depends on the regularization coefficient.

The concept of NMF introduced so far can be extended to factorizing a given tensor $\mathbf{V} = [v^{j}_{kl}] \in \mathbb{R}^{K \times L \times J}_{+}$. Representative methods include the multi-channel NMF (MNMF) algorithm [33] and the non-negative tensor factorization (NTF) algorithm [69]. The MNMF and NTF models are respectively given by

$$v^{j}_{kl} \approx \hat{v}^{j}_{kl} = \begin{cases} a^{j}_{k} \sum_{m=1}^{M} w_{km} h_{ml} & : \text{MNMF} \\ \sum_{m=1}^{M} a^{j}_{m} w_{km} h_{ml} & : \text{NTF} \end{cases} \tag{2.24}$$

where a^{j}_{k} and a^{j}_{m} are the mixing coefficients in the MNMF and NTF models, respectively. The parameters of the MNMF model based on the IS-divergence can be estimated via the following MU rules[2] [33]:

$$a_k^j \leftarrow a_k^j \frac{\sum_{l=1}^{L} \left[(\hat{v}_{kl}^j)^{-2} v_{kl}^j \sum_{m=1}^{M} w_{km} h_{ml} \right]}{\sum_{l=1}^{L} \left[(\hat{v}_{kl}^j)^{-1} \sum_{m=1}^{M} w_{km} h_{ml} \right]} \tag{2.25}$$

$$w_{km} \leftarrow w_{km} \frac{\sum_{j=1}^{J} a_k^j \left[\sum_{l=1}^{L} (\hat{v}_{kl}^j)^{-2} v_{kl}^j h_{ml} \right]}{\sum_{j=1}^{J} a_k^j \left[\sum_{l=1}^{L} (\hat{v}_{kl}^j)^{-1} h_{ml} \right]} \tag{2.26}$$

$$h_{ml} \leftarrow h_{ml} \frac{\sum_{j=1}^{J} \sum_{k=1}^{K} a_k^j w_{km} (\hat{v}_{kl}^j)^{-2} v_{kl}^j}{\sum_{j=1}^{J} \sum_{k=1}^{K} a_k^j w_{km} (\hat{v}_{kl}^j)^{-1}}. \tag{2.27}$$

The MNMF and NTF algorithms, aiming at factorizing a given tensor, can be applied to multi-channel speech enhancement problem.

2.2 Application of NMF to Speech Enhancement

We next briefly introduce a general framework for the application of the supervised NMF-based algorithms to the problems of the single and multi-channel speech enhancement.

2.2.1 Single-channel application

The frequency-domain representation is commonly used for audio and speech signal processing to better exploit spectral characteristics. To this end, a popular choice is the short-time Fourier transform (STFT), which consists of separate discrete Fourier transforms (DFT) applied to the signal under consideration over successive time windows or frames. Let $y[i]$, where index $i = 0, 1, 2, ...$, denote the discrete-time samples of a noisy speech signal, as obtained by passing a microphone output signal through an analog-to-digital converter (ADC)

[2] See [33] for the EM-based parameter estimation of the MNMF model, and [69] for the MU-based parameter estimation of the NTF model.

with sampling rate of f_s [Hz][3]. The STFT coefficients of the noisy speech are computed as [70, 71]

$$Y_{kl} = \sum_{i=0}^{L_w-1} y[i+(l-1)L_h]w_a[i]e^{-\jmath\frac{2\pi}{F}(k-1)i} \tag{2.28}$$

where $w_a[\imath]$ is an analysis window of length L_w defined in the interval of $0 \leq i \leq L_w - 1$, L_h is the frame advance (also referred to as hop size), F is the DFT size[4], $k \in \{1, ..., K\}$ is the frequency bin index and $l \in \{1, ..., L\}$ is the time frame index. Due to the conjugate symmetry of the spectral coefficients, only half of them are considered in general for practical implementation, resulting in $K = F/2 + 1$ for even F and $K = (F+1)/2$ for odd F.

Let us consider an additive noise model, i.e.,

$$y[i] = s[i] + n[i] \tag{2.29}$$

where $s[i]$ and $n[i]$ respectively denote the clean speech and noise signals in the discrete-time domain. Under this assumption, the noisy speech spectrum obtained via STFT is given by the sum of the clean speech and noise spectra , i.e.,

$$Y_{kl} = S_{kl} + N_{kl} \tag{2.30}$$

where Y_{kl}, S_{kl} and N_{kl} respectively denote the STFT coefficients of the noisy speech, clean speech and noise at the k-th frequency bin and l-th time frame.

Different choices of NMF cost functions have been presented in Section 2.1. The KL-

[3] In practice, f_s will range from 8 kHz (toll quality speech) to 44.1 kHz (high-quality audio).

[4] The DFT size F is often chosen to be larger than the window length L_w (equivalent to using the well-known zero-padding method), to limit temporal aliasing. The latter may occur when converting the enhanced speech spectrum, obtained by filtering the noisy speech spectrum, into the time-domain [75]. Throughout the thesis, however, we simply use $F = L_w$ by assuming that such effect is negligible.

based NMF algorithms when applied to speech enhancement are known to be better suited for handling the magnitude spectral coefficients. Hence, one commonly assumes that the magnitude spectrum of the noisy speech can be approximated by the sum of the clean speech and noise magnitude spectra [32, 34, 40, 72], i.e., $|Y_{kl}| \approx |S_{kl}| + |N_{kl}|$. In contrast, the IS-based NMF algorithms are usually applied to the power spectral coefficients [38, 40], i.e., $|Y_{kl}|^2 \approx |S_{kl}|^2 + |N_{kl}|^2$. In single-channel source separation and speech enhancement applications, the KL-based approach is a more popular choice and widely used since it has been shown to provide better performance compared to using other measures, such as the Euclidean distance or the IS-divergence [73, 74]. Hence, we focus on the KL-divergence as the main cost function in this thesis.

A supervised NMF-based speech enhancement framework consists of two stages. During the training stage, the basis matrices of the clean speech and noise, $\mathbf{W}_S = [w_{km}^S] \in \mathbb{R}_+^{K \times M_S}$ and $\mathbf{W}_N = [w_{km}^N] \in \mathbb{R}_+^{K \times M_N}$ respectively, are obtained by applying the NMF update rules to the corresponding training data separately. During the enhancement stage, by fixing $\mathbf{W}_Y = [\mathbf{W}_S\, \mathbf{W}_N] \in \mathbb{R}_+^{L \times (M_S + M_N)}$, the activation matrix of the noisy speech $\mathbf{H}_Y = [\mathbf{H}_S^T \;\; \mathbf{H}_N^T]^T \in \mathbb{R}_+^{(M_S+M_N) \times L}$ is estimated[5] by applying the activation update to $\mathbf{V}_Y = [|Y_{kl}|] \in \mathbb{R}_+^{K \times L}$. Once the activation matrix is computed, the clean speech spectrum can be estimated from the noisy speech spectrum via Wiener filtering as [1]

$$\hat{S}_{kl} = \frac{\hat{p}_{kl}^S}{\hat{p}_{kl}^S + \hat{p}_{kl}^N} Y_{kl} \tag{2.31}$$

[5] Besides this so-called batch approach, we can consider alternative implementation: *online* or *mini-batch online* approach. In the former case, the activation vector $\mathbf{h}_l^Y = [(\mathbf{h}_l^S)^T \;\; (\mathbf{h}_l^N)^T]^T \in \mathbb{R}_+^{(M_S+M_N)}$ is estimated from the instantaneous target vector $\mathbf{V}_l^Y = [|Y_{kl}|] \in \mathbb{R}_+^K$ for the l-th time frame. In the latter case, the activation matrix $\mathbf{H}_{l_b}^Y = [(\mathbf{H}_{l_b}^S)^T \;\; (\mathbf{H}_{l_b}^N)^T]^T \in \mathbb{R}_+^{(M_S+M_N) \times L_b}$ is obtained from a target matrix, $\mathbf{V}_{l_b}^Y = [|Y_{kl}|] \in \mathbb{R}_+^{K \times L_b}$ where $l_b = 1, 2, ...$ is the mini-batch index and L_b is the mini-batch size, obtained from consecutive time frames $l \in \{(l_b - 1)L_b + 1, ..., l_b L_b\}$.

where $\hat{p}_{kl}^S$ and $\hat{p}_{kl}^N$ respectively denote the estimated power spectral densities (PSD) of the clean speech and noise. The latter are typically obtained via temporal smoothing of the NMF-based periodograms as [56]

$$\hat{p}_{kl}^S = \tau_S \hat{p}_{k,l-1}^S + (1-\tau_S)\left(\sum_{m=1}^{M_S} w_{km}^S h_{ml}^S\right)^2 \tag{2.32}$$

$$\hat{p}_{kl}^N = \tau_N \hat{p}_{k,l-1}^N + (1-\tau_N)\left(\sum_{m=1}^{M_N} w_{km}^N h_{ml}^N\right)^2 \tag{2.33}$$

where τ_S and $\tau_N \in (0,1)$ are the smoothing factors for the clean speech and noise.

Finally, the time-domain enhanced speech signal is obtained by applying the inverse DFT (IDFT) to the enhanced speech spectrum for each time frame, followed by the overlap-add method [71]. Specifically, the IDFT values for the l-th time frame are obtained as

$$\tilde{s}_l[i] = \begin{cases} \dfrac{1}{F} \sum_{k=1}^{F} \hat{S}_{kl} e^{\jmath \frac{2\pi}{F}(k-1)i}, & i = 0, ..., F-1 \\ 0, & \text{else} \end{cases} . \tag{2.34}$$

Samples from successive frames are reassembled into a single time sequence via the overlap-add method:

$$\hat{s}[i] = \sum_{l=1}^{\infty} \tilde{s}_l[i-(l-1)L_h] w_s[i-(l-1)L_h] \tag{2.35}$$

where $w_s[i]$ is a synthesis window defined in the interval of $0 \leq i \leq L_w - 1$. For perfect reconstruction, the analysis and synthesis windows should satisfy [76]:

$$\sum_{l=1}^{L_w} w_a[i-(l-1)L_h] w_s[i-(l-1)L_h] = 1 \tag{2.36}$$

for all $i \geq 0$. In practice, we can consider a rectangular synthesis window $w_s[i] = 1/W_a$, where W_a is a constant value such that $W_a = \sum_l w_a[i-(l-1)L_h]$ for all $i \geq 0$. This latter condition can be satisfied by selecting an appropriate analysis window[6], e.g., a sine or Hanning window.

2.2.2 Multi-channel application

We can extend the single-channel application discussed in the previous subsection to a mutli-channel application, which utilizes a multi-channel noisy speech signal obtained through a microphone array. The latter consists of a fixed spatial arrangement of J microphone elements, indexed with $j \in \{1, ..., J\}$ where the outputs are sampled at a sampling rate f_s and fed to a multi-channel processor. Let $y^j[i]$ denote the discrete-time samples of a noisy speech signal recorded at the j-th microphone. By taking into account the convolutive nature of the acoustic medium, as represented by an acoustic impulse response between the speech source and the microphone (also known as a mixing filter), the noisy speech signal can be written in the time-domain as

$$\begin{aligned} y^j[i] &= z_S^j[i] + z_N^j[i] \\ &= \tilde{a}_S^j[i] * s[i] + z_N^j[i] \end{aligned} \tag{2.37}$$

where $z_S^j[i]$ is the so-called clean speech image source component[7], $z_N^j[i]$ is the additive noise components, $s[i]$ is the clean speech point source signal and $\tilde{a}_S^j[i]$ is the mixing filter for the clean speech.

Assuming that the mixing filter length is shorter than the STFT analysis window length,

[6] In practice, due to the shape of $w_a[i]$, this condition cannot be satisfied for several initial values of the time index i, i.e., for $i = 0, 1, ..., L_h - 1$. This can be simply handled by padding zero values prior to the signal before implementing STFT, and discarding them after reconstruction.

[7] The term "image" refers to the convolution operation in (2.38), which can be interpreted as a sum of scaled and delayed samples of $s[i]$.

the multi-channel convolutive noisy speech signal $y^j[i]$ in (2.38) can be expressed in the STFT domain as [33, 62, 69]

$$Y^j_{kl} = Z^j_{S,kl} + Z^j_{N,kl} = \tilde{A}^j_{S,k} S_{kl} + Z^j_{N,kl} \tag{2.38}$$

where $Z^j_{S,kl}$ and $Z^j_{N,kl}$ respectively denote the STFT coefficients of the convolutive clean speech and noise signals, $\tilde{A}^j_{S,k}$ is the acoustic transfer function (ATF) for the clean speech (obtained from the DFT coefficients of the mixing filter $\tilde{a}^j_S[i]$), S_{kl} is the STFT coefficient of the clean speech point source signal, and $k \in \{1, ..., K\}$, $l \in \{1, ..., L\}$ and $j \in \{1, ..., J\}$ are the frequency bin, time frame and microphone indices.

It is worth noting that we assume that the noise spectrum can be also expressed in terms of an ATF and point source spectrum as $Z^j_{N,kl} = \tilde{A}^j_{N,k} N_{kl}$, where $\tilde{A}^j_{N,k}$ is the ATF for the noise and N_{kl} is the noise point source spectrum. Although this noise model is theoretically valid only for a noise signal generated by a point source (e.g., when a small fan is placed in a room [17]), it is widely used in NMF-based framework for multi-channel speech enhancement in practice, e.g., [35, 61, 62]. Besides, we can directly consider the noise image source magnitude spectrum $|Z^j_{N,kl}|$, and estimate the basis vectors by applying a single-channel NMF algorithm to each channel to obtain the basis vectors, e.g., [77]. However, such an approach results in a large size of the basis matrix (i.e., proportional to the number of microphones J), which may increase the computational cost during the enhancement stage. Moreover, it can hardly handle the dynamic of the ATF, especially when the microphone configurations used while acquiring the training and test data are different. In contrast, the point source model suggests an efficient representation of the noise signal. That is, we can estimate a single basis matrix for a noise signal as well as we can capture the dynamic characteristic of the acoustic environment explicitly by means of the mixing coefficients. In addition, the point source model enables a possible post-processing of the beamformer output, e.g., the application of NMF-based single-channel Wiener filtering, specified by the basis and activation elements, to the MVDR beamformer output.

We can consider the MNMF or NTF model in (2.24) as an application of a tensor factorization algorithm for multi-channel speech enhancement. However, the NTF model employs frequency-independent mixing coefficients a_m^j. This intrinsically implies that the NTF model is suited for a linear instantaneous mixture signal and hence, the model is inadequate to handle the convolutive effects specified by the ATFs [33]. Therefore, we consider the MNMF model in this thesis. As in the single-channel application, we apply the MNMF algorithm to a given tensor based on the magnitude spectral coefficients, where the parameters in (2.24) can be interpreted as follows. Considering the clean speech image spectrum $Z_{S,kl}^j = \tilde{A}_{S,k}^j S_{kl}$ for instance, by setting the notation as $v_{kl}^{S,j} = |Z_{S,kl}^j|$, the basis and activation elements w_{km} and h_{ml} become related to the point source spectrum (i.e., $|S_{kl}| = \sum_m w_{km} h_{ml}$), while the mixing coefficient corresponds to the magnitude value of the ATF (i.e., $a_k^{S,j} = |\tilde{A}_{S,k}^j|$).

Similar to the single-channel application developed in the previous subsection, a supervised NMF-based multi-channel speech enhancement framework consists of two stages. During the training stage, the basis matrices of the clean speech and noise, $\mathbf{W}_S = [w_{km}^S] \in \mathbb{R}_+^{K \times M_S}$ and $\mathbf{W}_N = [w_{km}^N] \in \mathbb{R}_+^{K \times M_N}$, are obtained from the tensor-based training data. During the enhancement stage, by fixing the basis matrices, we estimate the magnitude values of the ATFs of the clean speech and noise (i.e., $\mathbf{A}_S = \{a_k^{S,j}\}$ and $\mathbf{A}_N = \{a_k^{N,j}\}$) and activation matrix $\mathbf{H}_Y \in \mathbb{R}_+^{(M_S+M_N) \times L}$ from the noisy speech magnitude spectrum[8]

[8] As explained in Footnote 4, we can consider an online or mini-batch online approach, which enables to handle the case of slowly moving sources (i.e., time-varying ATFs).

$\mathbf{V}_Y = [|Y_{kl}^j|] \in \mathbb{R}_+^{K \times L \times J}$.

There are two main reconstruction targets in multi-channel applications, i.e., the image source and point source estimation, which are explained below.

1) Image source estimation: The first one is to estimate the clean speech image source spectral coefficients $Z_{S,kl}^j$, which can be obtained by applying a single-channel Wiener filter to the noisy speech STFT coefficients Y_{kl}^j for each channel as [61, 62, 69]

$$\hat{Z}_{S,kl}^j = \frac{|\hat{A}_{S,k}^j|^2 \hat{p}_{kl}^S}{|\hat{A}_{S,k}^j|^2 \hat{p}_{kl}^S + |\hat{A}_{N,k}^j|^2 \hat{p}_{kl}^N} Y_{kl}^j \tag{2.39}$$

where $\hat{p}_{kl}^S$ and $\hat{p}_{kl}^N$ are given by (2.32) and (2.33)[9], $\hat{A}_{S,k}^j = a_k^{S,j} \varphi_k^{S,j}$ and $\hat{A}_{N,k}^j = a_k^{N,j} \varphi_k^{N,j}$ ($a_k^{S,j} \triangleq |\hat{A}_{S,k}^j|$, $\varphi_k^{S,j} \triangleq \exp(\jmath \angle \hat{A}_{S,k}^j)$, $a_k^{N,j} \triangleq |\hat{A}_{N,k}^j|$ and $\varphi_k^{N,j} \triangleq \exp(\jmath \angle \hat{A}_{N,k}^j)$) are the estimated complex-valued ATFs for the clean speech and noise. The phase-related components $\varphi_k^{S,j}$ and $\varphi_k^{N,j}$ can be obtained based on the noisy speech phase for the l-th time frame, i.e., $\varphi_k^{S,j} = \varphi_k^{N,j} = \exp(\jmath \angle Y_{kl}^j)$.

The clean speech image spectrum can be estimated alternatively via multi-channel Wiener filtering [61]:

$$\hat{\mathbf{Z}}_{S,kl} = \left[\left(\mathbf{R}_{kl}^S + \mathbf{R}_{kl}^N \right)^{-1} \mathbf{R}_{kl}^S \right]^H \mathbf{Y}_{kl} \tag{2.40}$$

where $\hat{\mathbf{Z}}_{S,kl} = [Z_{S,kl}^1, ..., Z_{S,kl}^J]^T$ and $\mathbf{Y}_{kl} = [Y_{kl}^1, ..., Y_{kl}^J]^T$, and $\mathbf{R}_{kl}^S \in \mathbb{C}^{J \times J}$ and $\mathbf{R}_{kl}^N \in \mathbb{C}^{J \times J}$ are the clean speech and noise correlation matrices. The latter are obtained via temporal smoothing, e.g., [78], as

$$[\mathbf{R}_{kl}^S]_{ab} = \tau_S [\mathbf{R}_{k,l-1}^S]_{ab} + (1 - \tau_S) \hat{A}_{S,k}^a \left(\hat{A}_{S,k}^b \right)^* \left(\sum_{m=1}^{M_S} w_{km}^S h_{ml}^S \right)^2 \tag{2.41}$$

[9] Note that $\hat{p}_{kl}^S$ and $\hat{p}_{kl}^N$ are obtained based on the basis and activation elements (i.e., related to the point source signal) and hence, do not depend on the microphone index j.

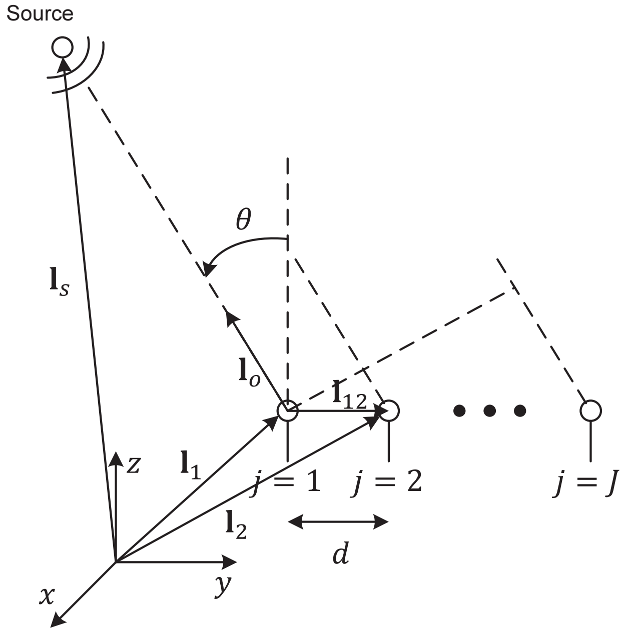

Fig. 2.2 Geometric illustration of the propagation delay between a source and microphone array.

$$[\mathbf{R}^N_{kl}]_{ab} = \tau_N[\mathbf{R}^N_{k,l-1}]_{ab} + (1-\tau_N)\hat{A}^a_{N,k}\left(\hat{A}^b_{N,k}\right)^*\left(\sum_{m=1}^{M_N} w^N_{km}h^N_{ml}\right)^2 \tag{2.42}$$

where τ_S and τ_N $(0 \leq \tau_S, \tau_N \leq 1)$ are the smoothing constants, and $a, b \in \{1, ..., J\}$.

2) Point source estimation: The second target is to estimate the clean speech point source spectral coefficients $\hat{S}_{kl}$, which can be obtained via the MVDR beamforming technique [78], as

$$\hat{S}_{kl} = \left(\frac{(\mathbf{R}^N_{kl})^{-1}\,\mathbf{b}_k}{\mathbf{b}^H_k(\mathbf{R}^N_{kl})^{-1}\,\mathbf{b}_k}\right)^H \mathbf{Y}_{kl} \tag{2.43}$$

where $\mathbf{R}^N_{kl}$ is the noise correlation matrix given by (2.42) and $\mathbf{b}_k = [b^j_k] \in \mathbb{C}^J$ is the steering vector. Specifically, assuming the far field model in which the wavefront of the transmitted sound appears planar when impinging on the microphone array, the steering vector is obtained in terms of the time difference-of-arrival (TDoA). The concept of the TDoA illustrated in Figure 2.2. Let us denote by $\zeta_j = ||\mathbf{l}_s - \mathbf{l}_j||_2/c$ the time delay for acoustic wave propagation between the source and the j-th microphone, where $\mathbf{l}_j \in \mathbb{R}^3$ and $\mathbf{l}_s \in \mathbb{R}^3$ respectively denote the j-th microphone and source position vectors, and c is the speed of sound. The TDoA of the source signal at the j-th microphone with respect to a reference position $\mathbf{l}_r$ is given by

$$\zeta_{rj} = \zeta_j - \zeta_r \tag{2.44}$$

where $\zeta_r = ||\mathbf{l}_s - \mathbf{l}_r||_2/c$ is the propagation time delay between the source and the reference. The j-th element of the steering vector is then given by

$$b_k^j = e^{-\jmath 2\pi f_k \zeta_{rj}} \tag{2.45}$$

where $f_k = (k-1)f_s/F$ is the continuous frequency [Hz] corresponding to the k-th frequency bin with sampling rate f_s [Hz] and DFT size F. Alternatively, the TDoA for the j-th microphone can be written as

$$\zeta_{rj} = -\mathbf{l}_{rj}^T \mathbf{l}_o / c \tag{2.46}$$

where $\mathbf{l}_{rj} = \mathbf{l}_j - \mathbf{l}_r$ is the relative position vector of microphone j with respect to the reference position and $\mathbf{l}_o = (\mathbf{l}_s - \mathbf{l}_r)/||\mathbf{l}_s - \mathbf{l}_r||_2$ is the unit look direction vector of the source from the reference position. Considering the microphone array in Figure 2.2 as a particular example, where the first microphone is taken as the reference position (i.e., $\mathbf{l}_r = \mathbf{l}_1$) and the microphones are equally spaced along a line (i.e., uniform linear microphone array), the steering vector is found to be

$$\mathbf{b}_k = \begin{bmatrix} e^{-\jmath 2\pi f_k(-(\mathbf{l}_1-\mathbf{l}_1)^T\mathbf{l}_o)/c} \\ e^{-\jmath 2\pi f_k(-(\mathbf{l}_2-\mathbf{l}_1)^T\mathbf{l}_o)/c} \\ \vdots \\ e^{-\jmath 2\pi f_k(-(\mathbf{l}_J-\mathbf{l}_1)^T\mathbf{l}_o)/c} \end{bmatrix} = \begin{bmatrix} 1 \\ e^{-\jmath 2\pi f_k(d/c)\sin\theta} \\ \vdots \\ e^{-\jmath 2\pi f_k((J-1)d/c)\sin\theta} \end{bmatrix} \tag{2.47}$$

where θ is the DoA in radian.

To further reduce the residual noise components in the enhanced speech obtained via the MVDR beamformer, we can apply a single-channel enhancement algorithm to the beamformer output as a post-processor [19, 20], e.g., Wiener filtering:

$$\hat{S}^{SC}_{kl} = \frac{\hat{p}^S_{kl}}{\hat{p}^S_{kl} + \hat{p}^N_{kl}}\hat{S}_{kl} \tag{2.48}$$

where $\hat{p}^S_{kl}$ and $\hat{p}^N_{kl}$ are given by (2.32) and (2.33). Again, once the clean speech spectrum is estimated, the time-domain enhanced speech signal is obtained via inverse STFT, followed by the overlap-add method as explained in Subsection 2.2.1.

Chapter 3

Regularized NMF with Gaussian Mixtures and Masking Model

In this chapter, we introduce single-channel supervised speech enhancement algorithms based on regularized NMF[1]. In the proposed framework, the LLFs of the magnitude spectra for both the clean speech and noise, based on GMM, are included as regularization terms in the NMF cost function. By using this proposed regularization as *a priori* information in the enhancement stage, we can exploit the statistical properties of both the clean speech and noise signals. For further improvement of the enhanced speech quality, we also incorporate a masking model of the human auditory system in our approach. Specifically, we construct a WWF where the PSDs of the clean speech and noise are estimated from the above mentioned NMF algorithm with the proposed regularization. The weighting factor in the WWF is selected based on a masking threshold which is obtained from the estimated PSD of the enhanced speech. Experimental results show that the proposed speech enhancement

[1] Parts of this chapter have been presented at the 2014 IEEE International Workshop on Machine Learning for Signal Processing in Reims, France [79]; and have been published in the Speech Communication [80].

algorithms (i.e., regularized NMF with and without the masking model) provide better enhancement performance than the benchmark algorithms.

This chapter is organized as follows. In Section 3.1, we address the research motivation and a brief overview of the proposed algorithms. The proposed NMF training stage with GMM parameter estimation is described in Section 3.2. In Section 3.3, the proposed modifications to the enhancement stage, including NMF algorithm with regularization, masking threshold estimation and perceptually motivated NMF algorithm for speech enhancement are explained. Experimental results are presented in Section 3.4.

3.1 Research Motivations and Contributions

One of the main problems of the NMF-based supervised speech enhancement algorithms is the existence of a mismatch between the characteristics of the training and test data, which in turn leads to a decreased quality of the estimated source signals. One possible remedy to this problem is to add explicit regularization terms to the NMF cost function. Based on a classical approach, we can simply consider the l_1 or l_2-norm of the activation matrix [58, 64]. In order to account for the temporal dependency of the successive time frames, [38] models the activations by means of Markov chain. Employing the regularization terms that incorporate some prior knowledges has been also introduced. The authors in [81] and [54] use a HMM, while [40] uses GMMs that help the activations to follow certain patterns. In [79], both the speech and noise spectra are modeled by a GMM, and their LLFs are used as regularization terms.

Besides the speech enhancement or source separation algorithms which mainly focus on the perspective of signal estimation and reconstruction, several algorithms incorporating modeling aspects of the human auditory system have been proposed in order to improve the perceptual quality of the estimated source signals. Specifically, these refined algorithms exploit a psychoacoustical property called auditory masking which refers to a process whereby one sound is rendered inaudible due to the presence of another sound [82]. In the case of frequency domain (or simultaneous) masking, the threshold which models this effect has been used for selecting parameters in spectral subtraction [4], subspace decomposition [10], Wiener filtering [83] and MMSE-based estimator [84, 85]. In the NMF-based algorithms, weighted NMF update rules have been proposed by applying a weighting matrix based on the masking threshold to the NMF cost function [86, 87]. For speech enhancement, the masking threshold which determines the amount of the noise reduction is usually calculated from the estimated PSD of the clean speech. This suggests that a more accurate estimation scheme may lead to further improvement of the enhanced speech quality when applying a masking threshold.

In this chapter, we introduce single-channel supervised speech enhancement algorithms based on regularized NMF which are extensions of our earlier work [79]. The proposed framework seeks to exploit the statistical properties of *both* the clean speech and noise, an approach which is widely used in traditional speech enhancement algorithms. This is achieved in two ways: i) by representing the corresponding magnitude spectra, which capture the general characteristics of the signals, with the help of GMMs motivated by [22] and [24], and ii) by adding regularization terms that incorporate this *a priori* information to the NMF cost function in the enhancement stage. The proposed method, therefore, can be interpreted as a combination of the NMF and statistical model-based approaches. During the training stage, by using an isolated training set for each type of clean speech and noise, we estimate the basis matrices in the NMF model via multiplicative update rules [37] and the parameters of the GMMs via the EM algorithm [66,67]. For the GMM, we propose to use normalized spectral values in order to handle the magnitude difference between the training and test data, similar to the work of [40]. In the enhancement stage, the LLFs of the clean speech and noise magnitude spectra are added as regularization terms to the NMF cost function and the activation matrix of the noisy speech is estimated. Consequently, the PSDs of the clean speech and noise are obtained and the enhanced speech is reconstructed using Wiener filtering.

For further improvement of the enhanced speech quality, we incorporate the masking effects of the human auditory system in our approach. Specifically, we construct a WWF where the PSDs of the speech and noise are estimated from the above mentioned NMF algorithm with the proposed regularization. The weighting factor in the WWF is selected based on a masking threshold which is obtained from the estimated PSD of the speech based on [88].

3.2 Proposed Training Stage

In the proposed framework, *a priori* knowledge about the magnitude spectra of the clean speech and noise is captured by distinct GMMs. As a brief overview of the training stage, we first estimate the basis and activation matrices, i.e., $\mathbf{W}$ and $\mathbf{H}$, for the clean speech and noise independently using isolated training data. To this end, we consider the KL-divergence given in (2.2) and apply the resulting update rules in (2.5), leading to factorizations for the clean speech and noise magnitude spectra $\mathbf{V}_S \approx \mathbf{W}_S \mathbf{H}_S$ and $\mathbf{V}_N \approx \mathbf{W}_N \mathbf{H}_N$. Subsequently, the GMM parameters for the speech and noise are estimated from the corresponding NMF parameters. The details of this computation, which is identical for the speech and noise, are further developed below where for convenience in notation, the subscripts S and N are dropped.

In [22] and [24], the probability density function (PDF) of the clean speech spectrum is modeled by a GMM. Motivated by this approach, we model the PDFs of the magnitude spectra for *both* the clean speech and noise by distinct GMMs[2]. Therefore, we can expect that a more detailed and accurate statistical description is provided for the noise as well as the clean speech. In the proposed algorithm, we consider the product $\mathbf{W}\,\mathbf{H}$, which is an approximation of $\mathbf{V}$, as the observation matrix for the parameter estimation of the magnitude spectrum PDF[3], since we intend to introduce a clear connection with the regularization term shown in (2.23). Specifically, by expressing the observation as $\mathbf{W}\,\mathbf{H}$, we can directly differentiate the regularization term with respect to $\mathbf{H}$ while deriving the update rule given by (2.22) during the enhancement stage (a detailed derivation will be presented in Section 3.3). Moreover, in order to handle the magnitude difference between the training and test data, we consider normalized observations where the columns of $\mathbf{W}\,\mathbf{H}$ are normalized by their l_1-norm, similar to [40]. Specifically, we define the normalized column of the observation matrix as,

$$\bar{\mathbf{V}}_l \triangleq \frac{[\mathbf{W}\,\mathbf{H}]_l}{\sum_m h_{ml}} \tag{3.1}$$

where $[\cdot]_l$ denotes the l-th column of its matrix argument. Note that the l_1-norm of $[\mathbf{W}\,\mathbf{H}]_l$, i.e., $\sum_k [\mathbf{W}\,\mathbf{H}]_{kl}$, simply turns into $\sum_m h_{ml}$ since the basis vectors are normalized with respect to the l_1-norm, i.e., $\sum_k w_{km} = 1$ for $m \in \{1, ...M\}$. The GMM is defined in terms of the following parametric model for the PDF of $\bar{\mathbf{V}}_l$

$$p(\bar{\mathbf{V}}_l|\,\boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\bar{\mathbf{V}}_l|\,\mathbf{z}) = \sum_{i=1}^{I} g_i\,\mathcal{N}(\bar{\mathbf{V}}_l|\,\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{3.2}$$

[2] Alternatively, we can model the PDF of the magnitude spectra by a RMM (e.g., [21]) or gamma mixture model (e.g., [55]), which remain an interesting avenue for our future explorative work.

[3] Indeed, we could verify through independent experiments that there was no significant difference in the enhancement performance when considering either $\mathbf{V}$ or $\mathbf{W}\,\mathbf{H}$ as the observation matrix.

where I is the number of Gaussian components, $\mathbf{z} = [z_1, ..., z_I]^T$ is an I-dimensional vector of discrete latent variables $z_i \in \{0, 1\}$ with $\sum_i z_i = 1$, and the set $\boldsymbol{\theta} \triangleq \{g_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^I$ consists of the GMM parameters. The marginal distribution over $\mathbf{z}$ is specified in terms of the mixing coefficients $g_i \triangleq p(z_i = 1)$. The conditional PDF of $\bar{\mathbf{V}}_l$ given a particular value for the latent variable z_i is a K-dimensional Gaussian distribution such that $p(\bar{\mathbf{V}}_l|z_i = 1) = \mathcal{N}(\bar{\mathbf{V}}_l|\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ where $\boldsymbol{\mu}_i = [\mu_{i,k}]$ is the mean vector and $\boldsymbol{\Sigma}_i$ is the covariance matrix. In this work, we ignore possible correlations between different spectral components and therefore consider diagonal covariance matrices for simplicity, i.e., $\boldsymbol{\Sigma}_i = \text{diag}\{\sigma_{i,k}^2\}$. Recall that the entries of the observation matrix $\bar{\mathbf{V}} = [\bar{v}_{kl}]$ are magnitude spectral values which are strictly non-negative, while the GMM can in theory assign non-zero probability to negative values. Nevertheless, modeling matrix $\bar{\mathbf{V}}$ by a GMM is perfectly reasonable if the mean value of its entries exceed the corresponding standard deviation by a significant margin. More specifically, if say $\mu_{i,k} \geq 3\sigma_{i,k}$ for every Gaussian component $i = 1, ..., I$, then we can safely assume that $P_r[\bar{v}_{kl} < 0] \approx 0$. In effect, we have been able to verify that this condition is generally satisfied in our experimental work.

The parameter set $\boldsymbol{\theta} = \{g_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^I$ can be estimated using the EM algorithm [66,67]. For a given observation $\bar{\mathbf{V}} = [\bar{\mathbf{V}}_1, \bar{\mathbf{V}}_2, ..., \bar{\mathbf{V}}_L] = [\bar{v}_{kl}]$, where the column vectors $\bar{\mathbf{V}}_l$ are assumed to be drawn independently, the LLF can be written as,

$$\begin{aligned}
\mathcal{L}(\bar{\mathbf{V}}|\,\boldsymbol{\theta}) &\triangleq \ln p(\bar{\mathbf{V}}|\,\boldsymbol{\theta}) \\
&= \sum_{l=1}^{L} \ln \left\{ \sum_{i=1}^{I} g_i \,\mathcal{N}(\bar{\mathbf{V}}_l|\,\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right\} \\
&\geq \sum_{l=1}^{L} \sum_{i=1}^{I} q(z_i) \ln \left\{ \frac{g_i \,\mathcal{N}(\bar{\mathbf{V}}_l|\,\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{q(z_i)} \right\} \triangleq \mathcal{L}_B(\bar{\mathbf{V}}|\,\boldsymbol{\theta}) \qquad (3.3)
\end{aligned}$$

where $q(z_i)$ is an arbitrary probability distribution. The inequality holds for any choice of $q(z_i)$ due to Jensen's inequality [24, 41]. Note that $\mathcal{L}_B(\bar{\mathbf{V}}|\,\boldsymbol{\theta})$ defines a lower bound on Ł$(\bar{\mathbf{V}}|\,\boldsymbol{\theta})$ where the equality holds for $q(z_i) = p(z_i = 1|\bar{\mathbf{V}}_l, \boldsymbol{\theta})$, which is the posterior distribution of latent variable z_i given the observation $\bar{\mathbf{V}}_l$. The EM algorithm is an iterative procedure which consists of two steps. During the expectation step (E-step), the posterior distribution of each latent variable given the observation is calculated, which is shown as

$$\gamma_{il}^{(r)} \triangleq p(z_i = 1|\bar{\mathbf{V}}_l, \boldsymbol{\theta}^{(r)}) = \frac{g_i^{(r)} \mathcal{N}(\bar{\mathbf{V}}_l|\,\boldsymbol{\mu}_i^{(r)}, \boldsymbol{\Sigma}_i^{(r)})}{\sum_{i=1}^{I} g_i^{(r)} \mathcal{N}(\bar{\mathbf{V}}_l|\,\boldsymbol{\mu}_i^{(r)}, \boldsymbol{\Sigma}_i^{(r)})} \tag{3.4}$$

where the superscript (r) denotes the r-th iteration. In the maximization step (M-step), by *fixing* the posterior distribution to $\gamma_{il}^{(r)}$, the parameter set $\boldsymbol{\theta}$ which maximizes Ł$_B(\bar{\mathbf{V}}|\,\boldsymbol{\theta})$ is determined. In effect, since $\gamma_{il}^{(r)}$ in (3.4) does not depend on $\boldsymbol{\theta}$, this is equivalent to the maximization criterion of the expectation of the complete data LLF with respect to the posterior distribution,

$$\mathcal{L}_C(\bar{\mathbf{V}}|\,\boldsymbol{\theta}) \triangleq \sum_{l=1}^{L} \sum_{i=1}^{I} \gamma_{il}^{(r)} \ln\{g_i \mathcal{N}(\bar{\mathbf{V}}_l|\,\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}. \tag{3.5}$$

The solution of the M-step can be obtained in closed form as,

$$\begin{aligned} g_i^{(r+1)} &= \frac{1}{L} \sum_{l=1}^{L} \gamma_{il}^{(r)}, \\ \mu_{i,k}^{(r+1)} &= \frac{\sum_{l=1}^{L} \gamma_{il}^{(r)} \bar{v}_{kl}}{\sum_{l=1}^{L} \gamma_{il}^{(r)}}, \\ \sigma_{i,k}^{2\;\;(r+1)} &= \frac{\sum_{l=1}^{L} \gamma_{il}^{(r)} (\bar{v}_{kl} - \mu_{i,k}^{(r+1)})^2}{\sum_{l=1}^{L} \gamma_{il}^{(r)}}. \end{aligned} \tag{3.6}$$

As for the initialization of $\boldsymbol{\theta}$, we apply k-means clustering to $\bar{\mathbf{V}}$, which is an iterative algorithm aiming to partition the observations into clusters, such that each observation belongs to the cluster with the nearest mean [67]. The number of clusters is set equal to I, the number of Gaussian components in the GMM, while the cluster mean values are initialized randomly.

At this point, we emphasize the main difference between the above proposed training algorithm and the one presented in our earlier work [79]. In the latter, we considered joint training of $\mathbf{W}$, $\mathbf{H}$ and $\boldsymbol{\theta}$, where we used a regularized cost function as in (2.23) in which the regularization term was the expected LLF given by (3.5). We observed that the regularization coefficient λ not only determines the convergence behavior of the iterative update but that it also affects the enhancement performance. Hence, selecting an appropriate value for this coefficient is difficult. In addition, the iterative update using the joint training converges slowly and hence requires a more extensive computational effort. For these reasons, we chose to consider here instead a *sequential* form of training, which is found to be simpler and more efficient in both terms of computation and enhancement performance.

3.3 Proposed Enhancement Stage

In this section, we introduce the proposed regularized NMF algorithms. The LLF of the magnitude spectra for both the clean speech and noise based on distinct GMMs are included as regularization terms in the NMF cost function, which will be discussed in Subsection 3.3.1. For further improvement of enhancement performance, we incorporate a masking model of the human auditory system in our approach, which will be provided in Subsection 3.3.2. Specifically, we construct a WWF where the PSDs of the speech and noise are estimated by using the method in Subsection 3.3.1, and the weighting factor in the WWF is selected based on a masking threshold which is obtained from the estimated PSD of the clean speech.

3.3.1 Regularized NMF with Gaussian mixtures

In the proposed enhancement stage, the activation matrix of the noisy speech $\mathbf{H}_Y = [\mathbf{H}_S^T \ \mathbf{H}_N^T]^T$ is estimated using the regularized NMF algorithm based on (2.22) and (2.23), by fixing the basis matrices $\mathbf{W}_Y = [\mathbf{W}_S \ \mathbf{W}_N]$ and the GMM parameter sets of the clean speech and noise, $\boldsymbol{\theta}_S = \{g_i^S, \boldsymbol{\mu}_i^S, \boldsymbol{\Sigma}_i^S\}_{i=1}^{I_S}$ and $\boldsymbol{\theta}_N = \{g_i^N, \boldsymbol{\mu}_i^N, \boldsymbol{\Sigma}_i^N\}_{i=1}^{I_N}$, which are obtained during the training stage. Specifically, the LLFs of the clean speech and noise based on (3.3), i.e., $\mathcal{L}(\bar{\mathbf{V}}_S|\,\boldsymbol{\theta}_S)$ and $\mathcal{L}(\bar{\mathbf{V}}_N|\,\boldsymbol{\theta}_N)$, are used as regularization terms. The proposed regularized cost function is shown as,

$$\mathcal{J} = \mathcal{D}_{KL}(\mathbf{V}_Y, \mathbf{W}_Y\,\mathbf{H}_Y) - \mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y) \tag{3.7}$$

where $\mathcal{D}_{KL}(\cdot)$ is the KL-divergence given in (2.2) and $\mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y)$ is the proposed regularization term written as,

$$\mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y) = \lambda_S\,\mathcal{L}(\bar{\mathbf{V}}_S|\,\boldsymbol{\theta}_S) + \lambda_N\,\mathcal{L}(\bar{\mathbf{V}}_N|\,\boldsymbol{\theta}_N) \tag{3.8}$$

where $\mathcal{L}(\cdot|\cdot)$ is given in (3.3) and $\bar{\mathbf{V}}_S$, $\bar{\mathbf{V}}_N$ are the normalized clean speech and noise spectra defined by (3.1). The values $\lambda_S > 0$ and $\lambda_N > 0$ are the regularization coefficients for the clean speech and noise, respectively. The optimal choices for λ_S and λ_N depend on the input SNR as well as the speaker, the type of noise and regularization term. In this work, however, we do not consider such dependencies (except the type of regularization term), and use constant values for simplicity, as we found indeed that the optimal choices mostly depend on the regularization term. Note that a negative sign is applied to the regularization term in (3.7), since the latter will represent a reward as opposed to a penalty.

For the derivation of the update rule of $\mathbf{H}_Y$, we first compute the gradient of

$\mathcal{D}_{KL}(\mathbf{V}_Y, \mathbf{W}_Y \mathbf{H}_Y)$ with respect to $\mathbf{H}_Y$. This gradient is shown as

$$\nabla_{\mathbf{H}_Y} \mathcal{D}_{KL} = \nabla^+_{\mathbf{H}_Y} \mathcal{D}_{KL} - \nabla^-_{\mathbf{H}_Y} \mathcal{D}_{KL} \tag{3.9}$$

where the dependence of $\mathcal{D}_{KL}(\mathbf{V}_Y, \mathbf{W}_Y \mathbf{H}_Y)$ on $\mathbf{V}_Y$ and $\mathbf{W}_Y \mathbf{H}_Y$ is omitted for notational convenience, and the values on the right-hand side are

$$\nabla^+_{\mathbf{H}_Y} \mathcal{D}_{KL} = \mathbf{W}_Y^T \mathbf{1} \tag{3.10}$$

$$\nabla^-_{\mathbf{H}_Y} \mathcal{D}_{KL} = \mathbf{W}_Y^T (\mathbf{V}_Y / (\mathbf{W}_Y \mathbf{H}_Y)) \tag{3.11}$$

where $\mathbf{1}$ is a $K \times Lc_Y$ matrix with all entries equal to one. Note that (3.10) and (3.11) appear respectively in the denominator and numerator in (2.5). Next, we derive the gradient of the regularization term $\mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y)$ in (3.8) with respect to $\mathbf{H}_Y$. Note that by using the equality in (3.3), i.e., $\mathcal{L}(\bar{\mathbf{V}}|\,\boldsymbol{\theta}) = \mathcal{L}_B(\bar{\mathbf{V}}|\,\boldsymbol{\theta})$ for $q(z_i) = \gamma_{il}$, the gradient of $\mathcal{L}(\bar{\mathbf{V}}|\,\boldsymbol{\theta})$ is identical to that of $\mathcal{L}_B(\bar{\mathbf{V}}|\,\boldsymbol{\theta})$, which is equivalent to the gradient of $\mathcal{L}_C(\bar{\mathbf{V}}|\,\boldsymbol{\theta})$. Consequently, the gradient of (3.8) can be shown in terms of the gradients of $\mathcal{L}_C(\bar{\mathbf{V}}_S|\,\boldsymbol{\theta}_S)$ and $\mathcal{L}_C(\bar{\mathbf{V}}_N|\,\boldsymbol{\theta}_S)$ with respect to $\mathbf{H}_S$ and $\mathbf{H}_N$, respectively, as,

$$\nabla_{\mathbf{H}_Y} \mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y) = \begin{bmatrix} \lambda_S \nabla_{\mathbf{H}_S} \mathcal{L}_C(\bar{\mathbf{V}}_S|\,\boldsymbol{\theta}_S) \\ \lambda_N \nabla_{\mathbf{H}_N} \mathcal{L}_C(\bar{\mathbf{V}}_N|\,\boldsymbol{\theta}_N) \end{bmatrix} \tag{3.12}$$

where $\mathcal{L}_C(\cdot|\cdot)$ is the expected LLF given in (3.5). As we can see from (3.1), the observations $\bar{\mathbf{V}}_S$ and $\bar{\mathbf{V}}_N$ are expressed in terms of the corresponding basis and activation matrices. Hence, using (3.5), we can derive the gradients of the expected LLF with respect to the activation matrix in (3.12), which is shown as

$$\nabla_{\mathbf{H}} \mathcal{L}_C = \nabla^+_{\mathbf{H}} \mathcal{L}_C - \nabla^-_{\mathbf{H}} \mathcal{L}_C \tag{3.13}$$

where $\mathbf{H}$ stands for either $\mathbf{H}_S$ or $\mathbf{H}_N$, and the dependence of $\mathcal{L}_C(\bar{\mathbf{V}}|\,\boldsymbol{\theta})$ on $\bar{\mathbf{V}}$ and $\boldsymbol{\theta}$ is omitted for convenience. In (3.13), the entries of the gradient terms on the right-hand side are

$$[\nabla^+_{\mathbf{H}} \mathcal{L}_C]_{ml} = \sum_{k=1}^{K} \sum_{i=1}^{I} \gamma_{il} \sigma_{i,k}^{-2} \Big(\mu_{i,k} \frac{w_{km}}{c_l} + \frac{([\mathbf{W}\,\mathbf{H}]_{kl})^2}{c_l^3} \Big) \tag{3.14}$$

$$[\nabla^-_{\mathbf{H}} \mathcal{L}_C]_{ml} = \sum_{k=1}^{K} \sum_{i=1}^{I} \gamma_{il} \sigma_{i,k}^{-2} (w_{km} + \mu_{i,k}) \frac{[\mathbf{W}\,\mathbf{H}]_{kl}}{c_l^2} \tag{3.15}$$

where γ_{il} is the posterior distribution given in (3.4) and $c_l = \sum_m h_{ml}$ is the normalizing factor. Specifically, γ_{il} is computed based on $\mathbf{W}_S$ and $\mathbf{W}_S$ obtained during the training stage and $\mathbf{H}_Y$ estimated in the previous multiplicative update iteration. Note that, based on the concept of the lower bound in (3.3) and the objective used in the M-step given by (3.5), the posterior γ_{il} is considered as a fixed constant value during the derivations of (3.14) and (3.15).

Based on the heuristic MU rules given in (2.22), the update rule of $\mathbf{H}_Y$ can be written as,

$$\hat{\mathbf{H}}_Y \leftarrow \hat{\mathbf{H}}_Y \otimes \frac{\nabla^-_{\mathbf{H}_Y} \mathcal{D}_{KL}(\mathbf{V}_Y, \mathbf{W}_Y\,\hat{\mathbf{H}}_Y) + \nabla^+_{\mathbf{H}_Y} \mathcal{R}_Y(\mathbf{W}_Y, \hat{\mathbf{H}}_Y)}{\nabla^+_{\mathbf{H}_Y} \mathcal{D}_{KL}(\mathbf{V}_Y, \mathbf{W}_Y\,\hat{\mathbf{H}}_Y) + \nabla^-_{\mathbf{H}_Y} \mathcal{R}_Y(\mathbf{W}_Y, \hat{\mathbf{H}}_Y)} \tag{3.16}$$

where $\nabla^+_{\mathbf{H}_Y} \mathcal{D}_{KL}(\mathbf{V}_Y, \mathbf{W}_Y\,\mathbf{H}_Y)$ and $\nabla^-_{\mathbf{H}_Y} \mathcal{D}_{KL}(\mathbf{V}_Y, \mathbf{W}_Y\,\mathbf{H}_Y)$ are given in (3.10) and (3.11). The components $\nabla^+_{\mathbf{H}_Y} \mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y)$ and $\nabla^-_{\mathbf{H}_Y} \mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y)$ are easily found by substituting (3.13) into (3.12). That is,

$$\nabla^{+}_{\mathbf{H}_Y} \mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y) = \begin{bmatrix} \lambda_S \nabla^{+}_{\mathbf{H}_S} \mathcal{L}_C(\bar{\mathbf{V}}_S | \boldsymbol{\theta}_S) \\ \lambda_N \nabla^{+}_{\mathbf{H}_N} \mathcal{L}_C(\bar{\mathbf{V}}_N | \boldsymbol{\theta}_N) \end{bmatrix} \tag{3.17}$$

$$\nabla^{-}_{\mathbf{H}_Y} \mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y) = \begin{bmatrix} \lambda_S \nabla^{-}_{\mathbf{H}_S} \mathcal{L}_C(\bar{\mathbf{V}}_S | \boldsymbol{\theta}_S) \\ \lambda_N \nabla^{-}_{\mathbf{H}_N} \mathcal{L}_C(\bar{\mathbf{V}}_N | \boldsymbol{\theta}_N) \end{bmatrix} \tag{3.18}$$

where $\nabla^{+}_{\mathbf{H}_{(\cdot)}} \mathcal{L}_C(\cdot|\cdot)$ in (3.17) and $\nabla^{-}_{\mathbf{H}_{(\cdot)}} \mathcal{L}_C(\cdot|\cdot)$ in (3.18) are given in (3.14) and (3.15), respectively.

It is easy to show that the update rule given in (3.16) takes on non-negative values. In fact, since the posterior distribution and all elements of the mean vector and the diagonal entries of the covariance matrix are non-negative, the values given in (3.14) and (3.15) are non-negative. Moreover, the values in (3.10) and (3.11) are also non-negative, and therefore the activation matrix is updated under the non-negative elements constraint.

After estimating the activation matrix of the noisy speech, the smoothed PSDs of both the clean speech and noise, $\hat{\mathbf{P}}_S = [\hat{p}^S_{kl}]$ and $\hat{\mathbf{P}}_N = [\hat{p}^N_{kl}]$, are obtained by using (2.32) and (2.33). Then the clean speech spectrum is estimated by Wiener filtering as given in (2.31). This proposed algorithm based on regularized NMF with Gaussian mixtures will be referred to as RNG.

3.3.2 Weighted Wiener filtering based on masking threshold

In this subsection, we describe our second method which uses a WWF. The masking threshold estimation is described first, followed by the proposed WWF.

Masking threshold estimation The masking effect, which is a psychoacoustical property of the human auditory system, has been employed in diverse applications such as audio and speech coding [88] and speech enhancement [4, 10, 83]. Masking refers to a process where one sound is rendered inaudible (maskee) due to the presence of another sound (masker) [82]. The masking properties are modeled using a masking threshold, where the components below the threshold are not perceived. There are two main masking phenomena, simultaneous (spectral) and non-simultaneous (temporal) masking. The former occurs whenever two or more stimuli are simultaneously presented to the auditory system. The latter takes place in the time domain, where the masking occurs both prior and after the onset and offset of the masker with finite duration [82]. In the proposed framework, we only consider the simultaneous masking effect.

Simultaneous masking can be explained in terms of critical band analysis which is a central mechanism in the inner ear. The critical band is specified by means of the so-called Bark scale, which is a perceptual measure relating acoustical frequency to the nonlinear perceptual resolution, in which one Bark covers one critical band. The analytical expression of the mapping function from the frequency f [kHz] to the Bark frequency B [Bark] is shown as

$$B_f = 13 \text{ arctan}(0.76f) + 3.5 \text{ arctan}[(f/7.5)^2]. \tag{3.19}$$

We followed the procedure introduced in [88] for evaluating the masking threshold in the l-th time frame, where we here briefly summarize the different steps involved in the computation; further implementation details are given in [88].

1) *Spectral analysis and normalization*: The PSD is normalized and presented in dB scale as,

$$\bar{p}_{kl} = 90.302 + 10\log_{10}[\hat{p}^S_{kl}/L_w^2] \tag{3.20}$$

where L_w denotes the analysis window length for the STFT, the constant 90.302 is used for the power compensation, and $\hat{p}^S_{kl}$ is the estimated clean speech PSD given in (2.32).

2) *Identification of tonal and non-tonal maskers*: Tonal maskers are identified according to the local maxima of the normalized PSD, $\bar{p}_{kl}$. A single non-tonal (noise-like) masker for each critical band is then identified by summing the energy of the spectral components which have not contributed to a tonal masker.

3) *Reorganization of maskers*: Any tonal or non-tonal maskers below the absolute hearing threshold (AHTH) are discarded, where the AHTH in dB versus frequency f [kHz] is shown as

$$T^A_f = 3.65f^{-0.8} - 6.5e^{-0.6(f-3.3)^2} + 10^{-3}f^4 \tag{3.21}$$

Next, any pair of maskers within a distance of 0.5 Bark are replaced by the stronger of the two.

4) *Individual masking threshold*: The individual masking threshold at frequency bin i due to a tonal masker at frequency bin j is given in dB as

$$T^{tm}_{ij} = \bar{p}^{tm}_j - 0.275\ B_{f_j} + \mathrm{SF}_{ij} - 6.025 \tag{3.22}$$

where $\bar{p}^{tm}_j$ is the level of tonal masker, f_j [kHz] is the corresponding frequency of the j-th bin, B_{f_j} denotes the Bark frequency given in (3.19) and SF_{ij} is the spreading function which accounts for the inter-band masking. The latter is given by

$$\mathrm{SF}_{ij} = \begin{cases} 17\Delta_B - 0.4\bar{p}_j^{tm} + 11, & -3 \leq \Delta_B < -1 \\ (0.4\bar{p}_j^{tm} + 6)\Delta_B, & -1 \leq \Delta_B < 0 \\ -17\Delta_B, & 0 \leq \Delta_B < 1 \\ (0.15\bar{p}_j^{tm} - 17)\Delta_B - 0.15\bar{P}_j^{tm}, & 1 \leq \Delta_B < 8 \end{cases} \tag{3.23}$$

where $\Delta_B = B_{f_i} - B_{f_j}$. Similarly, the masking threshold of a non-tonal masker is given by

$$T_{ij}^{nm} = \bar{p}_j^{nm} - 0.175\ B_{f_j} + \mathrm{SF}_{ij} - 2.025 \tag{3.24}$$

where $\bar{p}_j^{nm}$ is the non-tonal masker level. The spreading function used in (3.24) is identical to (3.23) where $\bar{p}_j^{tm}$ is replaced by $\bar{p}_j^{nm}$. The above computation of the masking thresholds T_{ij}^{tm} for tonal maskers and T_{ij}^{nm} for non-tonal ones are repeated for each frame; whenever such a computed threshold value falls below the AHTH, it is replaced by the latter.

5) *Global masking threshold*: Finally, the resulting individual masking thresholds are summed linearly along with the AHTH to obtain the global masking threshold in dB in the k-th frequency bin, which is shown as,

$$T_{kl}^{g} = 10\log_{10}\left(10^{0.1T_{f_k}^{A}} + \sum_{n=1}^{N_{tm}} 10^{0.1T_{k,j_n}^{tm}} + \sum_{n=1}^{N_{nm}} 10^{0.1T_{k,j_n}^{nm}}\right) \tag{3.25}$$

where N_{tm} and N_{nm} respectively denote the number of tonal and non-tonal maskers and j_n is the frequency bin location of the n-th masker. An example of the global masking threshold is illustrated in Figure 3.1, where we considered a speech signal of a female speaker.

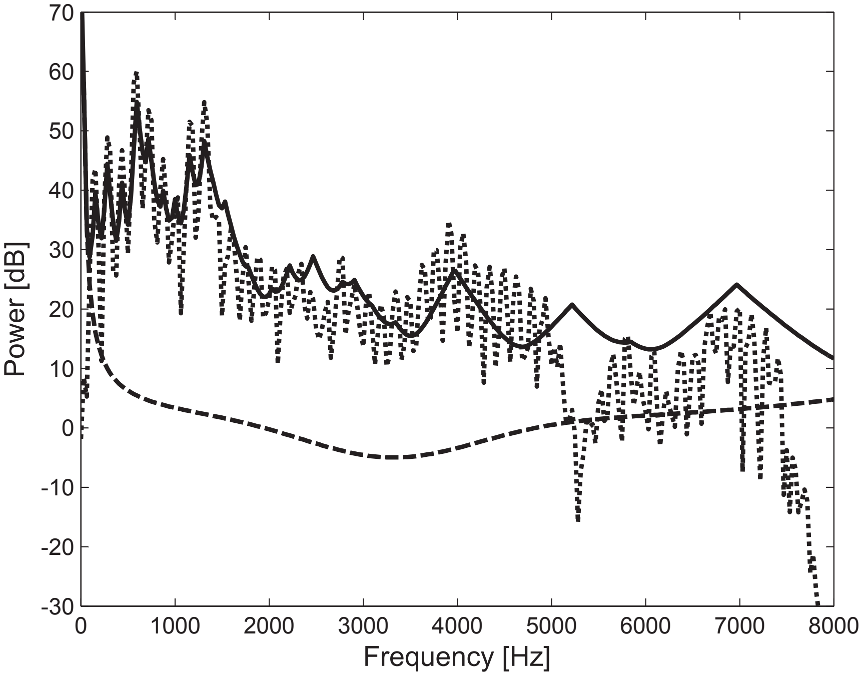

Fig. 3.1 Example of masking threshold (dotted: normalized power spectrum of a female speaker, solid: masking threshold, dashed: absolute hearing threshold).

Weighted Wiener filtering A generalized Wiener filtering has been introduced in [1], which is shown as,

$$\hat{S}_{kl} = \left(\frac{\hat{p}^S_{kl}}{\hat{p}^S_{kl} + \eta \hat{p}^N_{kl}} \right)^{\nu} Y_{kl} \tag{3.26}$$

where η and ν are tuning parameters. For simplicity, we will fix ν to 1 in the proposed framework, and refer to the resulting method as weighted Wiener filtering [89]. The weighting factor η is known to control the trade-off between noise reduction and speech distortion. For a large η, for instance, more noise reduction is performed at the expense of increased speech distortion, and vice versa. This phenomenon is illustrated in Figure 3.2 where we computed different objective measures while varying η from 1 to 20. The objective measures considered are the source-to-interference ratio (SIR), source-to-artifact ratio (SAR)

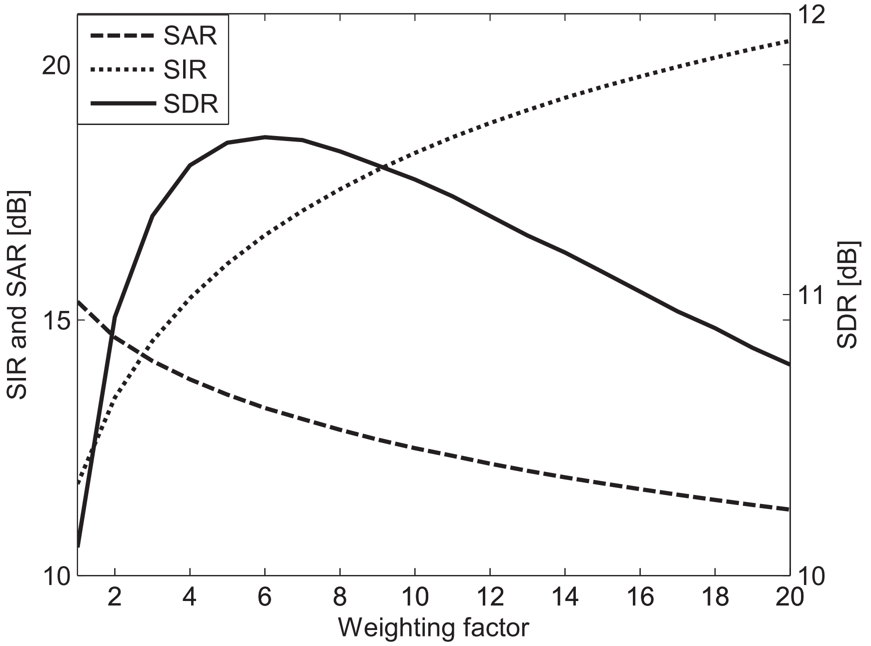

Fig. 3.2 SDR, SIR and SAR values for different weighting factors in WWF.

and source-to-distortion ratio (SDR) [99][4]. The noisy speech was generated by adding a factory noise to selected clean speech files[5] at a 5 dB input SNR, and the results were obtained by averaging over different speakers. For each noisy speech, the clean speech and noise PSDs were computed from the proposed RNG method introduced in Subsection 3.3.1, followed by temporal smoothing given in (2.32) and (2.33). As we can see from Figure 3.2, the results obtained for the different objective measures vary greatly as a function of η and therefore, an appropriate selection of the weighting factor is necessary.

In contrast to using a constant value as the weighting factor in (3.26), it has been proposed to select different weighting factor for each time-frequency bin, i.e., η_{kl}, based on

[4] For a given target source, the interference refers to unwanted signal components such as noise, whereas the artifact refers to components caused by other phenomena, such as e.g., forbidden distortion. In speech enhancement applications, these measures can be interpreted as follows: the SIR and SAR are proportional to the amount of noise reduction and inversely proportional to the speech distortion, respectively, while SDR measures the overall quality of the enhanced speech [34].

[5] Further details about various speech and noise files used in our experimental work are described in detail in Section 3.4.

the masking threshold computed for each of these bins. [90] proposed a heuristic approach where the linear estimator of the clean speech spectrum was derived, aiming to mask the distortion of the residual noise which is defined as the difference between the actual and residual noise powers. This estimator was extended in [83] by solving an optimization problem which minimizes a related error criterion. The authors in [91] proposed to use an exponential function to map the so-called noise-to-mask ratio (NMR) into the weighting factor, where the NMR in dB, Φ_{kl}, is defined as the log distance from the minimum masking threshold in one critical band to the noise level [88]:

$$\Phi_{kl} = \bar{p}_{kl}^{N} - \min_{k \in C_b} T_{kl}^{g} \tag{3.27}$$

where C_b is the set of frequency bins for the b-th critical band and $\bar{p}_{kl}^{N}$ is the normalized PSD given in (3.20).

For all these algorithms, a zero weighting factor is applied when the noise power is lower than the masking threshold, i.e., $\eta_{kl} = 0$ for $T_{kl}^{g} > \bar{p}_{kl}^{N}$. However, this strict condition limits the performance, since the masking threshold is calculated from an inaccurate estimate of the clean speech PSD. Although we can expect that a more accurate clean speech PSD can be obtained by using the proposed RNG method, we further suggest to relax this strict condition by taking into account in a continuous way the case where the noise power is even lower than the masking threshold. This approach can be regarded as a *soft* decision on the weighting factor.

In advance of describing the proposed method, we summarize several intuitive aspects, which should be considered for selecting the weighting factors in the WWF, as follows. When T_{kl}^{g} is low, the noise signal (maskee) is easily perceived due to the low masking capability of the speech signal (masker). The emphasis then should be put on reducing this

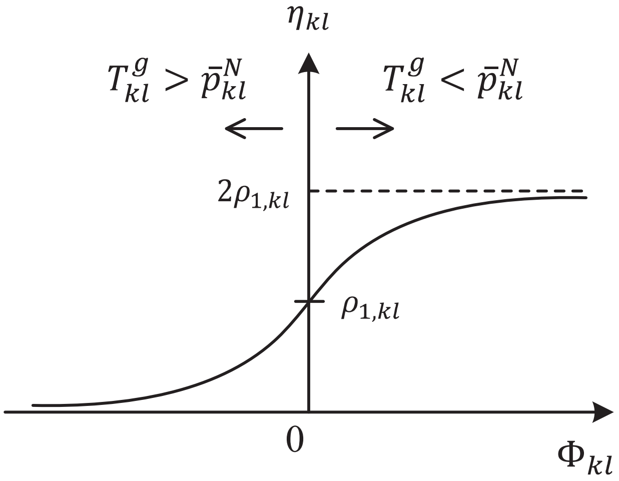

Fig. 3.3 Proposed mapping function from NMR, Φ_{kl}, to weighting factor, η_{kl}, based on a sigmoid function.

perceivable noise. Consequently, a high weighting factor is necessary in the WWF. On the contrary, if T_{kl}^{g} is high, the noise is easily masked by the speech. Hence, a small weighting factor is selected. Note that these aspects hold for both the cases where the NMR is either positive or negative. The difference is that a much smaller weighting factor for the case of negative NMR is necessary compared to the positive NMR.

In the proposed WWF, the weighting factor is selected through a heuristic approach using a sigmoid function as a mapping from the NMR to the weighting factor. The motivation for using the logistic function is to limit the range of the weighting factor to be selected, therefore avoiding extreme values that could lead to instability (Figure 3.3). The proposed mapping function is given by

$$\eta_{kl} = \frac{2\rho_{1,kl}}{1 + \exp(-\rho_{2,kl}\Phi_{kl})} \tag{3.28}$$

where $\rho_{1,kl}, \rho_{2,kl} > 0$ are tuning parameters and the NMR, Φ_{kl}, is given in (3.27). The value $\rho_{1,kl}$ defines the range of $\eta_{kl} \in (0, 2\rho_{1,kl})$ and $\rho_{2,kl}$ determines the slope of the sigmoid function. For simplicity of the implementation, we consider a constant slope, i.e., $\rho_{2,kl} = \rho_2$, and identical values of $\rho_{1,kl}$ across the frequency bins for a given time frame, i.e., $\rho_{1,kl} = \rho_{1,l}$.

The value $\rho_{1,l}$ is calculated using the following function

$$\rho_{1,l} = \xi_1 e^{-\xi_2 R_l} \tag{3.29}$$

where $\xi_1, \xi_2 > 0$ are tuning parameters and R_l is defined as

$$R_l = 10 \log_{10} \frac{\sum_k \hat{p}_{kl}^S}{\sum_k \hat{p}_{kl}^N}. \tag{3.30}$$

The underlying motivation for using the form given in (3.29) and (3.30) is similar to the approach introduced in [92]. That is, a small weighting factor is selected for a high input SNR. Specifically in the proposed method, the input SNR for a given time frame of the noisy speech is estimated from R_l given in (3.30), which is then applied to determine the range of η_{kl} through $\rho_{1,l}$ given in (3.29).

The proposed enhancement algorithm based on the regularized NMF with Gaussian mixtures and weighted Wiener filtering will be referred to as RNG-WWF. A simplified block diagram of both the RNG and RNG-WWF methods is illustrated in Figure 3.4. We note that for both algorithms, the same training approach as described in Section 3.2 is employed.

3.4 Experiments

In this section, a performance evaluation of the proposed methods is presented.

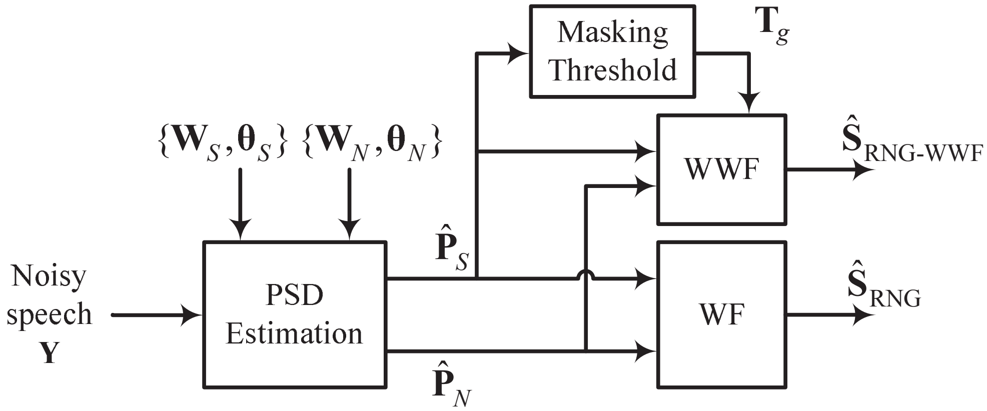
Masking Threshold
$\mathbf{T}_g$
$\{\mathbf{W}_S, \boldsymbol{\theta}_S\}$
$\{\mathbf{W}_N, \boldsymbol{\theta}_N\}$
WWF
$\hat{\mathbf{S}}_{\text{RNG-WWF}}$
Noisy speech $\mathbf{Y}$
PSD Estimation
$\hat{\mathbf{P}}_S$
WF
$\hat{\mathbf{S}}_{\text{RNG}}$
$\hat{\mathbf{P}}_N$

Fig. 3.4 Simplified block diagrams of RNG and RNG-WWF methods.

3.4.1 Methodology

We used clean speech from the TSP [95] and Grid Corpus [96] databases and noise from the NOISEX database [97], where the sampling rate of all signals was adjusted[6] to 16 kHz. For the clean speech, 20 speakers (10 males and 10 females) were selected from the TSP and 34 speakers (17 males and 17 females) from the Grid Corpus databases for a total of 54 speakers. For the noises, we selected the Buccaneer 1, Hfchannel, Babble and Factory 1 noises from the NOISEX database. Each clean speech and noise signal was divided into three disjoint groups: i) *training data*, used for estimating the NMF and GMM parameters, ii) *validation data*, used for selecting the regularization coefficients and tuning parameters, and iii) *test data*, used for final verification. Specifically, the training data consisted of approximately 2 minutes (50 sentences) and 8 minutes (350 utterances) of long speech segments for each speaker from the TSP and Grid Corpus databases, respectively, as well as 3 minutes segment for the noises. The validation data consisted of 12 seconds (5 sentences) and 20 seconds (15 utterances) of speech for each speaker from the TSP and Grid Corpus databases, respectively, and 30 seconds of noise from the NOISEX database.

[6] The original noise signals with 8 kHz sampling rate were upsampled to 16 kHz.

The same partitioning was used for the test data. The noisy speech signals were generated from the test and validation signals by scaling and adding the noise to the clean speech (based on the estimated variances of the time-domain signals) to obtain input SNRs of 0, 5 and 10 dB[7]. The STFT analysis was implemented by using a Hanning window of 512 samples with 50 % overlap. After enhancement, the estimated clean speech signal in the time-domain was reconstructed by applying the inverse STFT on its spectrum followed by the overlap-add method.

Regarding the implementation of the proposed algorithms, we considered a speaker-dependent (SD) application, where one basis matrix and associated GMM parameter set were trained for each speaker. We used $M = 80$ basis vectors and $I = 8$ Gaussian components in the GMM for both the clean speech and noise. The values of $(\tau_S, \tau_N) = (0.4, 0.9)$ were chosen empirically using the validation set and used as the temporal smoothing factors in (2.32) and (2.33). For the regularization coefficients λ_S and λ_N in (3.8), we examined different values from 0.0005 to 0.1 and obtained good results in the range $[0.005, 0.01]$. Hence, we selected $(\lambda_S, \lambda_N) = (0.005, 0.01)$. We also examined several choices for the tuning parameters in the proposed weighting function (3.28), i.e. ξ_1, ξ_2 and ρ_2. We first fixed ξ_1 to 4, 5 and 6, based on the results shown in Figure 3.2. For each value of ξ_1, we then considered various choices of ρ_1 and ξ_2 and determined the ones that gave the highest SDR values. Good results for both ρ_2 and ξ_2 were found around $[0.005, 0.1]$. Ultimately, we chose $\rho_2 = 0.01$ and $(\xi_1, \xi_2) = (5, 0.1)$ for the experiments.

We used the perceptual evaluation of speech quality (PESQ) [98], SDR [99], as well as the segmental SNR (SSNR) as the objective measures of performance. The PESQ attempts to predict overall perceptual quality in mean opinion score (MOS) and the SDR measures

[7] For a given source speech file, the desired input SNR values were obtained by scaling the noise signal level. In this case, we assume that there is no long pause in the speech signal, which is indeed justified for the data we used in the experiments.

the overall quality of the enhanced speech in dB by considering both the speech distortion and noise reduction as explained in Subsection 3.3.2. For all the measures, a higher value indicates a better result.

3.4.2 Benchmark algorithms

To evaluate the speech enhancement performance of the newly proposed algorithms, we compared them against several algorithms from the literature. Basic settings such as the STFT analysis and synthesis, number of basis vectors and Gaussian components in the GMM, and masking threshold calculations, when applicable, were kept identical for all the benchmark and proposed algorithms. Also, we considered the SD application for all NMF-based algorithms.

The benchmark algorithms were categorized into two groups. The purpose of the first group was only to compare the enhancement performance of the proposed WWF (i.e., RNG-WWF) to that of other perceptually-motivated and/or weighting methods. Specifically, we considered the algorithms proposed by [83, 90–92]; in the sequel, we shall refer to each algorithm using the names of its authors for simplicity. Although the algorithms in [91] and [92] were proposed for multi-channel speech enhancement, they can still be applied in the current single-channel framework. We used the following tuning parameters for these algorithms: a trade-off control parameter $\zeta = 0.1$ in [90], $(\gamma, \delta, \epsilon) = (0.2, 0.9, 0.9)$ in [91] and $(\alpha, \beta) = (1, 2)$ in [92] (see the references for the meaning of these notations). For all the benchmark algorithms and RNG-WWF method, we employed identical PSDs of the clean speech and noise, which were estimated using the RNG method. The salient features of the benchmarks and proposed algorithms are summarized in Table 3.1.

The purpose of the second group was to compare the enhancement performance of the proposed algorithms with that of various speech enhancement algorithms, which are given

Table 3.1 A comparison between different perceptually-motivated and/or weighting methods

Reference	Gain function, G_{kl} ($\hat{S}_{kl} = G_{kl} Y_{kl}$)	Description
Gustafsson et al. [90]	$\min\left(\sqrt{\frac{T^g_{kl}}{\hat{p}^N_{kl}}} + \zeta, 1\right)$	Heuristic gain function, aiming to mask the distortion of the residual noise
Hu et al. [83]	$\left(1 + \max\left(\sqrt{\frac{\hat{p}^N_{kl}}{T^g_{kl}}} - 1, 0\right)\right)^{-1}$	Gain function obtained by minimizing an error criterion (extension of [90])
Defraene et al. [91]	$\frac{\hat{p}^S_{kl}}{\hat{p}^S_{kl} + \eta_{kl}\hat{p}^N_{kl}}$	Heuristic mapping from the NMR to η_{kl} (*hard* desicion)
Kodrasi et al. [92]		Curvature-based optimization for the estimation of η_{kl}
Proposed		Heuristic mapping from the NMR to η_{kl} (*soft* desicion)

below. Note that, for all NMF-based algorithms, except the proposed RNG-WWF method which requires a weighting factor, we used the same reconstruction method introduced in Subsection 2.2.1, i.e., computing smoothed PSDs and Wiener filtering, for fair comparison.

1) *Short-time spectral amplitude estimator (STSA)*: We implemented the well-known classical MMSE-STSA estimator proposed by [5]. A smoothing factor of 0.98 in the decision-directed (DD) method for *a priori* SNR estimation was used. The noise PSD was estimated using an algorithm described in [93] with a value of 0.8 for the smoothing factor.

2) *Spectral subtraction with masking properties (SSM)*: We considered a spectral subtraction algorithm with masking properties proposed in [4]. The noise PSD in this approach was also estimated using the algorithm from [93] with 0.8 for the smoothing factor.

3) *Standard NMF*: The standard NMF algorithm based on KL-divergence introduced in Chapter 2 was evaluated, which will be referred to as NMF.

4) *Regularized NMF*: In order to compare with other regularization-based NMF algorithms, we chose an algorithm proposed by [40], where the column vectors of the activation matrix of the clean speech and noise are modeled by distinct GMMs. We employed the sequential form of training, and used the regularization coefficients of

$(\lambda_S, \lambda_N) = (0.005, 0.001)$ in our experiments as they provided good results. This method will be referred to as RNMF-AGM.

5) *Weighted NMF (WNMF)*: We evaluated a perceptually weighted NMF (WNMF) algorithm introduced in [87], where the perceptual weighting matrix was constructed (based on the masking threshold) as in [94]. Although the WNMF algorithm was originally proposed for an unsupervised application, we applied it in a supervised manner. That is, the basis matrices for the clean speech and noise were obtained independently during the training stage. In the enhancement stage, the WNMF activation update was applied to the noisy speech, where the masking threshold was calculated from the noisy speech. Although the masking threshold can be obtained from the estimated clean speech PSD by first applying a simple speech enhancement scheme, e.g., [4, 91], we followed the original paper, since we observed similar results when using the masking threshold either computed from the noisy or estimated clean speech PSD.

3.4.3 Results

We first illustrate an example of the proposed weighting factor η_{kl} for different input SNRs in Figure 3.5. In this particular example, a male speech is degraded with Buccaneer 1 noise at 0, 5 and 10 dB input SNR. We can make the following observations:

- The values of η_{kl} around 3 kHz, which corresponds to the intense ringing sound of the buccaneer 1 noise, are larger compared to the other frequencies;
- For a given time-frequency bin, η_{kl} decreases as the input SNR increases from 0 to 10 dB;
- The values of η_{kl} at the time frame of 2.1s (a speech-absence period) are larger than the ones at 1.3s (a speech-presence period).

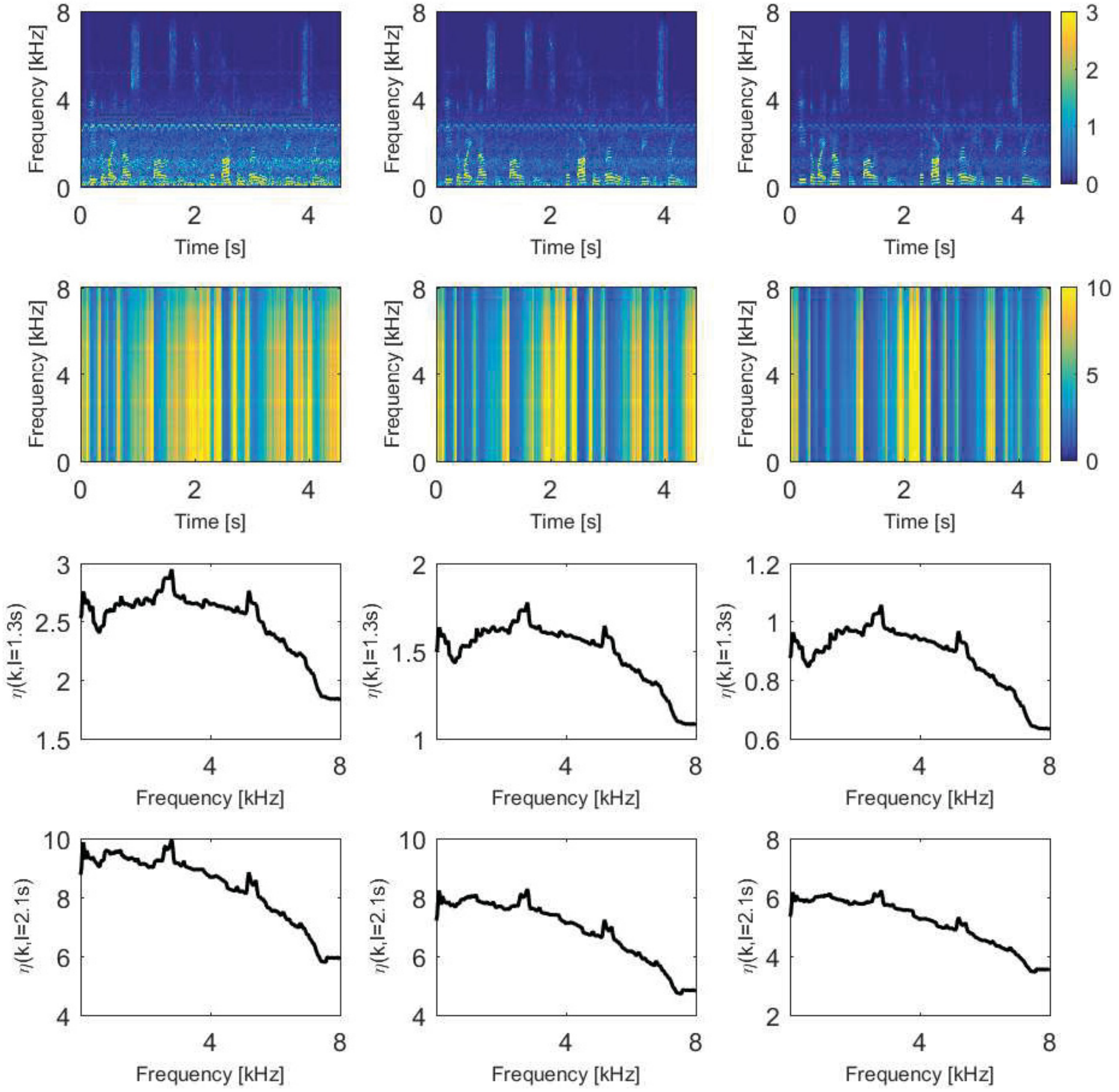

Fig. 3.5 Examples of proposed weighting factor. Each column from left to right respectively correspond to input SNR of 0, 5 and 10 dB. Each row from top to bottom shows the noisy speech magnitude spectrum, time-frequency representation of the proposed weighting factor and the weighting factor at the time frame of 1.3s and 2.1s.

These phenomena are essentially due to the estimated input SNR R_l given by (3.30). That is, as we intended, a larger value of η_{kl} is selected based on (3.28) and (3.29), for a lower value of R_l. Consequently, the noise components will be further suppressed in the corresponding time-frequency bins.

We compared the proposed RNG-WWF method with other methods in the first group of benchmark algorithms in order to verify the performance of the proposed weighting method. Average SDR and SSNR values over all speakers for Factory 1 and Hfchannel noises, with 0, 5 and 10 dB input SNRs, are displayed in Figure 3.6. We can see that in all cases, the proposed weighting scheme provides the best results. It is worth noting that the perceptually-motivated benchmark algorithms showed a worse performance than using a constant weighting factor of $\eta = 2$, and tend to show similar quality to using $\eta = 0.1$. This is mainly due to the hard decision on the weighting factor such that $\eta_{kl} = 0$ for $\bar{p}_{kl}^N < T_{kl}^g$, which leads to $\hat{S}_{kl} = Y_{kl}$, i.e., the noise components are not reduced in such time-frequency bins. Therefore, it is verified through experiments that employing soft decision on the weighting factor, i.e., applying non-zero value on η_{kl} for $\bar{p}_{kl}^N < T_{kl}^g$, improves the enhancement performance. Similar results were also found for the Babble and Buccaneer 1 noises.

Regarding the benchmark algorithms in the second group and the proposed algorithms, the average results over all speakers of the three objective measures (i.e., PESQ, SDR and SSNR) are shown for each noise type, respectively, in Table 3.2 to 3.5. As it can be observed, the best enhancement results were obtained with the proposed RNG-WWF method for all the different noise types and input SNRs. Moreover, the RNG method generally provided better results than the benchmark algorithms except in specific cases, e.g., SSNR for the Factory 1 noise at 0 dB input SNR. Among the benchmark algorithms, the STSA and SSM which used no training data provided reasonable results for Babble and Factory 1 noises

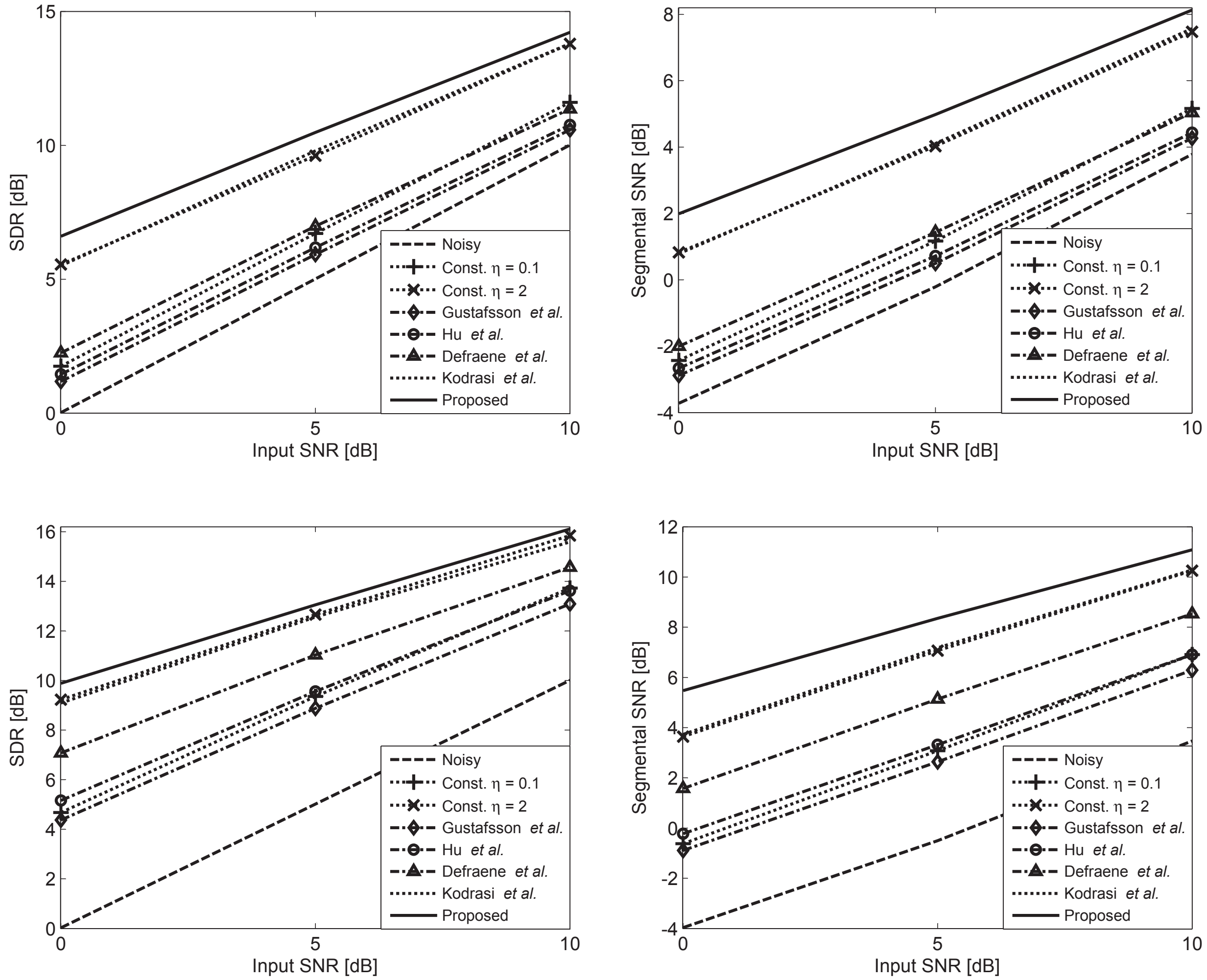

Fig. 3.6 SDR and SSNR comparisons for Factory 1 (top) and Hfchannel (bottom) noises.

compared to the NMF-based algorithms. However, they resulted in poorer performances for Buccaneer 1 and Hfchannel noises. Among the NMF-based benchmark algorithms, which used training data to obtain some prior knowledge of the clean speech and noise, it was found in general that the RNMF-AGM provided slightly better results compared to the NMF and WNMF methods (except in some cases, e.g., slightly better PESQ results using the WNMF method for the Buccaneer 1 and Factory 1 noises). If we only compare between the two proposed methods, the RNG-WWF method provided much better results

Table 3.2 Average results for Buccaneer 1 noise

Input SNR	Eval.	Noisy	STSA	SSM	NMF	WNMF	RNMF -AGM	RNG	RNG -WWF
0 dB	PESQ	1.25	1.58	1.61	1.79	1.83	1.81	1.98	**2.22**
	SDR	0.02	4.31	4.25	5.25	5.74	5.43	6.13	**7.92**
	SSNR	-3.97	-0.27	-0.56	0.13	1.15	0.28	1.79	**3.18**
5 dB	PESQ	1.54	1.94	1.99	2.18	2.21	2.20	2.35	**2.47**
	SDR	5.01	8.56	8.79	9.75	9.63	9.92	10.59	**11.38**
	SSNR	-0.49	2.79	2.78	3.58	4.07	3.75	4.40	**6.17**
10 dB	PESQ	1.89	2.32	2.39	2.53	2.55	2.55	2.64	**2.69**
	SDR	10.01	12.43	12.97	13.80	13.23	13.91	14.59	**14.85**
	SSNR	3.48	6.14	6.47	7.14	7.28	7.33	8.06	**9.19**

Table 3.3 Average results for Hfchannel noise

Input SNR	Eval.	Noisy	STSA	SSM	NMF	WNMF	RNMF -AGM	RNG	RNG -WWF
0 dB	PESQ	1.23	1.50	1.59	1.78	1.71	1.79	2.01	**2.30**
	SDR	0.03	7.11	7.62	7.32	6.97	7.51	8.31	**9.88**
	SSNR	-3.97	1.95	2.35	1.64	2.16	1.81	2.56	**5.46**
5 dB	PESQ	1.45	1.92	2.04	2.15	2.08	2.16	2.35	**2.51**
	SDR	5.02	10.80	11.66	11.50	10.85	11.66	12.37	**13.05**
	SSNR	-0.50	4.96	5.78	5.12	5.22	5.30	6.20	**8.35**
10 dB	PESQ	1.75	2.31	2.46	2.50	2.43	2.52	2.63	**2.70**
	SDR	10.01	14.12	15.19	15.12	14.44	15.22	15.91	**16.11**
	SSNR	3.47	7.91	9.03	8.58	8.48	8.74	9.67	**11.09**

than the RNG method, which further validates that using the proposed weighting factor improves the enhanced speech quality.

Figure 3.7 illustrates the magnitude spectra of clean, noisy and enhanced speech for several benchmark and proposed algorithms. In this particular example, a female speech is degraded with Buccaneer 1 noise at 0 dB input SNR. As we can see, the proposed RNG-WWF method could reduce the background noise significantly, and especially during the speech-absence periods where the noise is further reduced.

Informal listening tests were also conducted to compare the performance of the benchmark algorithms in the second group and the proposed algorithms. It was generally found that the latter, and especially the RNG-WWF method offered the best performance, both in terms of noise reduction and speech distortion. More specifically, the STSA and SSM

Table 3.4 Average results for Babble noise

Input SNR	Eval.	Noisy	STSA	SSM	NMF	WNMF	RNMF-AGM	RNG	RNG-WWF
0 dB	PESQ	1.52	1.68	1.62	1.77	1.72	1.78	1.81	**1.84**
	SDR	0.02	2.76	2.69	3.06	2.52	3.18	3.36	**4.55**
	SSNR	-3.48	-0.57	-0.65	-0.36	-0.34	-0.32	-0.29	**1.28**
5 dB	PESQ	1.86	2.05	2.02	2.16	2.11	2.17	2.20	**2.24**
	SDR	5.01	7.39	7.53	7.70	6.80	7.89	8.12	**8.53**
	SSNR	0.05	2.44	2.58	2.79	2.54	2.94	3.09	**4.06**
10 dB	PESQ	2.22	2.42	2.43	2.53	2.47	2.55	2.56	**2.59**
	SDR	10.01	11.52	11.90	11.53	10.38	11.73	12.17	**12.21**
	SSNR	4.05	5.84	6.23	5.91	5.66	6.16	6.66	**7.07**

Table 3.5 Average results for Factory 1 noise

Input SNR	Eval.	Noisy	STSA	SSM	NMF	WNMF	RNMF-AGM	RNG	RNG-WWF
0 dB	PESQ	1.36	1.68	1.66	1.74	1.80	1.76	1.80	**1.98**
	SDR	0.02	4.44	4.16	4.34	4.29	4.54	4.49	**6.60**
	SSNR	-3.72	0.28	0.17	-0.14	0.28	0.12	-0.10	**1.99**
5 dB	PESQ	1.70	2.09	2.10	2.15	2.18	2.16	2.19	**2.34**
	SDR	5.01	8.62	8.69	9.07	8.53	9.24	9.27	**10.48**
	SSNR	-0.21	3.21	3.34	3.33	3.19	3.53	3.42	**4.99**
10 dB	PESQ	2.07	2.45	2.50	2.53	2.52	2.54	2.54	**2.64**
	SDR	10.01	12.49	12.91	13.33	12.42	13.37	13.61	**14.22**
	SSNR	3.78	6.48	6.91	6.91	6.46	6.96	7.12	**8.13**

gave an enhanced speech with reasonable quality for the Babble and Factory 1 noises although some musical noise was found in the SSM method. However, they both failed to remove high frequency components in the Buccaneer 1 noise which resulted in a highly annoying ringing sound. The enhanced speech with the benchmark NMF algorithms, i.e., NMF, RNMF-AGM and WNMF, was perceived as being similar to that obtained with the STSA and SSM for Babble and Factory 1 noises, but of better quality for Buccaneer 1 and Hfchannel noises. Focusing on the proposed algorithms, the RNG method could remove more low frequency noise than the benchmark algorithms, whereas the high frequency components were further removed using the RNG-WWF method. Consequently, the enhanced speech using the RNG-WWF method was perceived as having much better quality than the one using the RNG method.

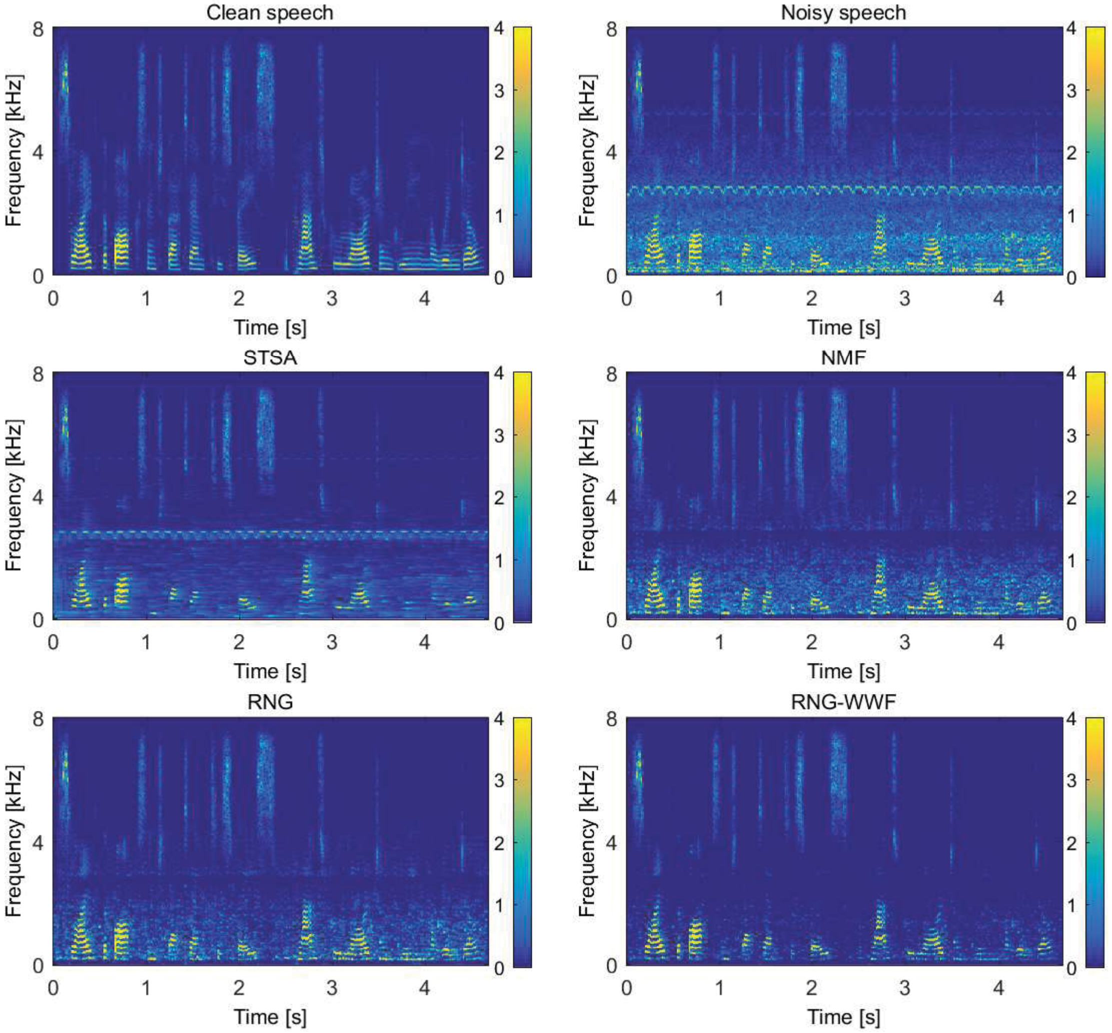

Fig. 3.7 Example of magnitude spectra of the clean, noisy and estimated clean speech for the benchmark and proposed algorithms. A female speech is degraded with Buccaneer 1 noise at 0 dB input SNR

3.5 Summary

New single-channel speech enhancement algorithms based on regularized NMF have been introduced in this Chapter. In the proposed framework, *a priori* knowledge about the magnitude spectra of the clean speech and noise is captured by distinct GMMs, where normalized spectra are employed to handle the magnitude difference between the training and test data. The corresponding LLFs are included as regularization terms in the NMF cost function during the enhancement stage. Further improvement of the enhance speech quality was obtained by exploiting the masking effects of the human auditory system. Specifically, we constructed a weighted Wiener filter where the weighting factor is selected based on the masking threshold calculated from the estimated clean speech PSD. In addition to informal listening tests and visual inspection of spectrograms, experimental results using three different objective measures (PESQ, SDR and SSNR) showed that the proposed speech enhancement algorithms could provide better performance than the benchmark algorithms for several types of noises and input SNRs.

Chapter 4

Training and Compensation of Class-conditioned NMF Bases

In this chapter, we introduce a training and compensation algorithm of the class-conditioned basis vectors in the NMF model for single-channel speech enhancement[1]. The main goal is to estimate the basis vectors of different signal sources in a way that prevents them from representing each other, in order to reduce the residual noise components that have features similar to the speech signal. During the proposed training stage, the basis matrices for the clean speech and noises are estimated jointly by constraining them to belong to different classes. To this end, we employ the PGM of classification, specified by class-conditional densities, as an *a priori* distribution for the basis vectors. The update rules of the NMF and the PGM parameters of classification are jointly obtained by using the VBEM algorithm, which guarantees convergence to a stationary point. Another goal of the proposed algorithm

[1] Parts of this chapter have been presented at the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing in Shanghai, China [100], and 2017 IEEE International Conference on Acoustics, Speech and Signal Processing in New Orleans, USA [102]; and have been publishted in the IEEE Signal Processing Letters [101], and Neurocomputing [103].

is to handle a mismatch between the characteristics of the training and test data. This is accomplished during the enhancement stage, where we implement a basis compensation scheme. Specifically, we use extra free basis vectors to capture the features which are not included in the training data. Objective experimental results for different combinations of speaker and noise types show that the proposed algorithm can provide better speech enhancement performance than the benchmark algorithms under various conditions.

This chapter is organized as follows. In Section 4.1, we address the research motivation and contributions of the proposed method. In Section 4.2, we introduce the PGMs of the NMF and classification models. The propose training stage is derived in Section 4.3, and the proposed enhancement stage is explained in Section 4.4. Experimental results are presented in Section 4.5.

4.1 Research Motivations and Contributions

In a supervised NMF-based framework, the basis vectors are typically obtained *a priori* for each source by independently using isolated training data during the training stage. However, there are two main problems in such a framework. The first one is that the basis vectors of the different signal sources, e.g., speech and noise, may share similar characteristics. For example, the basis vectors of the speech spectrum can represent the noise spectrum and hence, the enhanced speech may contain residual noise components which have features similar to the speech signal. One possible remedy is to train the basis vectors of each source in a way that prevents them from representing each other. In [104], the cross-coherence of the basis vectors is added as a penalty term to the NMF cost function, whereas the cross-reconstruction error terms are considered in [105]. The authors in [57–59] propose to use additional training data which are generated by mixing, e.g., adding or concatenating, the isolated training data of each source. However, the approaches in [57, 59] are based on heuristic MU rules which do not guarantee the convergence of the NMF in general [38, 60]. Moreover, the basis vectors in [58, 59] are obtained indirectly by means of the activation matrix estimated from the mixed training data and hence, lack an explicit interpretation in terms of discrimination.

The second problem in a supervised framework is the existence of a mismatch between the characteristics of the training and test data. As mentioned in Chapter 3, a common approach to overcome this problem is to add explicit regularization terms to the NMF cost function that incorporate some prior knowledge, such as the temporal continuity [54] or statistical characteristics of the magnitude spectra [80]. In these algorithms, however, the basis vectors are fixed during the enhancement stage, which limits the performance when there is a large mismatch between the training and test data. One alternative approach is to use a basis adaptation scheme during the enhancement stage. In [55], the basis vectors are adapted based on prior distributions modeled by gamma mixtures. The authors in [106] employ extra validation data for speaker adaptation in a speech-music separation task. In [56], the basis vectors are adapted by using a combination of the original and pre-processed noisy speech samples, the latter being obtained via a classical MMSE-based speech enhancement algorithm. In these algorithms, however, the basis vectors are adapted from the mixtures of multiple sources, e.g., noise and speech, such that the resulting basis vectors may still exhibit features of different sources. Consequently, the enhanced speech may contain some residual noise components and hence, adapting the complete set of basis vectors may limit the enhancement performance.

In this chapter, to overcome these limitations, we introduce a training and compensation algorithm of the class-conditioned basis vectors in the NMF model for single-channel speech enhancement, which is an extension of our earlier works on training class-conditioned basis vectors in [101], and basis compensation in [100]. In the proposed framework herein, we consider the PGM of classification specified by class-conditional densities [67], along with the NMF model [41]. Specifically, the PGM of classification is used as an explicit *a priori* distribution for the basis vectors. During the proposed training stage, the basis matrices for all the clean speech and noise sources are estimated jointly by constraining them to belong to one of several speech and noise classes. In our earlier work [101], we used a traditional Gaussian-distributed class-conditional density [67], and the model parameters were obtained through a MAP estimator using the EM algorithm. In this chapter, we make two key modifications. First, we employ a gamma-distributed class-conditional density to bring more coherence into the NMF model. Second, the update rules of the NMF model and the PGM parameters for classification are jointly obtained via the VBEM algorithm, which can be considered as an extension of the EM algorithm [41, 67, 107].

The proposed enhancement stage consists of two steps. First, we perform noise classification based on the posterior class probability (PCP), in order to determine which type of noise is included in the noisy speech. Second, we implement a basis compensation algorithm by adopting the approach in [100]. That is, we use extra free basis vectors for both the clean speech and noise to capture the features which cannot be explained by the limited set of basis vectors due to the hard decision on the noise type as well as features which are not included in the training data. The PGM parameters for classification are employed while inferring the free basis vectors as well as during the noise classification. Previously in [100], the free basis vectors were estimated by using the MU rules, whereas we use the VBEM algorithm.

4.2 Probabilistic Generative Models

In this section, we introduce two underlying PGMs for the proposed framework: the PGM of NMF, where the LLF corresponds to the KL-divergence, is described in Subsection 4.2.1, while the PGM of classification, which will be applied to the basis vectors, is presented in Subsection 4.2.2.

4.2.1 NMF model

We first briefly revisit the statistical interpretation of NMF, introduced in Section 2.1. The NMF model with KL-divergence is described within a statistical framework in [41]. Each entry of a non-negative matrix, $\mathbf{V} = [v_{kl}]$, is assumed to be a sum of M latent variables as

$$v_{kl} = \sum_{m=1}^{M} c_{kl}^{m}. \tag{4.1}$$

The m-th latent variable, c_{kl}^{m}, is assumed to be drawn from a Poisson distribution parameterized by w_{km} and h_{ml}

$$p(c_{kl}^{m}|w_{km}, h_{ml}) = \mathcal{P}(c_{kl}^{m}|w_{km}h_{ml}) \tag{4.2}$$

where $\mathcal{P}(c|u) = u^{c} \exp(-u)/(c!)$ is the Poisson distribution with mean u. Note that the approximation of v_{kl} as a sum of integer variables in (4.1) can be justified by assuming a large dynamic range for the former quantity, which in practice can be realized by a proper scaling of the magnitude spectra [34, 101, 109]. The ML estimates of the parameters w_{km} and h_{ml}, given the observation v_{kl}, are obtained via the EM algorithm, where the iterative solutions are given by (2.20) and (2.21).

4.2.2 Classification model

In the classification problem, the input vector $\mathbf{w} = [w_k] \in \mathbb{R}^K$ under test is assigned to one of I_C classes. The essential part of the classification is to find a partition of the observation space $\mathbb{R}^K$ into decision regions that will minimize the classification error, by using training data and their corresponding class labels. There are two main approaches to solve this problem: *PGM* and *discriminative modeling* [67, 110]. The former approach maximizes the likelihood based on the joint distribution of the input data and class labels, whereas the latter maximizes the PCP. In this work, we consider the PGM since it can provide the necessary *a priori* distributions to be used in the proposed training framework.

The PGM can be described by a class-conditional density based on a Gaussian distribution [67, 101] or a Gaussian mixture model [111]. In this work, we instead employ a gamma distribution, which is shown to be a conjugate prior to the Poisson model [41], to bring more coherence into the NMF model. By ignoring possible correlations between different entries in $\mathbf{w}$, the class-conditional density based on the gamma distribution can be expressed as

$$p(\mathbf{w}\,|d_i = 1) = \prod_{k=1}^{K} \mathcal{G}(w_k; \alpha^i_{w,k}, \beta_{w,k}) \tag{4.3}$$

where $\mathcal{G}(x; \alpha, \beta) = x^{\alpha-1}\beta^{-\alpha}\exp(-x/\beta)/\Gamma(\alpha)$ is the gamma distribution with mean $\alpha\beta$, $\Gamma(\cdot)$ is the gamma function, and α and β are referred to as the shape and scale parameters, respectively. Although we can use class-specific scales $\beta^i_{w,k}$, we consider a common value of $\beta_{w,k}$ for all classes [67], in order to avoid over-fitting.

For a given training set of $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_M]$ and $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_M]$, where $\mathbf{d}_m = [d_{im}]$ with $d_{im} \in \{0, 1\}$ (such that $\sum_i d_{im} = 1$) is an $I_C \times 1$ target class label vector, and assuming the columns $\mathbf{w}_m$ are independently drawn, the likelihood function is given by

$$p(\mathbf{W}, \mathbf{D}; \boldsymbol{\theta}_C) = \prod_{m=1}^{M} \prod_{i=0}^{I_C-1} \left[p(\mathbf{w}_m \,|d_i = 1)p_i\right]^{d_{im}} \tag{4.4}$$

where $\boldsymbol{\theta}_C = \{\{p_i, \{\alpha^i_{w,k}\}_{k=1}^K\}_{i=0}^{I_C-1}, \{\beta_{w,k}\}_{k=1}^K\}$ is a PGM parameter set for classification and $p_i \triangleq p(d_i = 1)$ is the prior class probability. The set $\boldsymbol{\theta}_C$ can be simply estimated via the ML criterion. Using Bayes' theorem, the PCP of class i, given the observation $\mathbf{w}$, can be expressed as

$$p(d_i = 1|\,\mathbf{w}) = \frac{p(\mathbf{w}\,|d_i = 1)p_i}{\sum_j p(\mathbf{w}\,|d_j = 1)p_j}. \tag{4.5}$$

4.3 Proposed Training Stage

In many applications of the EM algorithm, evaluating the posterior distribution or indeed computing expectations with respect to this distribution is analytically intractable. Consequently, it is highly demanding to derive a lower bound for the marginal likelihood of the observed data or to estimate the hyper-parameters. The VBEM algorithm overcomes this difficulty by computing an analytical and efficient approximation to the posterior distribution [67, 107], and also provides an effective estimation of the hyper-parameters. In general, the VBEM algorithm can be considered as an extension of the EM algorithm from the ML or MAP estimation of the single most probable value of each parameter to fully Bayesian estimation in which any unknown parameter is absorbed into the set of latent variables. We employ the VBEM method to develop the proposed training algorithm, as further explained below.

4.3.1 Prior structures

We first explicitly address the prior structures for the PGM in (4.2), which will be used in the proposed framework. We denote by M_i the number of basis vectors in class i (such that $M = \sum_i M_i$), and by L_i the number of time frames in class i. For the basis vectors, the likelihood function $p(\mathbf{W}, \mathbf{D}; \boldsymbol{\theta}_C)$ in (4.4), based on the class-conditional density given by (4.3), can be simply rearranged as

$$p(\mathbf{W}; \boldsymbol{\theta}_C) = \prod_{i=0}^{I_C-1} \prod_{m=1}^{M_i} \prod_{k=1}^{K} p_i \mathcal{G}(w_{km}^i; \alpha_{w,k}^i, \beta_{w,k}) \tag{4.6}$$

where we omit the dependence on $\mathbf{D}$ hereafter for convenience. For the activations, we follow the prior model based on the gamma distribution as introduced in [34, 41]:

$$p(h_{ml}^i; \alpha_{h,ml}^i, \beta_{h,ml}^i) = \mathcal{G}\left(h_{ml}^i; \alpha_{h,ml}^i, \frac{\beta_{h,ml}^i}{\alpha_{h,ml}^i} \right) \tag{4.7}$$

which provides an intuitive interpretation in terms of the mean value simply given by $\beta_{h,ml}^i$. Moreover, we consider constant values of $\alpha_{h,ml}^i$ and $\beta_{h,ml}^i$ for each class, i.e., $\alpha_{h,ml}^i = \alpha_h^i$ and $\beta_{h,ml}^i = \beta_h^i$, to avoid over-fitting [34, 41]. Assuming that the entries of $\mathbf{H}$ are independently distributed, the prior of $\mathbf{H}$ can be written as

$$p(\mathbf{H}; \boldsymbol{\alpha}_h, \boldsymbol{\beta}_h) = \prod_{i=0}^{I_C-1} \prod_{m=1}^{M_i} \prod_{l=1}^{L_i} p(h_{ml}^i; \alpha_h^i, \beta_h^i) \tag{4.8}$$

where $\boldsymbol{\alpha}_h = \{\alpha_h^i\}_{i=0}^{I-1}$ and $\boldsymbol{\beta}_h = \{\beta_h^i\}_{i=0}^{I-1}$. Note that employing the prior structure in (4.7) for the basis vectors specifies the class-specific scales in the PGM for classification and hence, limits the enhancement performance due to over-fitting.

4.3.2 VBEM algorithm

Let us denote by $\boldsymbol{\theta}_L = \{\mathbf{C}, \mathbf{W}, \mathbf{H}\}$ the set of latent variables, where $\mathbf{C} = \{c_{kl}^{m,i}\}$, $\mathbf{W} = \{w_{km}^{i}\}$, $\mathbf{H} = \{h_{ml}^{i}\}$, and by $\boldsymbol{\theta}_R = \{\boldsymbol{\theta}_C, \boldsymbol{\alpha}_h, \boldsymbol{\beta}_h\}$ the set of hyper-parameters. In the proposed framework, we use the class index $i = 0$ for the speech and $i = 1, ..., I_C - 1$ for the different noise types. For given training data sets of the clean speech and noise, $\mathbf{V} = \{\mathbf{V}^i\}$, the marginal LLF can be written as

$$\begin{aligned} \ln p(\mathbf{V}; \boldsymbol{\theta}_R) &\geq \sum_{\mathbf{C}} \int \int q(\mathbf{C}, \mathbf{W}, \mathbf{H}) \ln \frac{p(\mathbf{V}, \mathbf{C}, \mathbf{W}, \mathbf{H}; \boldsymbol{\theta}_R)}{q(\mathbf{C}, \mathbf{W}, \mathbf{H})} d\,\mathbf{W}\, d\,\mathbf{H} \\ &= \underbrace{\mathbb{E}_{q(\boldsymbol{\theta}_L)}[\ln p(\mathbf{V}, \boldsymbol{\theta}_L; \boldsymbol{\theta}_R)]}_{\triangleq\, \mathcal{L}_V(q(\boldsymbol{\theta}_L); \boldsymbol{\theta}_R)} - \underbrace{\mathbb{E}_{q(\boldsymbol{\theta}_L)}[\ln q(\boldsymbol{\theta}_L)]}_{\triangleq\, -\mathcal{L}_E(q(\boldsymbol{\theta}_L))} \\ &\triangleq \mathcal{L}_B(q(\boldsymbol{\theta}_L); \boldsymbol{\theta}_R) \end{aligned} \tag{4.9}$$

where $q(\cdot)$ is an arbitrary distribution (often referred to as a *variational distribution*) and $\mathbb{E}_{g(x)}[f(x)]$ indicates an expectation of $f(x)$ with respect to $g(x)$. The term $\mathcal{L}_B(q(\boldsymbol{\theta}_L); \boldsymbol{\theta}_R)$ defines the lower bound on $\ln p(\mathbf{V}; \boldsymbol{\theta}_R)$, where the equality holds for $q(\boldsymbol{\theta}_L) = p(\boldsymbol{\theta}_L \,|\, \mathbf{V}; \boldsymbol{\theta}_R)$ [41, 67]. A detailed expression of the lower bound is given in Appendix A.1. Analogous to the EM algorithm, the VBEM algorithm consists of two stages. During the E-step, the goal is to estimate $q(\boldsymbol{\theta}_L)$ which approximates the exact posterior distribution $p(\boldsymbol{\theta}_L \,|\, \mathbf{V}; \boldsymbol{\theta}_R)$. In the M-step, the hyper-parameters are obtained by maximizing the lower bound in (4.9) computed with a *fixed* $q(\boldsymbol{\theta}_L)$. That is, the term $\mathcal{L}_E(q(\boldsymbol{\theta}_L))$, which denotes the *entropy* of $q(\boldsymbol{\theta}_L)$, can be considered as a constant value and hence, maximizing the lower bound becomes equivalent to maximizing the *energy* $\mathcal{L}_V(q(\boldsymbol{\theta}_L); \boldsymbol{\theta}_R)$.

1) Variational E-step: Based on the *mean-field* approximation [67, 107], we assume that

$q(\mathbf{C}, \mathbf{W}, \mathbf{H})$ can be factorized as (e.g., [41, 43, 108])

$$q(\mathbf{C}, \mathbf{W}, \mathbf{H}) = q(\mathbf{C})q(\mathbf{W})q(\mathbf{H}) = \left(\prod_{i,k,l} q(\mathbf{c}_{kl}^i)\right)\left(\prod_{i,k,m} q(w_{km}^i)\right)\left(\prod_{i,m,l} q(h_{ml}^i)\right) \tag{4.10}$$

where $\mathbf{c}_{kl}^i = [c_{kl}^{1,i}, ..., c_{kl}^{M_i,i}]$. The resulting local optimal solutions can be found as [41, 67]:

$$q(\mathbf{C})^{(r+1)} \propto \exp\Big(\mathbb{E}_{q(\mathbf{W})^{(r)}q(\mathbf{H})^{(r)}}[\ln p(\mathbf{V}, \boldsymbol{\theta}_L; \boldsymbol{\theta}_R)]\Big) \tag{4.11}$$

$$q(\mathbf{W})^{(r+1)} \propto \exp\Big(\mathbb{E}_{q(\mathbf{C})^{(r+1)}q(\mathbf{H})^{(r)}}[\ln p(\mathbf{V}, \boldsymbol{\theta}_L; \boldsymbol{\theta}_R)]\Big) \tag{4.12}$$

$$q(\mathbf{H})^{(r+1)} \propto \exp\Big(\mathbb{E}_{q(\mathbf{C})^{(r+1)}q(\mathbf{W})^{(r+1)}}[\ln p(\mathbf{V}, \boldsymbol{\theta}_L; \boldsymbol{\theta}_R)]\Big) \tag{4.13}$$

where the superscript (r) denotes the r-th iteration. For convenience, we hereafter omit the superscript (r) and also drop the latent variables inside the subscript $q(\cdot)$ of the expectation operator, e.g., $\mathbb{E}_{q(w_{km}^i)}[w_{km}^i] = \mathbb{E}_q[w_{km}^i]$.

First, the distribution $q(\mathbf{c}_{kl}^i)$ in (4.11) is shown to be a multinomial distribution [41]:

$$\mathcal{M}(\mathbf{c}_{kl}^i; v_{kl}^i, \bar{\mathbf{p}}_{kl}^i) = \delta\left(v_{kl}^i - \sum_{m=1}^{M_i} c_{kl}^{m,i}\right) v_{kl}^i! \prod_{m=1}^{M_i} \frac{(\bar{p}_{kl}^{m,i})^{c_{kl}^{m,i}}}{c_{kl}^{m,i}!} \tag{4.14}$$

where $\delta(x)$ is the Kronecker delta function defined by $\delta(x) = 1$ when $x = 0$ and $\delta(x) = 0$ otherwise. The entries of $\bar{\mathbf{p}}_{kl}^i = [\bar{p}_{kl}^{m,i}]$ are given by

$$\bar{p}_{kl}^{m,i} = \frac{\exp\big(\mathbb{E}_q[\ln w_{km}^i] + \mathbb{E}_q[\ln h_{ml}^i]\big)}{\sum_{m=1}^{M_i} \exp\big(\mathbb{E}_q[\ln w_{km}^i] + \mathbb{E}_q[\ln h_{ml}^i]\big)}. \tag{4.15}$$

Next, the distribution $q(w_{km}^i)$ in (4.12) is obtained as

$$q(w_{km}^i) \;\; \propto \;\; \exp\left[\Big(\alpha_{w,k}^i + \sum_{l=1}^{L_i} \mathbb{E}_q[c_{kl}^{m,i}] - 1\Big) \ln w_{km}^i - \Big(\frac{1}{\beta_{w,k}} + \sum_{l=1}^{L_i} \mathbb{E}_q[h_{ml}^i]\Big) w_{km}^i\right]$$

$$\propto \;\; \mathcal{G}(w^i_{km}; \bar{\alpha}^i_{w,km}, \bar{\beta}^i_{w,km}) \tag{4.16}$$

where the parameters are given by

$$\bar{\alpha}^i_{w,km} = \alpha^i_{w,k} + \sum_{l=1}^{L_i} \mathbb{E}_q[c^{m,i}_{kl}], \;\; \bar{\beta}^i_{w,km} = \Big(\frac{1}{\beta_{w,k}} + \sum_{l=1}^{L_i} \mathbb{E}_q[h^i_{ml}]\Big)^{-1}. \tag{4.17}$$

Finally, the distribution $q(h^i_{ml})$ in (4.13) is also found to be a gamma distribution $\mathcal{G}(h^i_{ml}; \bar{\alpha}^i_{h,ml}, \bar{\beta}^i_{h,ml})$ [41], where the parameters are given by

$$\bar{\alpha}^i_{h,ml} = \alpha^i_h + \sum_{k=1}^{K} \mathbb{E}_q[c^{m,i}_{kl}], \;\; \bar{\beta}^i_{h,ml} = \Big(\frac{\alpha^i_h}{\beta^i_h} + \sum_{k=1}^{K} \mathbb{E}_q[w^i_{km}]\Big)^{-1}. \tag{4.18}$$

The sufficient statistics (expectations) are given below:

$$\mathbb{E}_q[c^{m,i}_{kl}] = v^i_{kl}\bar{p}^{m,i}_{kl} \tag{4.19}$$

$$\mathbb{E}_q[\ln w^i_{km}] = \Psi(\bar{\alpha}^i_{w,km}) + \ln \bar{\beta}^i_{w,km}, \;\; \mathbb{E}_q[w^i_{km}] = \bar{\alpha}^i_{w,km}\bar{\beta}^i_{w,km} \tag{4.20}$$

$$\mathbb{E}_q[\ln h^i_{ml}] = \Psi(\bar{\alpha}^i_{h,ml}) + \ln \bar{\beta}^i_{h,ml}, \;\; \mathbb{E}_q[h^i_{ml}] = \bar{\alpha}^i_{h,ml}\bar{\beta}^i_{h,ml} \tag{4.21}$$

where $\Psi(x) = d \ln \Gamma(x)/dx$ is the digamma function [41].

2) Variational M-step: The hyper-parameter set $\boldsymbol{\theta}_R$ is estimated by maximizing $\mathcal{L}_V(q(\boldsymbol{\theta}_L)^{(r+1)}; \boldsymbol{\theta}_R)$. Setting the partial derivative of $\mathcal{L}_V(q(\boldsymbol{\theta}_L)^{(r+1)}; \boldsymbol{\theta}_R)$ with respect to $\boldsymbol{\theta}_R$ to zero, the PGM parameters for classification, $\boldsymbol{\theta}_C$, are obtained as

$$\alpha^i_{w,k} \leftarrow \alpha^i_{w,k} - \frac{\Psi(\alpha^i_{w,k}) - \alpha^i_{qw}}{\Psi'(\alpha^i_k)} \tag{4.22}$$

$$\beta_{w,k} = \frac{\sum_{i=0}^{I-1}\sum_{m=1}^{M_i} \mathbb{E}_q[w^i_{km}]}{\sum_{i=0}^{I-1} M_i \alpha^i_{w,k}} \tag{4.23}$$

where $\alpha_{qw}^i = \sum_{m=1}^{M_i}(\mathbb{E}_q[\ln w_{km}^i] - \ln \beta_{w,k})/M_i$ and $\Psi'(x) = d\Psi(x)/dx$. The prior class probability is simply estimated by $p_i = M_i/M$. The hyper-parameters for the activations, $\boldsymbol{\alpha}_h$ and $\boldsymbol{\beta}_h$, are obtained as in [41]:

$$\alpha_h^i \leftarrow \alpha_h^i - \frac{\ln \alpha_h^i - \Psi(\alpha_h^i) + 1 - \alpha_{qh}^i}{1/\alpha_h^i - \Psi'(\alpha_h^i)} \tag{4.24}$$

$$\beta_h^i = \frac{1}{M_i L_i}\sum_{m=1}^{M_i}\sum_{l=1}^{L_i}\mathbb{E}_q[h_{ml}^i] \tag{4.25}$$

where $\alpha_{qh}^i = \sum_{m=1}^{M_i}\sum_{l=1}^{L_i}(\mathbb{E}_q[h_{ml}^i]/\beta_h^i - \mathbb{E}_q[\ln h_{ml}^i] + \ln \beta_h^i)/(M_i L_i)$.

The proposed training stage can be interpreted as follows. During the E-step, the basis vectors are adjusted based on their priors which define the classification boundaries. Hence, the basis vectors are estimated by constraining them to belong to different classes. During the M-step, the hyper-parameters (i.e., the PGM parameters for classification $\boldsymbol{\theta}_C$) are re-estimated, which define new classification boundaries.

4.3.3 Parameter initialization and normalization

For initialization, we generate positive random numbers and subsequently apply the MU rules in (2.5) to $\mathbf{V}$ for several iterations [34, 108], where we found that 10 iterations are sufficient. The resulting $\mathbf{W}^i$ and $\mathbf{H}^i$ are used as the initial values for the sufficient statistics, i.e., $\mathbb{E}_q[w_{km}^i]$, $\exp(\mathbb{E}_q[\ln w_{km}^i])$, $\mathbb{E}_q[h_{ml}^i]$ and $\exp(\mathbb{E}_q[\ln h_{ml}^i])$. To initialize $\boldsymbol{\theta}_C$, we apply (4.22) and (4.23) to the initial values of $\mathbb{E}_q[w_{km}^i]$ and $\mathbb{E}_q[\ln w_{km}^i]$. The hyper-parameters for the activations are initialized as $\alpha_h^i = 0.001$ and $\beta_h^i = 10$. We use 200 iterations for the VBEM algorithm, whereas 5 iterations are used for estimating the hyper-parameters in (4.22) and (4.24). To avoid scale indeterminacies in w_{km} and h_{ml} which appear as a product in the distribution (4.2), we include a normalization step. Motivated by [64], we normalize

$\mathbb{E}_q[w^i_{km}]$ and $\exp(\mathbb{E}_q[\ln w^i_{km}])$ such that they sum up to 1 with respect to k after computing (4.16).

4.4 Proposed Enhancement Stage

A number of attempts of combining the classical speech enhancement algorithms and the NMF-based framework have been made in the literature. In [56,100,112], a classical method is used as a pre-processor to first remove some stationary background noise, and the NMF-based algorithm is subsequently applied to further improve the enhancement performance. The authors in [113] implement the classical and NMF-based algorithms independently, and evaluate the geometric mean over them to estimate the clean speech spectrum. We combine both approaches and propose to use the weighted geometric mean (WGM) of the pre-processed noisy speech and its improvement via Wiener filtering. Regarding the pre-processor, we use the well-knwon MMSE short-time spectral amplitude (STSA) estimator [5], where the noise PSD is estimated based on [93]. The proposed enhancement stage consists of two steps[2], i.e., noise classification followed by basis compensation, which are explained in the following subsections. We denote by $\bar{\mathbf{S}}_{l_b} \in \mathbb{C}^{K \times L_b}$ the pre-processed noisy speech and by $\bar{\mathbf{N}}_{l_b} = \mathbf{Y}_{l_b} - \bar{\mathbf{S}}_{l_b}$ the pre-estimated noise.

4.4.1 Noise classification

In many NMF-based speech enhancement algorithms, the background noise type is assumed to be known *a priori*. In the proposed framework, we perform noise classification for the l_b-th mini-batch, to select a single noise type among different classes which has features similar to the noise included in the noisy speech. To this end, one possible approach is to

[2] In this chapter, we consider the mini-batch online approach, as explained in Footnote 1 in Subsection 2.2.1

apply the activation update given by (2.5) to $|\mathbf{Y}_{l_b}|$ for each noise type by fixing its corresponding basis matrix and observing the reconstruction error (i.e., KL-divergence), such as in [114]. However, this method requires additional iterations in which the computational cost increases with respect to the number of noise types.

In the proposed method, we use the PGM-based classifier given by (4.5). That is, we evaluate the PCP based on (4.5) and $\boldsymbol{\theta}_C$ for $i = \{1, ...I_C - 1\}$, and select the noise type with the highest PCP value. As a simple approach, we can first estimate a noise classification basis vector $\mathbf{w}_C = [w_k^C] \in \mathbb{R}_+^K$ by applying the MU rules in (2.5) to $|\bar{\mathbf{N}}_{l_b}|$, and use it as the input to the classifier. However, we can further reduce the computational cost by simply using the $|\bar{\mathbf{N}}_{l_b}|$ due to the property of NMF (i.e., the target matrix is represented as a linear combination of the basis vectors), since we can avoid additional iterations for computing $\mathbf{w}_C$. To further improve the classification performance, we consider both $\mathbf{Y}_{l_b}$ and $\bar{\mathbf{N}}_{l_b}$. That is, we compute the geometric mean of the magnitude spectra of the noisy speech and pre-estimated noise (i.e., $|\mathbf{Y}_{l_b} \otimes \bar{\mathbf{N}}_{l_b}|^{1/2} \in \mathbb{R}^{K \times L_b}$), to amplify the noise components. Subsequently, we average over the rows and normalize the resulting column vector using the l_1-norm, where the corresponding vector will be denoted by $\tilde{\mathbf{w}}_C \in \mathbb{R}_+^K$.

Regarding the classifier, we found that employing the gamma distribution in (4.3) directly for computing the PCP resulted in poor classification performance. One main reason is that the gamma distribution can lead to numerical instability, since $\Gamma(\alpha)$ rapidly approaches infinity as α increases. Hence, we instead use the approximated Gaussian distribution[3] as the class-conditional density, which is indeed simpler to compute than the gamma distribution:

$$p(\tilde{\mathbf{w}}_C | d_i = 1) \approx \mathcal{N}(\tilde{\mathbf{w}}_C; \tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Sigma}}_i) \tag{4.26}$$

[3] Note that this approximation is employed only for the noise classification. The inference on $q(w_{km}^i)$ does not suffer from the extreme value of the gamma function, i.e., the extreme value of the digamma function $(-\infty)$ appearing in $\mathbb{E}_q[\ln(\cdot)]$ in (4.20) and (4.21) is handled by the exponential in (4.15).

where $\tilde{\boldsymbol{\mu}}_i = [\tilde{\mu}_{ik}]$ and $\tilde{\boldsymbol{\Sigma}}_i = \mathrm{diag}\{\tilde{\sigma}_{ik}^2\}$ are the mean vector and diagonal covariance matrix of the Gaussian distribution with entries $\tilde{\mu}_{ik} = \alpha_{w,k}^i \beta_{w,k}$ and $\tilde{\sigma}_{ik}^2 = \alpha_{w,k}^i \beta_{w,k}^2$. The underlying motivation for using the form in (4.26) is similar to the application of the Laplace approximation [67], which aims at finding a Gaussian approximation to the original distribution. According to this approach, the mean and variance of the approximated Gaussian distribution are obtained based on the mode and second order derivative at the mode of the original distribution, respectively. However, since the mode of the gamma distribution is defined only for $\alpha > 1$, we instead use its mean and variance. Furthermore, we use the average value of $\tilde{\boldsymbol{\Sigma}}_i$ over all i for the covariance in (4.26), which leads to computing the (exponential of the squared) Mahalanobis distance. The latter is known to further reduce the computational cost compared to using the Gaussian model with class-specific variances [70].

4.4.2 Basis compensation

Once the noise type is determined, we implement a basis compensation scheme by adopting the approach proposed in [100]. That is, we use extra free basis vectors for both the clean speech and noise to capture the features which cannot be explained by the limited set of basis vectors due to the hard decision on a single noise type, as well as features which are not included in the training data. We denote by $\mathbf{W}_{l_b}^{SF} = [w_{km}^{SF}] \in \mathbb{R}_+^{K \times M_{SF}}$ and $\mathbf{W}_{l_b}^{NF} = [w_{km}^{NF}] \in \mathbb{R}_+^{K \times M_{NF}}$ (such that $M_{SF} < M_S$ and $M_{NF} < M_N$) the free basis matrices of the clean speech and noise, respectively.

For the l_b-th mini-batch, motivated by [56] and [100], we aim at factorizing $\mathbf{V}_{l_b} = [|\,\mathbf{Y}_{l_b}\,|\,|\bar{\mathbf{S}}_{l_b}|] \in \mathbb{R}_+^{K \times 2L_b}$ into the product of $\mathbf{W}_{l_b} = [\mathbf{W}_S \;\; \mathbf{W}_{l_b}^{SF} \;\; \mathbf{W}_N \;\; \mathbf{W}_{l_b}^{NF}] = [w_{km}] \in \mathbb{R}_+^{K \times M_Y}$ and $\mathbf{H}_{l_b} = [\mathbf{H}_{l_b}^{Y} \;\; \mathbf{H}_{l_b}^{\bar{S}}] = [h_{ml'}] \in \mathbb{R}_+^{M_Y \times 2L_b}$, where $M_Y = M_S + M_{SF} + M_N + M_{NF}$. We use the VBEM algorithm introduced in Subsection 4.3.2, to estimate the variational distributions

$q(\mathbf{W}^{SF}_{l_b})$, $q(\mathbf{W}^{NF}_{l_b})$ and $q(\mathbf{H}_{l_b})$. At each iteration, the distribution $q(\mathbf{C})$ is first inferred as (4.14), where the parameters are given by (4.15). Second, we estimate the parameters of $q(\mathbf{W}^{SF}_{l_b})$ and $q(\mathbf{W}^{NF}_{l_b})$, while fixing the parameters of $q(\mathbf{W}_S)$ and $q(\mathbf{W}_N)$. Specifically, the parameters of $q(w^{SF}_{km})$ and $q(w^{NF}_{km})$, which correspond to the ones in $q(w_{km})$ for the intervals $M_S < m \leq M_S + M_{SF}$ and $M_S + M_{SF} + M_N < m \leq M_Y$, respectively, are computed based on (4.17). The parameters of $q(\mathbf{H}_{l_b})$ are then simply obtained by using (4.18). Subsequently, the parameter of the noisy speech activation prior β_{h,l_b} is obtained by

$$\beta_{h,l_b} = \frac{\sum_{m=1}^{M_Y} \sum_{l'=1}^{2L_b} \alpha_{h,ml'}\, \mathbb{E}_q[h_{ml'}]}{\sum_{m=1}^{M_Y} \sum_{l'=1}^{2L_b} \alpha_{h,ml'}} \tag{4.27}$$

where $\alpha_{h,ml} = \alpha^S_h$ for $1 \leq m \leq M_S + M_{SF}$ and $\alpha_{h,ml} = \alpha^N_h$ for $M_S + M_{SF} + 1 \leq m \leq M_Y$. In contrast to the β_{h,l_b}, we fix the shape parameters of the clean speech and noise, α^S_h and α^N_h, which controls the degree of sparsity [41], mainly in order to reduce the computational cost since their updates require additional iterations as given by (4.24).

After estimating $q(\mathbf{W}^{SF}_{l_b})$, $q(\mathbf{W}^{NF}_{l_b})$ and $q(\mathbf{H}_{l_b})$, we compute the smoothed PSDs of the clean speech and noise based on (2.32) and (2.33), where the periodograms are obtained from the mean values[4] of $q(\mathbf{W}_{l_b})$ and $q(\mathbf{H}_{l_b})$. Specifically, the mini-batch clean speech PSD, $\hat{\mathbf{P}}^S_{l_b} = [\hat{p}^S_{kl}] \in \mathbb{R}^{K \times L_b}_+$, is computed by replacing $\mathbf{W}_S$ with $[\mathbb{E}_q[\mathbf{W}_S]\ \mathbb{E}_q[\mathbf{W}^{SF}_{l_b}]] \in \mathbb{R}^{K \times (M_S + M_{SF})}_+$ and $\mathbf{H}^S_{l_b}$ with the first $M_S + M_{SF}$ rows of $\mathbb{E}_q[\tilde{\mathbf{H}}_{l_b}] = (\mathbb{E}_q[\mathbf{H}^Y_{l_b}] + \mathbb{E}_q[\mathbf{H}^{\bar{S}}_{l_b}])/2 \in \mathbb{R}^{M_Y \times L_b}_+$. A similar procedure is carried out for the mini-batch noise PSD $\hat{\mathbf{P}}^N_{l_b} = [\hat{p}^N_{kl}] \in \mathbb{R}^{K \times L_b}_+$. Then, we estimate the clean speech spectrum where the magnitude is obtained via the WGM of $|\bar{\mathbf{S}}_{l_b}|$ and Wiener-filtered $|\bar{\mathbf{S}}_{l_b}|$, and the phase is taken from the noisy speech. Since

[4] Alternatively, based on [34], we can compute the smoothed PSD based on the sufficient statistics of $c^{m,i}_{kl}$ in (4.19) where $\bar{p}^{m,i}_{kl}$ is given by (4.15). However, we verified through experiments that using $\mathbb{E}_q[w^i_{km}]$ provided better enhancement performance as well as reduced complexity.

$\angle \mathbf{Y}_{l_b} = \angle \bar{\mathbf{S}}_{l_b}$ [5], the enhanced speech spectrum can be written as

$$\hat{\mathbf{S}}_{l_b} = \left(\left|\bar{\mathbf{S}}_{l_b}\right|^{\nu_{l_b}} \otimes \left| \frac{\hat{\mathbf{P}}^S_{l_b}}{\hat{\mathbf{P}}^S_{l_b} + \hat{\mathbf{P}}^N_{l_b}} \otimes \bar{\mathbf{S}}_{l_b} \right|^{1-\nu_{l_b}} \right) \otimes e^{j\angle \mathbf{Y}_{l_b}} = \left(\frac{\hat{\mathbf{P}}^S_{l_b}}{\hat{\mathbf{P}}^S_{l_b} + \hat{\mathbf{P}}^N_{l_b}} \right)^{1-\nu_{l_b}} \otimes \bar{\mathbf{S}}_{l_b} \tag{4.28}$$

where $0 \leq \nu_{l_b} \leq 1$ is the weighting factor. The motivation of using the WGM is to control the effect of pre-processing. For a high input SNR, for instance, the classical method tends to show a reasonable enhancement performance, which implies that Wiener filtering the pre-processed signal may further distort the enhanced speech quality. Hence, it is necessary to put more weight on $\bar{\mathbf{S}}_{l_b}$ by selecting a large ν_{l_b}. In contrast, the classical method results in a poor enhanced speech quality for a low input SNR and hence, further improvement is necessary. This can be specified by applying more weight on the Wiener-filtered $\bar{\mathbf{S}}_{l_b}$ by selecting a small ν_{l_b}. Based on these considerations, we use the logistic function for selecting ν_{l_b}:

$$\nu_{l_b} = \frac{\rho_1}{1 + \exp(-\rho_2 R_{l_b})} \tag{4.29}$$

where $R_{l_b} = 10 \log_{10}(\sum_k \sum_l \hat{p}^S_{kl} / \sum_k \sum_l \hat{p}^N_{kl})$ is the estimated input SNR in dB for the l_b-th mini-batch. The parameters ρ_1 and ρ_2 respectively define the range of $\nu_{l_b} \in (0, \rho_1)$ and the slope of the sigmoid function, where we use $\rho_1 = \rho_2 = 0.5$ through the experiments.

For the l_b-th mini-batch, the parameters of $q(\mathbf{W}^{NF}_{l_b})$ are initialized by applying the NMF algorithm to $|\bar{\mathbf{N}}_{l_b}|$ for 2 iterations. Specifically, since $M_{NF} > L_b$ (i.e., over-complete), we use the sparse NMF algorithm which is simply implemented by adding the sparsity parameter (we use 0.5) to the denominator of the activation update in (2.5). In contrast, the parameters of $q(\mathbf{W}^{SF}_{l_b})$ are initialized from the ones estimated in the previous mini-batch frame index. The parameters of $q(\mathbf{H}_{l_b})$ are initialized by generating positive random numbers. We use 5 iterations for the VBEM algorithm.

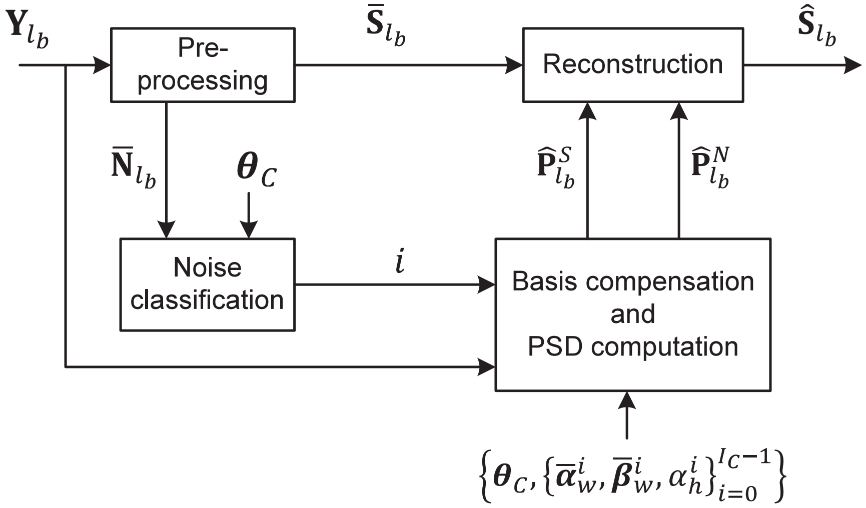
$\mathbf{Y}_{l_b}$
Pre-processing
$\bar{\mathbf{S}}_{l_b}$
Reconstruction
$\hat{\mathbf{S}}_{l_b}$
$\bar{\mathbf{N}}_{l_b}$
$\boldsymbol{\theta}_C$
$\hat{\mathbf{P}}^S_{l_b}$
$\hat{\mathbf{P}}^N_{l_b}$
Noise classification
i
Basis compensation and PSD computation
$\left\{\boldsymbol{\theta}_C, \{\bar{\boldsymbol{\alpha}}^i_w, \bar{\boldsymbol{\beta}}^i_w, \alpha^i_h\}_{i=0}^{I_C-1}\right\}$

Fig. 4.1 A simplified block diagram of the proposed VNCP-BC method.

The proposed algorithm, i.e., variational inference on the NMF model based on class probabilities and basis compensation, will be referred to as VNCP-BC. A simplified block diagram of the proposed method is illustrated in Figure 4.1, while the algorithm is summarized in Table 4.1. Recall that the terms $\bar{\boldsymbol{\alpha}}^i_w = [\bar{\alpha}^i_{w,km}] \in \mathbb{R}^{K \times M_i}$ and $\bar{\boldsymbol{\beta}}^i_w = [\bar{\beta}^i_{w,km}] \in \mathbb{R}^{K \times M_i}$ represent the parameters of the variational distribution in (4.16), and the sets $\boldsymbol{\theta}_C$ and $\{\alpha^i_h\}$ respectively denote the PGM parameters for classification and the shape parameters in the activation prior.

4.5 Experiments

The enhancement performance of the proposed method was assessed through objective experiments. Below, after describing the general methodology and benchmark algorithms, we present and discuss the experimental results.

Table 4.1 Algorithm summary of the proposed enhancement stage

```
for $l_b = 1, 2, ...$
    Estimate $\bar{\mathbf{S}}_{l_b}$ and $\bar{\mathbf{N}}_{l_b} = \mathbf{Y}_{l_b} - \bar{\mathbf{S}}_{l_b}$
    if $l_b = 1$
        Initialize $\hat{p}^S_{k,0} = \sum_l |\bar{S}_{kl}|^2 / L_b$ and $\hat{p}^N_{k,0} = \sum_l |\bar{N}_{kl}|^2 / L_b$
        Initialize $q(\mathbf{W}^{SF}_{l_b - 1})$ parameters by applying sparse NMF to $|\bar{\mathbf{S}}_{l_b}|$
    end
    Compute $\tilde{\mathbf{w}}_C$ by averaging and normalizing $| \mathbf{Y}_{l_b} \otimes \bar{\mathbf{N}}_{l_b}|^{1/2}$
    Select noise type $i \in \{1, ..., I_C - 1\}$ via (4.5) and (4.26)
    Initialize $q(\mathbf{W}^{SF}_{l_b})$ parameters by the one estimated at $l_b - 1$
    Initialize $q(\mathbf{W}^{NF}_{l_b})$ parameters by applying sparse NMF to $|\bar{\mathbf{N}}_{l_b}|$
    Initialize $q(\mathbf{H}_{l_b})$ parameters by generating positive random numbers
    for iter = 1:itermax
        Estimate $q(\mathbf{W}^{SF}_{l_b})$ and $q(\mathbf{W}^{NF}_{l_b})$ and normalize
        Estimate $q(\mathbf{H}_{l_b})$
        Update $\beta_{h,l_b}$ via (4.27)
    end
    Compute $\hat{\mathbf{P}}^S_{l_b} = [\hat{p}^S_{kl}]$ and $\hat{\mathbf{P}}^N_{l_b} = [\hat{p}^N_{kl}]$
    Compute $\nu_{l_b}$ via (4.29) and estimate $\hat{\mathbf{S}}_{l_b}$ via (4.28)
end
```

4.5.1 Methodology

We conducted the experiments using the 4th CHiME challenge corpus [115]. The speech and noise files were divided into two disjoint groups: i) *training data*, used for estimating the basis matrix for each class i during the training stage, and ii) *test data*, used during the enhancement stage to evaluate the enhancement performance. The clean speech training data of the CHiME database are from the Wall Street Journal (WSJ0) corpus, which consists of 101 speakers. We considered a speaker-independent (SI) application, where one *universal* basis matrix covering all speakers is estimated during the training stage. To this end, we randomly selected 40 utterances from each speaker and concatenated them to construct the clean speech training data ($i = 0$), resulting in a total of 8 hours long signal. Regarding the noise training data, we selected the Bus ($i = 1$), Pedestrian ($i = 2$) and Street ($i = 3$) noises, where each noise type consists of 2 hours long signal.

We used the reference clean speech from the test set of the CHiME corpus, which consists of 330 utterances. Regarding the test data for the noise signals, we categorized them into two groups, referred to as: *matched* and *mismatched* cases. The matched case assumes that the training data is available, whereas the purpose of the mismatched case is to evaluate the enhancement performance for an *unseen* noise type, i.e., when no training data is available. For both the matched and mismatched cases, we performed noise classification to select a single noise type which has characteristics similar to the actual noise included in the noisy speech.

We considered two types of the noisy speech signals for the test: *additive noise* and *filtered noisy speech*. The noisy speech signals for the former type were generated by scaling and adding the noise to the reference clean speech signal to obtain input SNRs of -5, 0, 5, and 10 dB. The filtered test set, provided by the CHiME organization (referred to as “simulated test data”), contains the noisy speech signals which were generated by artificially mixing the clean speech and noises. Specifically, the clean speech signals were filtered by the impulse responses (IR) between the speaker and microphone, estimated from the real recorded signals and hence, the filtered data exhibit a more realistic nature of the noisy speech (see [115] for more details about the database).

For both the additive and filtered data types, we considered the Bus ($i = 1$), Pedestrian ($i = 2$) and Street ($i = 3$) noises for the matched noise case and used the Cafe noise from the CHiME database for the mismatched noise case. Regaring the additive noise, we additionally selected the Factory 1 and Babble noises from the NOISEX database [97] for

Table 4.2 Summary of the test noise types

	Additive	Filtered
Matched	Bus, Pedestrian, Street (from CHiME)	
Mismatched	Cafe (from CHiME), Factory 1, Babble (from NOISEX)	Cafe (from CHiME)

the mismatched noise case. The sampling rate of all signals was set to 16 kHz. The noise types used for the test are summarized in Table 4.2.

Regarding the implementation, a Hanning window of 512 samples with 50% overlap was employed for the STFT analysis. We used $M_i = 60$ (for all i) and $M_{SF} = M_{NF} = 20$ basis vectors. The values of $(\tau_S, \tau_N) = (0.4, 0.9)$ were chosen as the temporal smoothing factors in (2.32) and (2.33). We used $L_b = 16$ for the mini-batch size. For the pre-processor, the value of 0.9 was used as the smoothing factor in the decision-directed (DD) method for the *a priori* SNR estimation in [5], whereas 0.85 was used as the smoothing factor for the noise PSD estimation in [93]. Regarding the shape parameters for the activation α_h^i, we obtained values around 0.02 using the training data (similar results were found when using different initial values, e.g., $\alpha_h^i = 0.1$). Although we can use such values during the enhancement stage, we found that instead using larger values resulted in slightly better enhancement performance, where we ultimately chose $\alpha_h^S = 0.1$ and $\alpha_h^N = 0.2$ in the experiments. The reason for this phenomenon can be explained as follows. The basis vectors in the proposed framework are estimated within a restricted decision boundary for each class, which may prevent them from properly representing the target magnitude spectrum. This becomes severe when the number of sources increases (i.e., resulting in smaller decision regions) and hence, may further limit the enhancement performance. Fortunately, the extra free basis vectors can handle such limitation by supporting the class-conditioned basis vectors to better represent the target observation $\mathbf{V}_{l_b}$. In particular, for a given class i, it is necessary to relax the dependency of the free basis vectors on their prior distribution so that they are able to be estimated beyond the decision boundaries. This can be specified by lowering the degree of sparsity of the activations, which corresponds to using a larger value of α_h^i [41].

We considered the PESQ [98], SDR [99] and SSNR as the objective measures of performance. The PESQ attempts to predict overall perceptual quality in MOS and the SDR measures the overall quality of the enhanced speech in dB by considering both the aspects of speech distortion and noise reduction. For all the measures, a higher value indicates a better result.

4.5.2 Benchmark algorithms

To evaluate the enhancement performance of the proposed VNCP-BC method, we implemented several benchmark algorithms, which are summarized below. Basic settings, such as the STFT analysis and synthesis, the mini-batch size L_b and the reconstruction method, were kept identical when applicable.

1) MMSE-STSA estimator: We implemented the MMSE-STSA estimator [5], where the noise PSD was estimated based on [93]. A smoothing factor of 0.85 in the DD method and 0.9 in the noise PSD estimation were used.

2) NMF: The standard NMF algorithm based on KL-divergence introduced in Chapter 2 was evaluated.

3) NMF model with distinct basis vectors: Among several NMF algorithms aiming at estimating the distinct basis vectors, we implemented two algorithms as representative benchmarks. The first one estimates the basis vectors based on the cross-coherence penalty (NCC) which is presented in [104]. The second one is our earlier work in [101], i.e., the NMF model based on class probabilities (NCP), where the class-conditioned basis vectors are obtained via the MAP estimator. A brief summary of the NCP method is given in

Appendix B.

4) NMF with basis compensation (NBC): The NMF algorithm with basis compensation proposed in [100] was evaluated, as a representative benchmark among several NMF algorithms proposed for handling the mismatch problem. We examined the NBC method with three different types of basis vectors, i.e., obtained via the conventional NMF, NCC and NCP methods. We used identical settings for the pre- and post-processing as in the proposed VNCP-BC method.

5) Bayesian NMF model (BNMF): To compare with a VBEM-based NMF algorithm, We implemented the BNMF method in [41]. The difference with the proposed VNCP (-BC) method is that the BNMF method estimates the basis matrix for each source independently as in the typical supervised NMF-based framework, whereas the proposed method estimates the basis matrices for all sources jointly.

In addition to the above mentioned benchmarks, we implemented the proposed method without employing the free basis vectors and pre-processing, which will be referred to as VNCP.

We used $M_i = 80$ basis vectors for all NMF-based benchmark algorithms (including the VNCP method) except for the NBC method, where we used $M_i = 60$ and $M_{SF} = M_{NF} = 20$. Hence, the same total number of basis vectors was employed for fair comparison. To perform the noise classification for the benchmark algorithms, we estimated the set $\boldsymbol{\theta}_C$ based on the Gaussian-distributed class-conditional density [67, 101]. For the NMF, NBC and BNMF methods, we first estimated the basis vectors for each class i, then we applied the ML criterion to the basis vectors [101]. The set $\boldsymbol{\theta}_C$ for the NCP method was jointly obtained with the NMF parameters. The noise classification was performed by following a strategy similar to the one introduced in Subsection 4.4.1. Note that the pre-processing was performed only for the noise classification in the NMF, NCP and VNCP methods, since these methods do not employ the pre-processed noisy speech during the reconstruction.

4.5.3 Results

Figure 4.2 shows the PCPs of the estimated basis vectors $\mathbb{E}_q[\mathbf{w}_m^i]$ (which will be simply denoted by $\mathbf{w}_m^i$). The x-axis indicates the m-th column vector of the matrix $[\mathbf{W}^0, ..., \mathbf{W}^3] = [\mathbf{w}_m]$, where each submatrix $\mathbf{W}^i$ consists of 80 basis vectors, i.e., $M_i = 80$ for all i. For each class i, the PCP values $p(d_i = 1|\,\mathbf{w}_m)$ should be close to one for the interval $iM_i + 1 \leq m \leq (i+1)M_i$, whereas the PCPs for the other intervals should be close to zero. Regarding the class $i = 0$, for example, the PCPs $p(d_0 = 1|\,\mathbf{w}_m)$ for the interval $1 \leq m \leq 80$ should be close to one, whereas the PCPs for the interval $81 \leq m \leq 320$ should be close to zero. We can see that the basis vectors are estimated to be distinct in terms of the PCP in general (although $p(d_2 = 1|\,\mathbf{w}_m)$ for the interval $1 \leq m \leq 80$ tend to be close to one since the Pedestrian noise contains a lot of speech components), which implies that the basis vectors of each source will be less likely to represent each other.

Figure 4.3 shows an example of the noise classification results using the method introduced in Subsection 4.4.1. In this particular example, a male speech signal was degraded with a noise at 0 dB input SNR. Specifically, the noise was generated by concatenating the Bus ($i = 1$), Street ($i = 3$) and Pedestrian ($i = 2$) noises where each noise signal was 3 seconds in duration. As we can see, the noise type is well estimated. The magnitude spectra of the clean speech, noisy speech and the enhanced speech using the proposed VNCP-BC method, for this particular example, are illustrated in Figure 4.4. As it can be observed, the background noise has been significantly reduced.

The average results over all utterances for the additive noises are shown in Tables 4.3 to 4.8, where the values in bold indicate the best performance along the corresponding row. Most of all, we can see that the proposed VNCP-BC method provided better enhance-

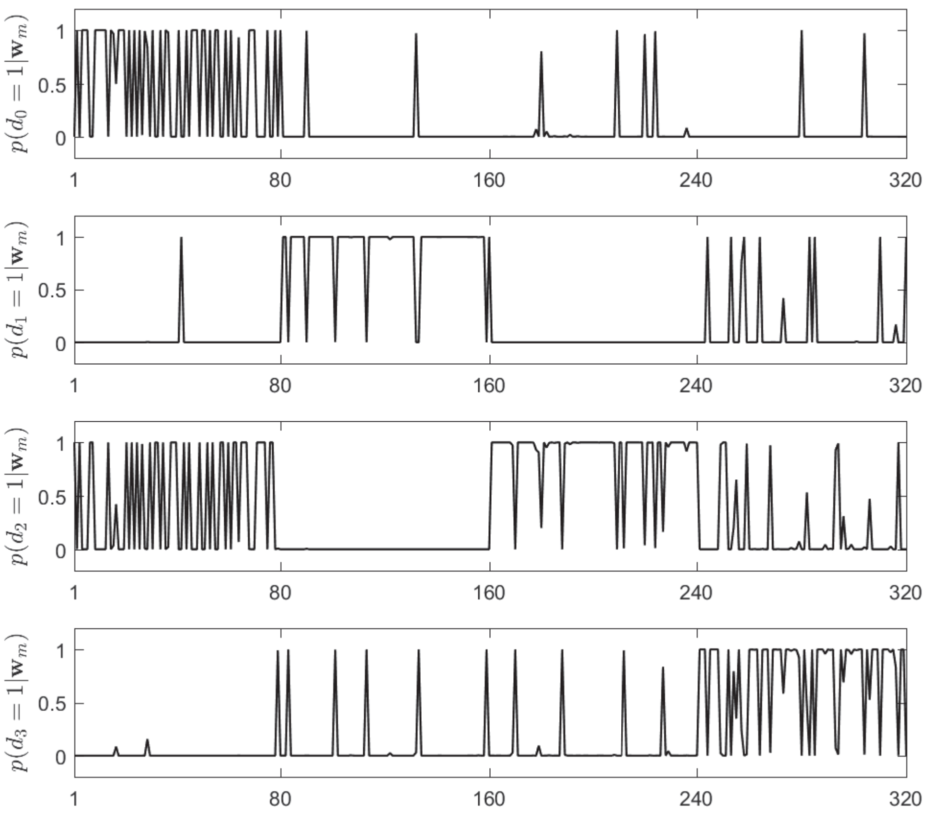

Fig. 4.2 The posterior class probabilities $p(d_i = 1|\, \mathbf{w}_m)$.

ment performance than the benchmark algorithms in general for both the matched and mismatched noise cases. Specifically, the proposed VNCP-BC method resulted in better performance compared to using the algorithms introduced in our previous works, i.e., the NCP and NBC methods. Moreover, the VNCP-BC method provided better results than the VNCP method, which further validates that implementing the basis compensation scheme improves the performance.

Regarding the matched noises, the results of the VBEM-based VNCP method were found to be better than the MAP-based NCP method. Comparing between the VBEM-based methods, the class-conditioned model-based VNCP method exhibited better performance than the independent source training-based BNMF method in general, whereas the

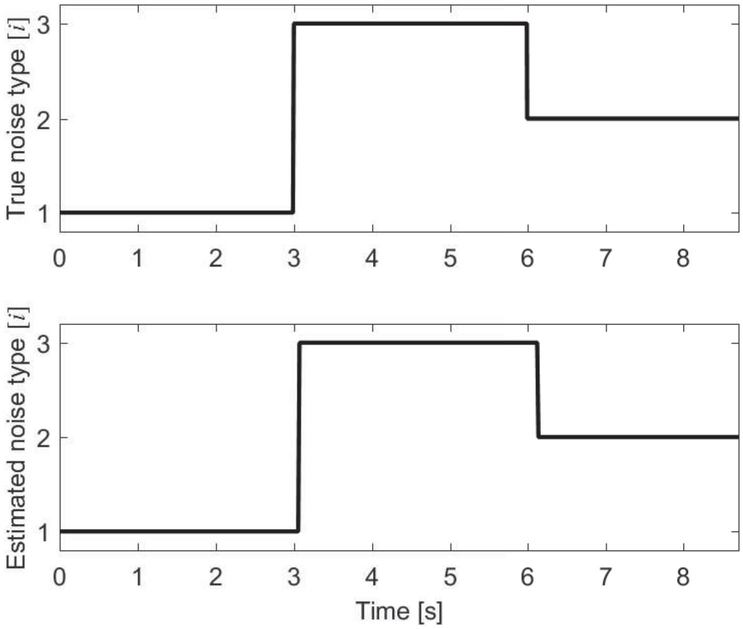

Fig. 4.3 An example of noise classification. Top shows the true noise type and bottom shows the estimated noise type using the proposed method.

BNMF method provided slightly better results for the Pedestrian noise. Among the NBC methods with different types of basis vectors, we can see that using the basis vectors obtained via the NCP method provided better results. We also conducted experiments for all benchmarks and proposed algorithms assuming that the noise type is known *a priori*, for the matched noise case. Although we do not report their objective results in this thesis, we have seen that there were no significant differences with the results obtained by including the noise classification. That is, the results increased by about 0.01 in PESQ and SDR for all methods when assuming that the noise type is known *a priori*.

The effectiveness of using the basis compensation scheme can be better verified from the results of the mismatched noises. In general, we can see that some NMF-based benchmark algorithms showed even worse performance than using the STSA estimator, whereas the NBC-based methods provided reasonable results. Specifically, although the NBC methods gave acceptable SDR and SSNR values for the Cafe and Babble noises under low input

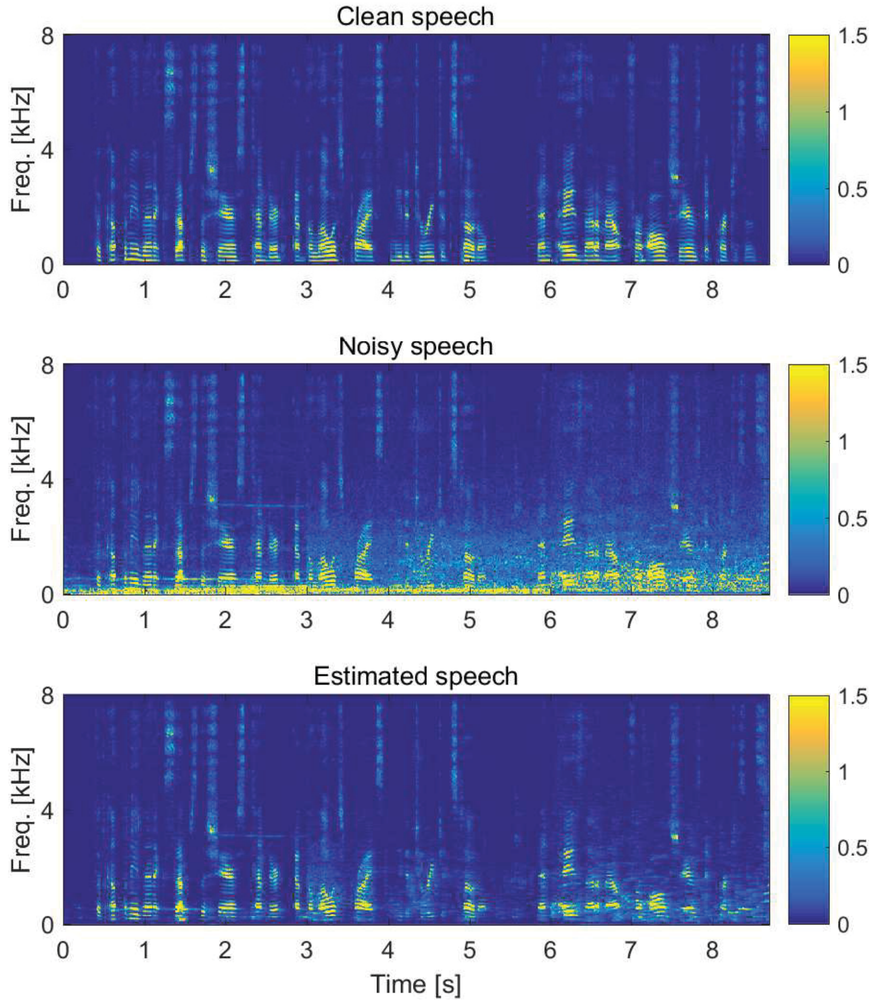

Fig. 4.4 Examples of magnitude spectrograms of the clean, noisy and estimated clean speech using the VNCP-BC method. A male speech is degraded by a noise consisting of different types as shown in Figure 4.3, at 0 dB input SNR.

SNRs, the proposed VNCP-BC method exhibited better than all benchmark algorithms in most cases.

The average results over all utterances for the filtered data set are shown in Table 4.9. Although the results showed slightly different pattern from the additive noise case (e.g., the STSA estimator gave even better results than some of the benchmarks for the Pedestrian noise), mainly due to the effect of the IR-filtered clean speech, we can see that the proposed

Table 4.3 Average results for additive Bus noise (matched)

Input SNR	Eval.	Noisy	STSA	NMF	NCC	NCP	NBC -NMF	NBC -NCC	NBC -NCP	BNMF	VNCP	VNCP -BC
-5 dB	PESQ	1.83	2.08	2.07	2.08	2.08	2.17	2.17	2.16	2.11	2.11	**2.27**
	SDR	-4.89	0.17	2.83	2.63	2.80	6.60	6.50	6.80	3.44	3.70	**7.54**
	SSNR	-13.57	-7.10	-4.50	-4.58	-4.49	-1.24	-1.22	-1.00	-3.54	-3.29	**-0.29**
0 dB	PESQ	2.20	2.43	2.41	2.42	2.42	2.49	2.49	2.48	2.42	2.42	**2.57**
	SDR	0.05	5.25	7.70	7.51	7.67	10.39	10.33	10.51	8.05	8.28	**11.13**
	SSNR	-8.56	-2.75	-0.78	-0.85	-0.85	1.57	1.58	1.73	0.06	0.25	**2.30**
5 dB	PESQ	2.55	2.76	2.74	2.74	2.74	2.78	2.78	2.77	2.74	2.74	**2.87**
	SDR	5.03	9.99	11.96	11.79	11.98	13.62	13.56	13.55	12.48	12.64	**14.27**
	SSNR	-3.56	1.43	2.38	2.31	2.21	4.12	4.13	4.20	3.62	3.74	**4.75**
10 dB	PESQ	2.90	3.07	3.04	3.04	3.05	3.04	3.04	3.03	3.06	3.06	**3.15**
	SDR	10.03	14.33	15.05	15.04	15.19	16.22	16.12	15.96	16.51	16.58	**17.07**
	SSNR	1.45	5.54	4.87	4.84	4.65	6.36	6.39	6.34	6.86	6.91	**7.09**

Table 4.4 Average results for additive Pedestrian noise (matched)

Input SNR	Eval.	Noisy	STSA	NMF	NCC	NCP	NBC -NMF	NBC -NCC	NBC -NCP	BNMF	VNCP	VNCP -BC
-5 dB	PESQ	1.22	1.34	1.33	1.35	1.33	1.30	1.33	1.32	1.36	1.37	**1.39**
	SDR	-4.88	-3.61	-4.19	-3.96	-3.93	-3.55	-3.44	-3.58	-3.71	-3.76	**-3.18**
	SSNR	-13.88	-9.02	-9.22	-9.21	-9.25	-6.51	-6.53	-6.50	-9.21	-9.27	**-5.97**
0 dB	PESQ	1.51	1.70	1.70	1.71	1.70	1.73	1.75	1.74	1.75	1.76	**1.86**
	SDR	0.06	1.95	1.11	1.31	1.33	2.09	2.16	2.05	1.94	1.90	**2.76**
	SSNR	-8.87	-4.70	-4.78	-4.80	-4.82	-2.67	-2.65	-2.66	-4.22	-4.28	**-1.70**
5 dB	PESQ	1.85	2.09	2.09	2.10	2.09	2.13	2.15	2.14	2.16	2.16	**2.26**
	SDR	5.04	7.07	6.02	6.18	6.26	6.86	6.94	6.85	7.37	7.31	**7.84**
	SSNR	-3.87	-0.43	-0.80	-0.86	-0.86	0.92	0.95	0.90	0.85	0.76	**1.93**
10 dB	PESQ	2.20	2.44	2.46	2.48	2.46	2.45	2.47	2.45	2.54	2.53	**2.61**
	SDR	10.03	11.88	10.00	10.16	10.28	10.69	10.76	10.59	**12.36**	12.25	12.30
	SSNR	1.14	3.87	2.47	2.38	2.39	4.02	4.05	3.91	5.11	4.99	**5.20**

Table 4.5 Average results for additive Street noise (matched)

Input SNR	Eval.	Noisy	STSA	NMF	NCC	NCP	NBC -NMF	NBC -NCC	NBC -NCP	BNMF	VNCP	VNCP -BC
-5 dB	PESQ	1.39	1.68	1.63	1.64	1.64	1.84	1.86	1.86	1.77	1.81	**2.06**
	SDR	-4.89	-0.35	0.76	1.06	0.89	4.07	3.80	4.72	4.05	4.58	**7.11**
	SSNR	-13.72	-6.80	-6.11	-6.05	-6.16	-2.73	-2.89	-2.40	-3.11	-2.65	**-0.21**
0 dB	PESQ	1.67	2.02	1.98	1.99	1.98	2.20	2.21	2.21	2.10	2.14	**2.40**
	SDR	0.05	4.87	5.77	6.06	5.91	8.37	8.18	8.83	8.32	8.71	**10.30**
	SSNR	-8.72	-2.61	-1.97	-1.89	-2.02	0.42	0.36	0.65	0.43	0.72	**2.10**
5 dB	PESQ	2.00	2.37	2.35	2.36	2.36	2.52	2.53	2.53	2.44	2.47	**2.67**
	SDR	5.03	9.63	10.17	10.43	10.37	11.83	11.81	12.13	12.32	12.55	**13.27**
	SSNR	-3.72	1.43	1.58	1.69	1.54	3.30	3.38	3.47	3.76	3.85	**4.31**
10 dB	PESQ	2.36	2.70	2.71	2.72	2.72	2.77	2.79	2.78	2.76	2.77	**2.92**
	SDR	10.03	14.03	13.49	13.76	13.80	14.46	14.59	14.59	16.06	16.15	**16.18**
	SSNR	1.29	5.41	4.39	4.57	4.38	5.67	5.94	5.80	6.66	6.63	**6.69**

VNCP-BC method provided the best results for all types of noises. Hence, it is verified that the proposed VNCP-BC method performs well under a more realistic environment.

Table 4.6 Average results for additive Cafe noise (mismatched)

Input SNR	Eval.	Noisy	STSA	NMF	NCC	NCP	NBC -NMF	NBC -NCC	NBC -NCP	BNMF	VNCP	VNCP -BC
-5 dB	PESQ	1.30	1.38	1.38	1.38	1.38	1.29	1.32	1.32	1.37	1.37	**1.39**
	SDR	-4.89	-3.40	-2.98	-2.99	-2.78	-2.22	-2.23	**-1.86**	-3.29	-3.20	-1.89
	SSNR	-14.48	-10.93	-10.08	-10.19	-9.75	**-7.99**	-8.37	-8.09	-10.68	-10.53	-8.01
0 dB	PESQ	1.56	1.68	1.69	1.70	1.71	1.67	1.69	1.68	1.70	1.71	**1.76**
	SDR	0.06	2.07	2.13	2.17	2.37	3.32	3.33	3.61	2.29	2.36	**4.07**
	SSNR	-9.47	-6.26	-5.56	-5.65	-5.27	-3.74	-4.06	-3.90	-5.82	-5.70	**-3.26**
5 dB	PESQ	1.87	2.00	2.03	2.04	2.06	2.04	2.05	2.04	2.09	2.10	**2.15**
	SDR	5.04	7.18	6.88	6.93	7.13	7.99	8.07	8.25	7.91	7.95	**9.16**
	SSNR	-4.47	-1.72	-1.46	-1.54	-1.25	0.06	-0.11	-0.09	-0.44	-0.40	**1.02**
10 dB	PESQ	2.20	2.35	2.37	2.39	2.41	2.39	2.40	2.38	2.50	2.51	**2.53**
	SDR	10.03	11.96	10.72	10.81	10.94	11.63	11.85	11.80	13.10	13.08	**13.31**
	SSNR	0.54	2.74	1.96	1.91	2.08	3.35	3.34	3.21	4.55	4.51	**4.68**

Table 4.7 Average results for additive Factory 1 noise (mismatched)

Input SNR	Eval.	Noisy	STSA	NMF	NCC	NCP	NBC -NMF	NBC -NCC	NBC -NCP	BNMF	VNCP	VNCP -BC
-5 dB	PESQ	1.23	1.44	1.33	1.34	1.33	1.47	1.49	1.48	1.40	1.41	**1.57**
	SDR	-4.90	-1.44	-3.33	-2.59	-3.17	0.56	0.51	0.83	-1.12	-0.92	**2.07**
	SSNR	-14.33	-8.33	-9.82	-9.22	-9.76	-5.36	-5.49	-5.13	-8.48	-8.56	**-4.15**
0 dB	PESQ	1.50	1.77	1.67	1.68	1.67	1.84	1.86	1.85	1.73	1.74	**1.97**
	SDR	0.05	3.96	1.92	2.64	2.12	5.53	5.51	5.73	4.21	4.32	**7.16**
	SSNR	-9.32	-4.05	-5.31	-4.75	-5.25	-1.81	-1.87	-1.64	-3.95	-4.04	**-0.56**
5 dB	PESQ	1.83	2.12	2.05	2.06	2.04	2.20	2.22	2.21	2.11	2.12	**2.34**
	SDR	5.03	8.83	6.68	7.34	6.90	9.52	9.57	9.59	9.44	9.51	**11.13**
	SSNR	-4.32	0.14	-1.32	-0.83	-1.28	1.50	1.53	1.58	0.71	0.56	**2.63**
10 dB	PESQ	2.18	2.48	2.43	2.44	2.42	2.51	2.53	2.51	2.50	2.50	**2.67**
	SDR	10.03	13.40	10.42	10.94	10.72	12.38	12.56	12.35	14.08	14.19	**14.52**
	SSNR	0.68	4.40	1.88	2.27	1.90	4.28	4.41	4.27	4.88	4.68	**5.44**

Table 4.8 Average results for additive Babble noise (mismatched)

Input SNR	Eval.	Noisy	STSA	NMF	NCC	NCP	NBC -NMF	NBC -NCC	NBC -NCP	BNMF	VNCP	VNCP -BC
-5 dB	PESQ	1.33	1.45	1.44	1.44	1.44	1.40	1.43	1.43	1.46	1.46	**1.53**
	SDR	-4.89	-2.72	-3.75	-3.67	-3.68	-1.94	-1.91	-1.71	-3.95	-3.97	**-1.63**
	SSNR	-14.26	-9.90	-10.85	-10.90	-10.82	**-7.53**	-7.85	-7.70	-12.06	-12.09	-8.09
0 dB	PESQ	1.63	1.79	1.78	1.78	1.78	1.77	1.78	1.79	1.80	1.80	**1.90**
	SDR	0.05	2.77	1.48	1.55	1.54	3.54	3.50	3.67	1.45	1.43	**4.44**
	SSNR	-9.25	-5.39	-6.17	-6.17	-6.12	-3.49	-3.69	-3.61	-6.83	-6.87	**-3.15**
5 dB	PESQ	1.96	2.12	2.14	2.15	2.14	2.12	2.13	2.14	2.19	2.19	**2.29**
	SDR	5.03	7.78	6.47	6.58	6.51	7.97	8.01	8.11	7.21	7.14	**9.54**
	SSNR	-4.24	-1.02	-1.84	-1.78	-1.81	0.18	0.12	0.10	-1.19	-1.29	**1.39**
10 dB	PESQ	2.31	2.46	2.50	2.50	2.50	2.45	2.46	2.46	2.58	2.57	**2.64**
	SDR	10.03	12.41	10.68	10.92	10.69	11.27	11.45	11.43	12.80	12.69	**13.38**
	SSNR	0.77	3.31	1.82	1.94	1.80	3.32	3.38	3.24	4.26	4.11	**4.97**

A comparison of the computational times for the enhancement stage of the various methods is reported in Table 4.10. Specifically, the table lists the total time needed (in

Table 4.9 Average results for filtered noisy speech

Input SNR	Eval.	Noisy	STSA	NMF	NCC	NCP	NBC -NMF	NBC -NCC	NBC -NCP	BNMF	VNCP	VNCP -BC
BUS (mat.)	PESQ	1.70	1.97	1.94	1.95	1.94	2.05	2.06	2.05	1.98	2.00	**2.16**
	SDR	-1.34	2.79	3.62	4.18	4.00	6.45	6.39	6.66	5.48	5.53	**7.98**
	SSNR	-10.75	-7.35	-6.61	-6.36	-6.44	-4.75	-4.88	-4.67	-5.63	-5.65	**-3.58**
PED. (mat.)	PESQ	1.50	1.72	1.67	1.67	1.67	1.76	1.78	1.76	1.72	1.72	**1.86**
	SDR	0.13	3.26	1.47	1.58	0.89	4.33	4.37	4.41	2.27	2.29	**5.40**
	SSNR	-10.58	-7.54	-7.32	-7.30	-7.36	-5.48	-5.60	-5.53	-7.12	-7.17	**-4.54**
STR. (mat.)	PESQ	1.51	1.76	1.73	1.74	1.74	1.85	1.86	1.85	1.81	1.82	**2.00**
	SDR	-1.76	2.08	1.77	2.10	1.98	4.69	4.69	4.96	3.64	3.67	**6.45**
	SSNR	-10.81	-7.40	-6.96	-6.89	-6.92	-5.10	-5.30	-5.02	-5.90	-5.94	**-3.64**
CAF. (mis.)	PESQ	1.52	1.71	1.68	1.69	1.67	1.72	1.74	1.73	1.72	1.72	**1.83**
	SDR	-0.18	2.54	1.02	0.80	0.74	3.41	3.52	3.56	2.13	2.03	**4.73**
	SSNR	-10.64	-7.80	-7.48	-7.57	-7.53	-5.84	-5.99	-5.91	-7.37	-7.43	**-4.84**

Table 4.10 Comparison of computational times (in seconds)

STSA	NMF	NBC-NMF	BNMF	VNCP-BC
0.09	0.29	0.64	0.30	0.68

seconds) to process an 8.15 seconds long noisy speech file with the corresponding algorithm when implemented in MATLAB and running on a 3.4 GHz CPU with 32 GB RAM. Note that some of the methods use identical implementation for their enhancement stage and so, we only report the processing times of representative methods. In particular: NMF includes NCC and NCP; NBC-NMF includes NBC-NCC and NBC-NCP; while BNMF includes VNCP. The STSA estimator showed the shortest running time, since it requires no iterations. The NBC and VNCP-BC methods took more time than the NMF, NCP and BNMF methods due to the additional computation of the free basis vectors.

Besides the experiments using the CHiME database, we also conducted additional experiments using the clean speech from the TSP database [95] and the noise from the NOISEX database [97]. The main purpose was to see whether the classification-based VNCP and VNCP-BC methods limits the enhancement performance when we increase the number of noise classes (from 3 to 8). Although we do not report the objective results in this thesis, we were able to verify that the VNCP-BC method still provided the best enhancement performance in general.

4.6 Summary

We introduced a training and compensation algorithm of the class-conditioned basis vectors in the NMF model for single-channel speech enhancement. We considered the PGM for both the NMF and classification models. The former is specified by a Poisson observation model, whereas the latter is specified by gamm class-conditional densities, which are used as *a priori* distribution for the basis vectors. During the training stage, the basis matrices for the clean speech and noises were estimated jointly by constraining them to belong to different classes. The parameters of the NMF model and PGM of classification were obtained by using the VBEM algorithm, which guarantees convergence to a stationary point. During the enhancement stage, we performed a noise classification followed by a basis compensation. The latter was implemented by using extra free basis vectors to capture features which are not included in the training data. The PGM parameters for classificaion were employed while estimating the free basis vectors as well as during the noise classification. Experiments showed that the proposed VNCP-BC method provided better enhancement performance than the benchmark algorithms in general.

Chapter 5

Multi-channel Extension of Bayesian NMF Model

In this chapter, we introduce a supervised multi-channel speech enhancement algorithm based on a Bayesian MNMF model. In the proposed framework, we consider the PGM of MNMF, specified by Poisson-distributed latent variables and gamma-distributed priors. During the proposed training stage, the MNMF parameters of different speech and noise sources are estimated from the tensor-based training data via the VBEM algorithm. During the enhancement stage, the clean speech point source signal is estimated via the MNMF-based MVDR beamforming technique, whose realization involves two main steps. First, the speech source location is determined by observing the output powers of a DS beamformer applied to the MNMF-based pre-processed noisy speech signal. Second, the noise correlation matrix is computed using the MNMF parameters for the magnitude components, and a combination of the noisy speech phase and steering vector for the phase components. Experimental results for different combinations of speaker and noise types show that the proposed Bayesian algorithm can provide better speech enhancement performance than the benchmark algorithms.

This chapter is organized as follows. In Section 5.1, we state the research motivation and explain the significance of the proposed method. The training stage of the new algorithm is derived in Section 5.2, whereas the proposed enhancement stage is explained in Section 5.3. Experimental results are presented and discussed in Section 5.4.

5.1 Research Motivations and Contributions

Numerous NMF-based multi-channel speech enhancement and source separation algorithms have been introduced. In [33], the authors derive both the MU rules and EM algorithm for estimating the MNMF parameters, based on the IS-divergence. To better exploit the spatial properties of the sources, the authors in [61] aim at factorizing the SCM of the observation in each frequency bin, which is specified by the channel covariance matrices of the individual sources. The extended SCM, formulated as a weighted superposition of multiple DoA kernels (i.e., differential steering matrices), is proposed in [62]. In [43], the complex-valued Gaussian-distributed latent variables are modeled by auto-regressive moving average (ARMA) processes to better handle reverberation effects in realistic environments. A joint localization and enhancement method, based on the probabilistic SRP model specified by the DoA, is presented in [35]. However, a main issue with the above algorithms (or approaches) is the computational complexity of their implementation. That is, the computational cost increases rapidly as the number of NMF basis vectors, microphones or the dimension of the search grid for the speaker localization increase. The authors in [77] apply a single-channel NMF algorithm to the beamformer output as a post-processor. Although this approach is computationally efficient, it employs a classical MVDR beamforming technique which limits performance due to the poor estimation of the noise correlation matrix and source localization parameters.

As an alternative to MNMF, the NTF framework has been introduced in [69], where the authors derived the MU rules for the KL and IS divergences. In [116], the authors derive the EM algorithm for estimating the NTF parameters, based on the IS-divergence. A Bayesian NTF algorithm for stereo source separation has been introduced in [117]. However, the NTF model employs frequency-independent mixing coefficients, which implies that it can handle a linear instantaneous signal mixture and inadequate to represent the convolutive effects specified by the ATFs. In contrast, the MNMF model, which employs frequency-dependent mixing coefficients, does not suffer from this limitation.

In this chapter, we introduce a novel supervised multi-channel speech enhancement algorithm based on a Bayesian MNMF model. We consider the PGM of MNMF that corresponds to the KL-divergence within a statisical framework, as specified by Poisson-distributed latent variables and gamma-distributed priors. During the proposed training stage, the MNMF parameters of different sources are estimated from the tensor-based training data via the VBEM algorithm, which can be considered as an extension of the EM algorithm [41, 67, 107]. Specifically, compared to using the complex-valued Gaussian-distributed PGM of MNMF [43], one main advantage of using the Poisson-distributed PGM is that we can reduce the computational cost, since we only need the marginal statistics while implementing the VBEM algorithm [41].

During the proposed enhancement stage, the clean speech point source signal is estimated via the MNMF-based MVDR beamforming technique, whose realization involves two main steps. First, the clean speech and noise locations are determined by observing the spatial output power of a low-complexity DS beamformer applied to the MNMF-based pre-processed noisy speech signal. Second, the noise correlation matrix is computed using the MNMF parameters for the magnitude components, and a combination of the noisy speech phase and steering vector for the phase components.

5.2 Proposed Training Stage

In this section, we first explain the PGM of MNMF that corresponds to KL-divergence. Then, we introduce the explicit prior distributions for the PGM, which will be used in the proposed framework. Finally, the VBEM algorithm for estimating the latent variables and hyper-parameters is presented.

5.2.1 PGM of multi-channel NMF with KL-divergence

For a given tensor $\mathbf{V} = [v_{kl}^{j}] \in \mathbb{R}_{+}^{K \times L \times J}$, the MNMF algorithm aims at factorizing it into a mixing matrix $\mathbf{A} = [a_{k}^{j}] \in \mathbb{R}_{+}^{K \times J}$, a basis matrix $\mathbf{W} = [w_{km}] \in \mathbb{R}_{+}^{K \times M}$ and an activation matrix $\mathbf{H} = [h_{ml}] \in \mathbb{R}_{+}^{M \times L}$. Specifically, it seeks to represent each entry of $\mathbf{V}$ in the form of [33]

$$v_{kl}^{j} \approx \hat{v}_{kl}^{j} = a_{k}^{j} \sum_{m=1}^{M} w_{km} h_{ml}. \tag{5.1}$$

From a statistical perspective, each entry of $\mathbf{V}$ can be constructed as a sum of M latent variables

$$v_{kl}^{j} = \sum_{m=1}^{M} c_{kl}^{m,j}. \tag{5.2}$$

According to [41], the m-th latent variable $c_{kl}^{m,j}$ can be assumed to be drawn from a Poisson distribution parameterized by a_{k}^{j}, w_{km} and h_{ml}. That is:

$$p(c_{kl}^{m,j} | a_{k}^{j}, w_{km}, h_{ml}) = \mathcal{P}(c_{kl}^{m,j} | a_{k}^{j} w_{km} h_{ml}) \tag{5.3}$$

where $\mathcal{P}(x|u) = u^x \exp(-u)/(x!)$ is the Poisson distribution with mean u. Assuming that v_{kl}^j are independently drawn, the LLF of $\mathbf{V}$ can be written as[1]

$$\begin{aligned}\ln p(\mathbf{V} \,|\, \mathbf{A}, \mathbf{W}, \mathbf{H}) &= \ln \prod_{j=1}^{J} \prod_{k=1}^{K} \prod_{l=1}^{L} \mathcal{P}(v_{kl}^j | \hat{v}_{kl}^j) \\ &= \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} \left(v_{kl}^j \ln \hat{v}_{kl}^j - \hat{v}_{kl}^j - \ln v_{kl}^j! \right).\end{aligned} \tag{5.4}$$

We can see that maximizing the LLF with respect to the mixing, basis and activation elements is equivalent to minimizing the KL-divergence given by

$$\mathcal{D}_{KL}(\mathbf{V}, \hat{\mathbf{V}}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} \left(v_{kl}^j \ln \frac{v_{kl}^j}{\hat{v}_{kl}^j} - v_{kl}^j + \hat{v}_{kl}^j \right) \tag{5.5}$$

where the entries of the tensor $\hat{\mathbf{V}} = [\hat{v}_{kl}^j] \in \mathbb{R}_+^{K \times L \times J}$ are given by (5.1).

5.2.2 Prior structures

Regarding the prior distributions for the MNMF parameters $\mathbf{A}$, $\mathbf{W}$ and $\mathbf{H}$, we consider the gamma distribution, which is shown to be the conjugate prior to the Poisson distribution [41]. Specifically, we employ the following priors based on [41]:

$$p(a_k^j; \alpha_{a,k}^j, \beta_{a,k}^j) = \mathcal{G}\left(a_k^j; \alpha_{a,k}^j, \frac{\beta_{a,k}^j}{\alpha_{a,k}^j} \right) \tag{5.6}$$

$$p(w_{km}; \alpha_{w,km}, \beta_{w,km}) = \mathcal{G}\left(w_{km}; \alpha_{w,km}, \frac{\beta_{w,km}}{\alpha_{w,km}} \right) \tag{5.7}$$

$$p(h_{ml}; \alpha_{h,ml}, \beta_{h,ml}) = \mathcal{G}\left(h_{ml}; \alpha_{h,ml}, \frac{\beta_{h,ml}}{\alpha_{h,ml}} \right) \tag{5.8}$$

[1] We note that the sum of independent Poisson random variables x_m with means μ_m is another Poisson random variable with mean $\sum_m \mu_m$.

where $\mathcal{G}(x; a, b) = x^{a-1} b^{-a} \exp(-x/b)/\Gamma(a)$ is the gamma distribution with mean ab, $\Gamma(\cdot)$ is the gamma function, and a and b are referred to as the shape and scale parameters, respectively. Note that the parameterization of the gamma distributions given by (5.6), (5.7) and (5.8) provide an appealing interpretation in terms of the mean of the distribution given by $\beta^j_{a,k}$, $\beta_{w,km}$ and $\beta_{h,ml}$, respectively. In the proposed framework, we consider constant values for the hyper-parameters for each type of matrix factor (e.g., $\alpha^j_{a,k} = \alpha_a$ and $\beta^j_{a,k} = \beta_a$ for all k and j), to avoid over-fitting [34, 41]. Assuming that the entries of $\mathbf{A}$, $\mathbf{W}$ and $\mathbf{H}$ are independently distributed, the priors can be written as

$$p(\mathbf{A}; \alpha_a, \beta_a) = \prod_{k=1}^{K} \prod_{j=1}^{J} \mathcal{G}\left(a^j_k; \alpha_a, \frac{\beta_a}{\alpha_a}\right) \tag{5.9}$$

$$p(\mathbf{W}; \alpha_w, \beta_w) = \prod_{k=1}^{K} \prod_{m=1}^{M} \mathcal{G}\left(w_{km}; \alpha_w, \frac{\beta_w}{\alpha_w}\right) \tag{5.10}$$

$$p(\mathbf{H}; \alpha_h, \beta_h) = \prod_{m=1}^{M} \prod_{l=1}^{L} \mathcal{G}\left(h_{ml}; \alpha_h, \frac{\beta_h}{\alpha_h}\right). \tag{5.11}$$

5.2.3 VBEM algorithm

Let us denote by $\boldsymbol{\theta}_L = \{\mathbf{C}, \mathbf{A}, \mathbf{W}, \mathbf{H}\}$ the set of latent variables, and by $\boldsymbol{\theta}_R = \{\alpha_a, \beta_a, \alpha_w, \beta_w, \alpha_h, \beta_h\}$ the set of hyper-parameters. For a set of training data represented by $\mathbf{V} = [v^j_{kl}] \in \mathbb{R}_+^{K \times L \times J}$, which consists of the magnitude spectral coefficients of the image source, our goal is to estimate the latent variables and hyper-parameters via the VBEM algorithm. Note that, since we consider the magnitude spectra as entries of tensor $\mathbf{V}$, the mixing coefficients a^j_k represent the magnitude value of the ATFs. The marginal LLF can be written as

$$\begin{aligned}
\ln p(\mathbf{V};\boldsymbol{\theta}_R) &\geq \sum_{\mathbf{C}} \iiint q(\mathbf{C},\mathbf{A},\mathbf{W},\mathbf{H}) \ln \frac{p(\mathbf{V},\mathbf{C},\mathbf{A},\mathbf{W},\mathbf{H};\boldsymbol{\theta}_R)}{q(\mathbf{C},\mathbf{A},\mathbf{W},\mathbf{H})} d\,\mathbf{A}\, d\,\mathbf{W}\, d\,\mathbf{H} \\
&= \underbrace{\mathbb{E}_{q(\boldsymbol{\theta}_L)}[\ln p(\mathbf{V},\boldsymbol{\theta}_L;\boldsymbol{\theta}_R)]}_{\triangleq\ \mathcal{L}_V(q(\boldsymbol{\theta}_L);\boldsymbol{\theta}_R)} - \underbrace{\mathbb{E}_{q(\boldsymbol{\theta}_L)}[\ln q(\boldsymbol{\theta}_L)]}_{\triangleq\ -\mathcal{L}_E(q(\boldsymbol{\theta}_L))} \\
&\triangleq \mathcal{L}_B(q(\boldsymbol{\theta}_L);\boldsymbol{\theta}_R) \qquad (5.12)
\end{aligned}$$

where $q(\cdot)$ is a variational distribution and $\mathbb{E}_{g(x)}[f(x)]$ indicates an expectation of $f(x)$ with respect to $g(x)$. The term $\mathcal{L}_B(q(\boldsymbol{\theta}_L);\boldsymbol{\theta}_R)$ defines the lower bound on $\ln p(\mathbf{V};\boldsymbol{\theta}_R)$, where the equality holds for $q(\boldsymbol{\theta}_L) = p(\boldsymbol{\theta}_L \,|\, \mathbf{V};\boldsymbol{\theta}_R)$ [41,67] (see Appendix A.2 for a detailed expression of the lower bound). The VBEM algorithm consists of two stages. During the variational E-step, the variational distribution $q(\boldsymbol{\theta}_L)$ which approximates the exact posterior distribution $p(\boldsymbol{\theta}_L \,|\, \mathbf{V};\boldsymbol{\theta}_R)$ is estimated. During the variational M-step, the hyperparameters are obtained by maximizing the lower bound in (5.12) computed with a *fixed* $q(\boldsymbol{\theta}_L)$.

1) Variational E-step: Based on the mean-field approximation [67,107], the variational distribution $q(\mathbf{C},\mathbf{A},\mathbf{W},\mathbf{H})$ can be expressed in a factorized form as (e.g., [41,43,108])

$$\begin{aligned}
q(\mathbf{C},\mathbf{A},\mathbf{W},\mathbf{H}) &= q(\mathbf{C})q(\mathbf{A})q(\mathbf{W})q(\mathbf{H}) \qquad (5.13) \\
&= \left(\prod_{j,k,l} q(\mathbf{c}_{kl}^{j})\right)\left(\prod_{j,k} q(a_k^j)\right)\left(\prod_{k,m} q(w_{km})\right)\left(\prod_{m,l} q(h_{ml})\right)
\end{aligned}$$

where $\mathbf{c}_{kl}^j = [c_{kl}^{1,j}, ..., c_{kl}^{M,j}]$. The resulting local optimal solutions can be found as:

$$q(\mathbf{C})^{(r+1)} \propto \exp\left(\mathbb{E}_{q(\mathbf{A})^{(r)}q(\mathbf{W})^{(r)}q(\mathbf{H})^{(r)}}[\ln p(\mathbf{V},\boldsymbol{\theta}_L;\boldsymbol{\theta}_R)]\right) \qquad (5.14)$$

$$q(\mathbf{A})^{(r+1)} \propto \exp\Big(\mathbb{E}_{q(\mathbf{C})^{(r+1)} q(\mathbf{W})^{(r)} q(\mathbf{H})^{(r)}} [\ln p(\mathbf{V}, \boldsymbol{\theta}_L; \boldsymbol{\theta}_R)] \Big) \tag{5.15}$$

$$q(\mathbf{W})^{(r+1)} \propto \exp\Big(\mathbb{E}_{q(\mathbf{C})^{(r+1)} q(\mathbf{A})^{(r+1)} q(\mathbf{H})^{(r)}} [\ln p(\mathbf{V}, \boldsymbol{\theta}_L; \boldsymbol{\theta}_R)] \Big) \tag{5.16}$$

$$q(\mathbf{H})^{(r+1)} \propto \exp\Big(\mathbb{E}_{q(\mathbf{C})^{(r+1)} q(\mathbf{W})^{(r+1)} q(\mathbf{W})^{(r+1)}} [\ln p(\mathbf{V}, \boldsymbol{\theta}_L; \boldsymbol{\theta}_R)] \Big) \tag{5.17}$$

where the superscript (r) denotes the r-th iteration. For convenience, we shall omit the superscript (r) and also drop the latent variables inside the subscript $q(\cdot)$ of the expectation operator, e.g., $\mathbb{E}_{q(w_{km})}[w_{km}] = \mathbb{E}_q[w_{km}]$.

First, the distribution $q(\mathbf{c}^j_{kl})$ in (5.14) is inferred as [41]:

$$\mathcal{M}(\mathbf{c}^j_{kl}; v^j_{kl}, \bar{\mathbf{p}}^j_{kl}) = \delta\left(v^j_{kl} - \sum_{m=1}^{M} c^{m,j}_{kl} \right) v^j_{kl}! \prod_{m=1}^{M} \frac{(\bar{p}^{m,j}_{kl})^{c^{m,j}_{kl}}}{c^{m,j}_{kl}!} \tag{5.18}$$

where $\delta(x)$ is the Kronecker delta function defined by $\delta(x) = 1$ when $x = 0$ and $\delta(x) = 0$ otherwise. The entries of $\bar{\mathbf{p}}^j_{kl} = [\bar{p}^{m,j}_{kl}]$ are given by

$$\bar{p}^{m,j}_{kl} = \frac{\exp\big(\mathbb{E}_q[\ln a^j_k] + \mathbb{E}_q[\ln w_{km}] + \mathbb{E}_q[\ln h_{ml}] \big)}{\sum_{m=1}^{M} \exp\big(\mathbb{E}_q[\ln a^j_k] + \mathbb{E}_q[\ln w_{km}] + \mathbb{E}_q[\ln h_{ml}] \big)}. \tag{5.19}$$

The distributions $q(a^j_k)$, $q(w_{km})$ and $q(h_{km})$ are found successively as

$$\begin{aligned} q(a^j_k) &\propto \exp\Bigg[\underbrace{\Big(\alpha_a + \sum_{m=1}^{M} \sum_{l=1}^{L} \mathbb{E}_q[c^{m,j}_{kl}] - 1 \Big)}_{\triangleq \bar{\alpha}^j_{a,k}} \ln a^j_k - \underbrace{\Big(\frac{\alpha_a}{\beta_a} + \sum_{m=1}^{M} \sum_{l=1}^{L} \mathbb{E}_q[w_{km}]\, \mathbb{E}_q[h_{ml}] \Big)}_{\triangleq (\bar{\beta}^j_{a,k})^{-1}} a^j_k \Bigg] \\ &\propto \mathcal{G}(a^j_k; \bar{\alpha}^j_{a,k}, \bar{\beta}^j_{a,k}) \end{aligned} \tag{5.20}$$

$$q(w_{km}) \propto \exp\left[\Big(\underbrace{\alpha_w + \sum_{j=1}^{J}\sum_{l=1}^{L}\mathbb{E}_q[c_{kl}^{m,j}] - 1}_{\triangleq \bar{\alpha}_{w,km}}\Big)\ln w_{km} - \Big(\underbrace{\frac{\alpha_w}{\beta_w} + \sum_{j=1}^{J}\sum_{l=1}^{L}\mathbb{E}_q[a_k^j]\,\mathbb{E}_q[h_{ml}]}_{\triangleq(\bar{\beta}_{w,km})^{-1}}\Big)w_{km}\right]$$

$$\propto \mathcal{G}(w_{km}; \bar{\alpha}_{w,km}, \bar{\beta}_{w,km}) \tag{5.21}$$

$$q(h_{ml}) \propto \exp\left[\Big(\underbrace{\alpha_h + \sum_{k=1}^{K}\sum_{j=1}^{J}\mathbb{E}_q[c_{kl}^{m,j}] - 1}_{\triangleq \bar{\alpha}_{h,ml}}\Big)\ln h_{ml} - \Big(\underbrace{\frac{\alpha_h}{\beta_h} + \sum_{k=1}^{K}\sum_{j=1}^{J}\mathbb{E}_q[a_k^j]\,\mathbb{E}_q[h_{ml}]}_{\triangleq(\bar{\beta}_{h,ml})^{-1}}\Big)h_{ml}\right]$$

$$\propto \mathcal{G}(h_{ml}; \bar{\alpha}_{h,ml}, \bar{\beta}_{h,ml}). \tag{5.22}$$

The sufficient statistics (expectations) are given below:

$$\mathbb{E}_q[c_{kl}^{m,j}] = v_{kl}^j \bar{p}_{kl}^{m,j} \tag{5.23}$$

$$\mathbb{E}_q[\ln a_k^j] = \Psi(\bar{\alpha}_{a,k}^j) + \ln \bar{\beta}_{a,k}^j, \qquad \mathbb{E}_q[a_k^j] = \bar{\alpha}_{a,k}^j \bar{\beta}_{a,k}^j \tag{5.24}$$

$$\mathbb{E}_q[\ln w_{km}] = \Psi(\bar{\alpha}_{w,km}) + \ln \bar{\beta}_{w,km}, \quad \mathbb{E}_q[w_{km}] = \bar{\alpha}_{w,km}\bar{\beta}_{w,km} \tag{5.25}$$

$$\mathbb{E}_q[\ln h_{ml}] = \Psi(\bar{\alpha}_{h,ml}) + \ln \bar{\beta}_{h,ml}, \qquad \mathbb{E}_q[h_{ml}] = \bar{\alpha}_{h,ml}\bar{\beta}_{h,ml} \tag{5.26}$$

where $\Gamma(x)$ is the gamma function and $\Psi(x) = d\ln\Gamma(x)/dx$ is the digamma function [41].

2) Variationl M-step: The hyper-parameter set $\boldsymbol{\theta}_R$ is estimated by maximizing $\mathcal{L}_V(q(\boldsymbol{\theta}_L)^{(r+1)}; \boldsymbol{\theta}_R)$. Setting the partial derivative of $\mathcal{L}_V(q(\boldsymbol{\theta}_L)^{(r+1)}; \boldsymbol{\theta}_R)$ with respect to $\boldsymbol{\theta}_R$ to zero, the hyper-parameters for the mixing coefficients, α_a and β_a, are obtained as

$$\alpha_a \leftarrow \alpha_a - \frac{\ln\alpha_a - \Psi(\alpha_a) + 1 - \alpha_{qa}}{1/\alpha_a - \Psi'(\alpha_a)} \tag{5.27}$$

$$\beta_a = \frac{1}{KJ} \sum_{k=1}^{K} \sum_{j=1}^{J} \mathbb{E}_q[a_k^j] \tag{5.28}$$

where $\alpha_{qa} = \sum_{k=1}^{K} \sum_{j=1}^{J} (\mathbb{E}_q[a_k^j]/\beta_a - \mathbb{E}_q[\ln a_k^j] + \ln \beta_a)/(KJ)$ and $\Psi'(x) = d\Psi(x)/dx$. The hyper-parameters for the bases, α_w and β_w, are obtained as

$$\alpha_w \leftarrow \alpha_w - \frac{\ln \alpha_w - \Psi(\alpha_w) + 1 - \alpha_{qw}}{1/\alpha_w - \Psi'(\alpha_w)} \tag{5.29}$$

$$\beta_w = \frac{1}{KM} \sum_{k=1}^{K} \sum_{m=1}^{M} \mathbb{E}_q[w_{km}] \tag{5.30}$$

where $\alpha_{qw} = \sum_{k=1}^{K} \sum_{m=1}^{M} (\mathbb{E}_q[w_{km}]/\beta_w - \mathbb{E}_q[\ln w_{km}] + \ln \beta_w)/(KM)$. The hyper-parameters for the activations, α_h and β_h, are obtained as

$$\alpha_h \leftarrow \alpha_h - \frac{\ln \alpha_h - \Psi(\alpha_h) + 1 - \alpha_{qh}}{1/\alpha_h - \Psi'(\alpha_h)} \tag{5.31}$$

$$\beta_h = \frac{1}{ML} \sum_{m=1}^{M} \sum_{l=1}^{L} \mathbb{E}_q[h_{ml}] \tag{5.32}$$

where $\alpha_{qh} = \sum_{m=1}^{M} \sum_{l=1}^{L} (\mathbb{E}_q[h_{ml}]/\beta_h - \mathbb{E}_q[\ln h_{ml}] + \ln \beta_h)/(ML)$.

5.2.4 Parameter initialization and normalization

For initialization, we first generate positive random numbers and subsequently apply the MU rules in (2.5) to $\bar{\mathbf{V}} = [\sum_{j=1}^{J} v_{kl}^j/J] \in \mathbb{R}_+^{K\times L}$ for several iterations [33, 34, 108], where we found 10 iterations are sufficient. The resulting $\mathbf{W}$ and $\mathbf{H}$ are used as the initial values for the sufficient statistics $\mathbb{E}_q[w_{km}]$, $\exp(\mathbb{E}_q[\ln w_{km}])$, $\mathbb{E}_q[h_{ml}]$ and $\exp(\mathbb{E}_q[\ln h_{ml}])$. The sufficient statistics $\mathbb{E}_q[a_k^j]$ and $\exp(\mathbb{E}_q[\ln a_k^j])$ are initialized to 1. The hyper-parameters are initialized as $\alpha_a = \alpha_w = \alpha_h = 0.001$ and $\beta_a = \beta_w = \beta_h = 10$. We use 200 iterations for the VBEM algorithm, whereas 5 iterations are used for estimating the hyper-parameters in (5.27), (5.29) and (5.31). To avoid scale indeterminacies in a_k^j, w_{km} and h_{ml} which appear as a product in the distribution (5.3), we include a parameter normalization step at each iteration. Motivated by [64], we normalize $\mathbb{E}_q[a_k^j]$ and $\exp(\mathbb{E}_q[\ln a_k^j])$ after computing (5.20), such that $\sum_j \mathbb{E}_q[a_k^j] = 1$ and $\sum_j \exp(\mathbb{E}_q[\ln a_k^j]) = 1$. Also, we normalize $\mathbb{E}_q[w_{km}]$ and $\exp(\mathbb{E}_q[\ln w_{km}])$ after computing (5.21), such that $\sum_k \mathbb{E}_q[w_{km}] = 1$ and $\sum_k \exp(\mathbb{E}_q[\ln w_{km}]) = 1$.

5.3 Proposed Enhancement Stage

In the enhancement stage, we propose an effective method of estimating the clean speech point source signal via the MVDR beamforming technique, whose realization involves two main steps. First, the speech source location is determined by observing the spatial output power of a DS beamformer applied to the MNMF-based pre-processed noisy speech signal. Second, the noise correlation matrix is computed using the MNMF parameters for the magnitude components, and a combination of the noisy speech phase and steering vector for the phase components. In the following subsections, we explain the MNMF parameter estimation, followed by the proposed source localization method and noise correlation matrix computation and finally, the proposed MVDR beamforming method.

5.3.1 MNMF parameter estimation

We consider a mini-batch online approach while enhancing the noisy speech signal. Let us denote by $\mathbf{Y}_{l_b} = [Y_{kl}^j] \in \mathbb{C}^{K\times L_b \times J}$ the tensor consisting of the noisy speech spectral coefficients of the time frames $l \in \{(l_b-1)L_b+1, ..., l_b L_b\} \triangleq C_{l_b}$, where $l_b = 1, 2...$ is the mini-batch frame index and L_b is the mini-batch size. For the l_b-th mini-batch frame, we aim at factorizing the tensor $\mathbf{V}_{l_b}^Y = [|Y_{kl}^j|] \in \mathbb{R}_+^{K\times L_b\times J}$ into $\mathbf{A}_Y = \{a_k^{S,j}, a_k^{N,j}\}$, $\mathbf{W}_Y =$

$[\mathbf{W}_S \;\; \mathbf{W}_N] = [w^Y_{km}] \in \mathbb{R}_+^{K \times M_Y}$ and $\mathbf{H}^Y_{l_b} = [h^Y_{ml}] \in \mathbb{R}_+^{M_Y \times L_b}$, where $M_Y = M_S + M_N$ and $a^{S,j}_k$ and $a^{N,j}_k$ are the magnitude values of the ATFs of the clean speech and noise $\tilde{A}^{S,j}_k$ and $\tilde{A}^{N,j}_k$ given by (2.38) (i.e., $a^{S,j}_k \triangleq |\tilde{A}^{S,j}_k|$ and $a^{N,j}_k \triangleq |\tilde{A}^{N,j}_k|$). The factorization model can be written as

$$|Y^j_{kl}| \approx \hat{v}^j_{kl} = a^{S,j}_k \sum_{m=1}^{M_S} w^S_{km} h^S_{ml} + a^{N,j}_k \sum_{m=1}^{M_N} w^N_{km} h^N_{ml}. \tag{5.33}$$

We use the VBEM algorithm introduced in Subsection 5.2.3 to infer the variational distributions $q(\mathbf{C})$, $q(\mathbf{W}_Y)$ and $q(\mathbf{H}_Y)$. That is, at each iteration, the parameters of $q(\mathbf{C})$ are estimated via (5.18) and (5.19). Next, we estimate the parameters of $q(\mathbf{A}_Y)$ based on (5.20). The parameters of $q(\mathbf{H}^Y_{l_b})$ are then simply obtained by using (5.22). Subsequently, the hyper-parameters of the mixing coefficients and noisy speech activation priors are obtained by

$$\beta^S_a = \frac{1}{KJ} \sum_{k=1}^{K} \sum_{j=1}^{J} \mathbb{E}_q[a^{S,j}_k], \qquad \beta^N_a = \frac{1}{KJ} \sum_{k=1}^{K} \sum_{j=1}^{J} \mathbb{E}_q[a^{N,j}_k] \tag{5.34}$$

$$\beta^S_h = \frac{1}{M_S L_b} \sum_{m=1}^{M_S} \sum_{l \in C_{l_b}} \mathbb{E}_q[h^Y_{ml}], \qquad \beta^N_h = \frac{1}{M_N L_b} \sum_{m=M_S+1}^{M_Y} \sum_{l \in C_{l_b}} \mathbb{E}_q[h^Y_{ml}] \tag{5.35}$$

In contrast to the parameters given above, we fix the shape parameters, mainly in order to reduce the computational cost since their updates require additional iterations as given by (5.27) and (5.31) [34].

5.3.2 Source localization and noise correlation matrix computation

One main step in the MVDR beamformer is the steering vector specified by the TDoA, which is computed based on the estimated speech source position. To this end, numerous methods for source localization have been proposed, such as angular spectrum-based and clustering-based methods. The former method constructs a function of the TDoA known as the angular spectrum for each time-frequency bin. The source TDoA is then estimated by the highest peak of that function. The latter method iteratively estimates the time-frequency bins associated to each source, e.g., the clean speech and noise, and the corresponding TDoAs by means of a clustering algorithm. A more detailed discussion of various source localization methods can be found in [118] and references therein.

In the proposed framework, we consider the angular spectrum-based approach mainly due to the simplicity of its implementation compared to the clustering-based approach. Specifically, the speech source position is determined by observing the spatial output power of a DS beamformer applied to the noisy speech signal [35,119]. To improve the localization performance, we apply the DS beamformer to the MNMF-based pre-processed noisy speech signal, and subsequently apply the single-channel Wiener filtering to the DS beamformer output as given by (2.48). A detailed processing of the proposed localization method is described below.

Let us denote by $\mathbf{L}_o = \{\mathbf{l}_o\}_{o=1}^{O}$ a set of unit-length look direction vectors $\mathbf{l}_o$ (i.e., the search grid for source localization). Based on (2.39), the NMF-based pre-processed noisy speech, $\bar{\mathbf{S}}_{kl} = [\bar{S}_{kl}^{1}, ..., \bar{S}_{kl}^{J}]^T \in \mathbb{C}^J$ and pre-estimated noise, $\bar{\mathbf{N}}_{kl} = [\bar{N}_{kl}^{1}, ..., \bar{N}_{kl}^{J}]^T \in \mathbb{C}^J$, are first obtained as

$$\bar{S}_{kl}^{j} = \frac{(a_k^{S,j})^2 \hat{p}_{kl}^{S}}{(a_k^{S,j})^2 \hat{p}_{kl}^{S} + (a_k^{N,j})^2 \hat{p}_{kl}^{N}} Y_{kl}^{j} \tag{5.36}$$

and $\bar{N}_{kl}^{j} = Y_{kl}^{j} - \bar{S}_{kl}^{j}$, where $\hat{p}_{kl}^{S}$ and $\hat{p}_{kl}^{N}$ are given by (2.32) and (2.33). Specifically, as explained in Subsection 4.2.2, the latter are computed based on the mean values of $q(\mathbf{W}_{l_b})$ and $q(\mathbf{H}_{l_b})$ and similarly, we use $\mathbb{E}_q[a_k^{S,j}]$ and $\mathbb{E}_q[a_k^{N,j}]$ in (5.36). Subsequently, the clean speech and noise spectra at the o-th DoA, $\hat{S}_{kl}^{o}$ and $\hat{N}_{kl}^{o}$, are estimated by applying the DS beamformer to $\bar{S}^j_{kl}$ and $\bar{N}^j_{kl}$, followed by single-channel Wiener filtering as:

$$\hat{S}^o_{kl} = \frac{\hat{p}^S_{kl}}{\hat{p}^S_{kl} + \hat{p}^N_{kl}} \hat{S}^D_{kl} \tag{5.37}$$

$$\hat{N}^o_{kl} = \frac{\hat{p}^N_{kl}}{\hat{p}^S_{kl} + \hat{p}^N_{kl}} \hat{N}^D_{kl} \tag{5.38}$$

where $\hat{S}^D_{kl} = (\mathbf{b}^o_k)^H \bar{\mathbf{S}}_{kl}$ and $\hat{N}^D_{kl} = (\mathbf{b}^o_k)^H \bar{\mathbf{N}}_{kl}$ are the DS beamformer outputs, and $\mathbf{b}^o_k$ is the steering vector given by (2.45) which is computed based on $\mathbf{l}_o$. At the l_b-th mini-batch, the look direction vectors for the clean speech and noise, $\mathbf{l}^S_{o,l_b}$ and $\mathbf{l}^N_{o,l_b}$, are then determined that give the highest input SNR (in dB)

$$R^o_{l_b} = 10 \log_{10} \frac{\sum_{k=1}^K \sum_{l \in C_{l_b}} |\hat{S}^o_{kl}|^2}{\sum_{k=1}^K \sum_{l \in C_{l_b}} |\hat{N}^o_{kl}|^2} \tag{5.39}$$

and the lowest input SNR, respectively.

To avoid a rapid change in the estimation of the source locations and to exploit extra DoAs which are not included in the search grid $\mathbf{L}_o$, we consider the smoothed $\mathbf{l}_o$ based on the input SNR as

$$\hat{\mathbf{l}}^S_{o,l_b} = \sigma(R_{l_b-1}) \hat{\mathbf{l}}^S_{o,l_b-1} + \sigma(R_{l_b}) \mathbf{l}^S_{o,l_b} \tag{5.40}$$

$$\hat{\mathbf{l}}^N_{o,l_b} = \sigma(-R_{l_b-1}) \hat{\mathbf{l}}^S_{o,l_b-1} + \sigma(-R_{l_b}) \mathbf{l}^N_{o,l_b} \tag{5.41}$$

where $\sigma(x) = \rho_1/(1 + \exp(-\rho_2 x))$ is the logistic function, and R_{l_b} is the estimated input SNR in dB given by

$$R_{l_b} = 10 \log_{10} \frac{\sum_{j=1}^J \sum_{k=1}^K \sum_{l \in C_{l_b}} (a^{S,j}_k)^2 \hat{p}^S_{kl}}{\sum_{j=1}^J \sum_{k=1}^K \sum_{l \in C_{l_b}} (a^{N,j}_k)^2 \hat{p}^N_{kl}}. \tag{5.42}$$

The estimated look direction vectors are then normalized to have unit length. Note that R_{l_b} differs from $R^o_{l_b}$ as follows. The former SNR is computed based on the image spectra and used for estimating the *smoothed* look direction vectors $\hat{\mathbf{l}}^S_{o,l_b}$ and $\hat{\mathbf{l}}^N_{o,l_b}$. The latter SNR is computed based on the DS beamformer output and used for estimating the *instantaneous* look direction vectors $\mathbf{l}^S_{o,l_b}$ and $\mathbf{l}^N_{o,l_b}$.

Another main step in the MVDR beamformer is the estimation of the noise correlation matrix $\mathbf{R}^N_{kl}$, which can be computed via temporal smoothing based on (2.42). Regarding the phase components, we can use the noisy speech phase as mentioned in Subsection 2.2.2. However, this may limit the performance especially for a high input SNR since the clean speech phase will dominate the noisy speech phase. Hence, we instead propose to use the combination of the noisy speech phase and steering vector, based on the estimated input SNR, as

$$\varphi^j_{N,k} = \sigma(R_{l_b}) b^{N,j}_k + (1 - \sigma(R_{l_b})) \exp(\jmath \angle Y^j_{kl}) \tag{5.43}$$

where $b^{N,j}_k$ is the k-th entry of the steering vector given by (2.45), computed based on the estimated noise look direction vector $\hat{\mathbf{l}}^N_{o,l_b}$. The estimated phase-related component $\varphi^{N,j}_k$ is then normalized to have unit magnitude, i.e., $\varphi^{N,j}_k / |\varphi^{N,j}_k|$, to ensure $|\exp(\jmath \angle \hat{A}^j_{N,k})| = 1$ as discussed in connection with image source estimation in Subsection 2.2.2.

5.3.3 NMF-based MVDR beamforming

The MVDR beamformer output is given by [120]

$$\hat{S}_{kl} = \left(\frac{(\mathbf{R}^N_{kl} + \lambda \mathbf{I}_J)^{-1} \mathbf{b}_k}{\mathbf{b}^H_k (\mathbf{R}^N_{kl} + \lambda \mathbf{I}_J)^{-1} \mathbf{b}_k} \right)^H \mathbf{Y}_{kl} \tag{5.44}$$

where λ is the diagonal loading factor, $\mathbf{I}_J$ is the $J \times J$ identity matrix, and $\mathbf{b}_k = [b^j_k] \in \mathbb{C}^J$ is the steering vector computed based on (2.45) and estimated DoA.

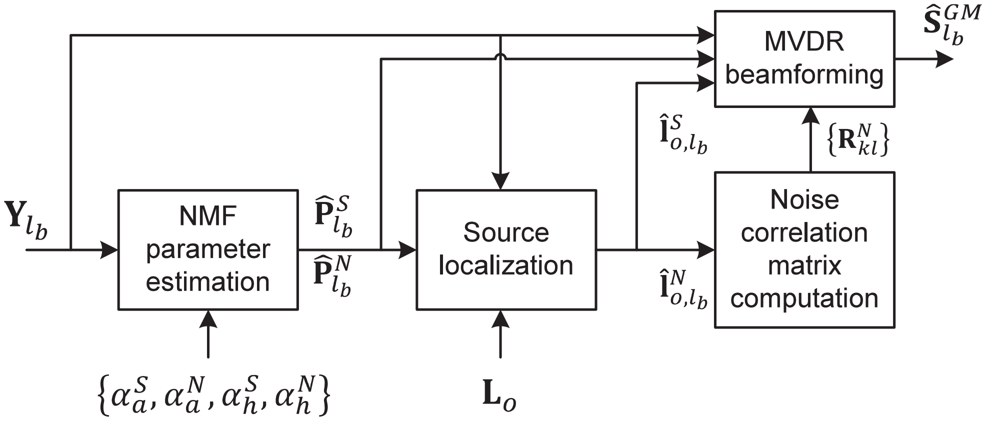
MVDR beamforming
$\hat{\mathbf{S}}_{l_b}^{GM}$
$\hat{\mathbf{l}}_{o,l_b}^{S}$
$\{\mathbf{R}_{kl}^{N}\}$
$\mathbf{Y}_{l_b}$
NMF parameter estimation
$\hat{\mathbf{P}}_{l_b}^{S}$
$\hat{\mathbf{P}}_{l_b}^{N}$
Source localization
$\hat{\mathbf{l}}_{o,l_b}^{N}$
Noise correlation matrix computation
$\{\alpha_a^S, \alpha_a^N, \alpha_h^S, \alpha_h^N\}$
$\mathbf{L}_o$

Fig. 5.1 A simplified block diagram of the proposed enhancement stage.

To further reduce the residual noise components in the enhanced speech obtained via the MVDR beamformer, we can apply a single-channel enhancement algorithm to the beamformer output as a post-processor, as given by (2.48) [19, 20]. However, the post-processing may further distort the clean speech components, especially for a high input SNR. Hence, motivated by [103], we consider the WGM of the magnitude components of the MVDR output and its improvement via single-channel Wiener filtering:

$$\hat{S}_{kl}^{GM} = \left(\left| \hat{S}_{kl} \right|^{\nu_{l_b}} \left| \frac{\hat{p}_{kl}^S}{\hat{p}_{kl}^S + \hat{p}_{kl}^N} \hat{S}_{kl} \right|^{1-\nu_{l_b}} \right) e^{\jmath \angle \hat{S}_{kl}} = \left(\frac{\hat{p}_{kl}^S}{\hat{p}_{kl}^S + \hat{p}_{kl}^N} \right)^{1-\nu_{l_b}} \hat{S}_{kl} \tag{5.45}$$

where $0 \le \nu_{l_b} \le 1$ is the weighting factor. Similar as in [103], we select the weighting factor by $\nu_{l_b} = \sigma(R_{l_b})$, where R_{l_b} is the estimated input SNR given by (5.42).

A simplified block diagram of the proposed method is illustrated in Figure 5.1, while the algorithm is summarized in Table 5.1. The proposed method, i.e., variational inference on the multi-channel Bayesian NMF model and the MNMF-based beamforming, will be referred to as VMNMF throughout this section.

Table 5.1 Algorithm summary of the proposed enhancement stage

```
% Initialize
Initialize $q(\mathbf{A}_Y)$ parameters to 1
Initialize $\hat{p}^S_{k,0}$ and $\hat{p}^N_{k,0}$ to 0
Initialize $R_{l_b}$ to 0

for $l_b = 1, 2, ...$
    % NMF parameter estimation
    Initialize $q(\mathbf{A}_Y)$ parameters to the ones estimated at $l_b - 1$
    Initialize $q(\mathbf{H}_{l_b})$ parameters by generating positive random numbers
    for iter = 1:itermax
        Estimate $q(\mathbf{A}_Y)$ and normalize
        Estimate $q(\mathbf{H}_{l_b})$
        Update $\beta^S_a$, $\beta^N_a$, $\beta^S_h$ and $\beta^N_h$ via (5.34) and (5.35)
    end
    Compute $\hat{\mathbf{P}}^S_{l_b} = [\hat{p}^S_{kl}]$ and $\hat{\mathbf{P}}^N_{l_b} = [\hat{p}^N_{kl}]$ via (2.32) and (2.33)
    Compute $\nu_{l_b}$ via (5.42)

    % Source localization
    Compute $\bar{\mathbf{S}}_{l_b} = [\bar{S}^j_{kl}]$ via (5.36) and $\bar{\mathbf{N}}_{l_b} = [\bar{N}^j_{kl}] = \mathbf{Y}_{l_b} - \bar{\mathbf{S}}_{l_b}$
    for $o = 1, ..., O$
        Compute $\hat{\mathbf{S}}^o_{l_b} = [\hat{S}^o_{kl}]$ and $\hat{\mathbf{N}}^o_{l_b} = [\hat{N}^o_{kl}]$ via (5.37) and (5.38)
        Compute $R^o_{l_b}$ via (5.39)
    end
    Compute $\hat{\mathbf{I}}^S_{o,l_b}$ and $\hat{\mathbf{I}}^N_{o,l_b}$ via (5.40) and (5.41)
    Compute $R_{l_b}$ via (5.42)

    % MVDR beamforming
    Compute $\mathbf{b}_k$ via (2.45) and (2.46)
    Compute $\{\mathbf{R}^N_{kl}\}$ via (2.42) and (5.43)
    Estimate $\hat{\mathbf{S}}^{GM}_{l_b} = [\hat{S}^{GM}_{kl}]$ via (5.44) and (5.45)
```

5.4 Experiments

The enhancement performance of the proposed method was assessed through objective experiments. Below, we describe the general methodology and benchmark algorithms, and then present the experimental results.

5.4.1 Methodology

We conducted the experiments using the 4-th CHiME challenge corpus [115], where the sampling rate of all signals was set to 16 kHz. The speech and noise files were divided into two disjoint groups: i) *training data*, used for estimating the basis matrix during the training stage, and ii) *test data*, used during the enhancement stage to evaluate the enhancement performance. The clean speech training data of the CHiME database are from the Wall Street Journal (WSJ0) corpus, which consists of 101 speakers. We considered a speaker-independent (SI) application, where one *universal* basis matrix covering all speakers is estimated during the training stage. To this end, we randomly selected 40 utterances from each speaker and concatenated them to construct the clean speech training data, resulting in an 8 hours long signal. Regarding the noise training data, we considered the Bus, Cafe, Pedestrian, and Street noises, where we used a 2 hours long signal for each noise type.

Regarding the test data, we used the development set of the CHiME corpus (referred to as "simulated development data"), which consists of 410 utterances of the 6-channel noisy speech signals, generated by artificially mixing the clean speech and noises. The multi-channel noisy speech signals were generated by scaling and adding the noise to the filtered clean speech signal to obtain input SNRs of -5, 0, 5, and 10 dB. Specifically, the clean speech signals were filtered by the *time-varying* impulse responses (IR) between the speaker and microphones. The IR is estimated from the real recorded signals (see [115] for more details about the database).

Regarding the implementation, a Hanning window of 1024 samples with 50% overlap was employed for the STFT analysis. We used $M = 80$ basis vectors for the clean speech and noises and $L_b = 32$ for the mini-batch size. Although we can use noise classification as introduced in [103] to determine which type of noise is included in the noisy speech, we assumed that the noise type is known *a priori* throughout the experiments. The values of $\tau_S = 0.4$ were chosen as the temporal smoothing factors in (2.32), and $\tau_N = 0.9$ in (2.33) and (2.42). We used $\lambda = 0.01$ for the diagonal loading factor for the MVDR beamformer. Regarding the shape parameters during the enhancement stage, we used $\alpha_a^S = \alpha_a^N = 1$ and $\alpha_h^S = \alpha_h^N = 0.1$. We used $\rho_1 = 2$ and $\rho_2 = 0.01$ for the parameters of the logistic function while computing the smoothed look direction vector sgiven by (5.40) and (5.41), whereas we used $\rho_1 = \rho_2 = 0.5$ while computing the noise correlation matrix in (5.43) and the clean speech spectrum estimation in (5.45). The geometry of the microphone array used for recording the CHiME data is described in [77], which can be expressed in the Cartesian coordinates by (x, y, z) = (-0.1,0.095,0), (0,0.095,-2), (0.1,0.095,0), (-0.1,-0.095,0), (0,-0.095,0), (0.1,-0.095,0) in meters. We used the origin, i.e., $(x, y, z) = (0, 0, 0)$, for the reference position $\mathbf{l}_r$. Considering the spherical polar coordinates, let us denote by θ the azimuth angle in the xy-plane from the x-axis with $0 \leq \theta < 2\pi$, and by ϕ the polar angle from the positive z-axis with $0 \leq \phi \leq \pi$ (see Figure 5.2). The unit-length look direction vector $\mathbf{l}_o \in \mathbb{R}^3$ then can be expressed in the Cartesian coordinates as $(x, y, z) = (\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi)$. Regarding the search grid for source localization (i.e., the set of look direction vectors $\mathbf{L}_o$), we considered the polar angles of $\phi = \{0, 15, 30, 45, 60\}$ in degrees, where the azimuth angles for the given polar angle were sampled by the angles of 360, 45, 30, 15 and 10 in degrees, respectively, resulting in a total of $O = 81$ look direction vectors.

We considered the PESQ [98], SDR [99] and overall composite value (Cov) [121] as the objective measures of performance. The PESQ attempts to predict overall perceptual quality in MOS and the SDR measures the overall quality of the enhanced speech in dB by considering both aspects of speech distortion and noise reduction. The Cov predicts overall quality in MOS, by taking into account of different objective measures, i.e.,

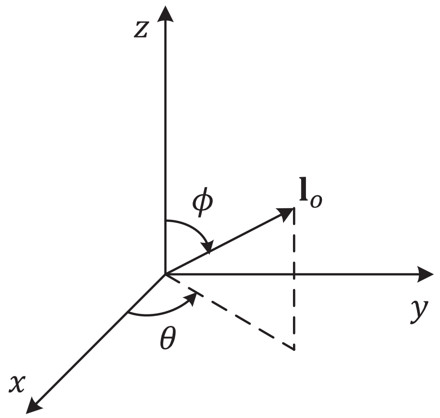

Fig. 5.2 Geometric illustration of the spherical coordinates.

PESQ, weighted-slope spectral distance (WSS) and log-likelihood ratio (LLR). For all the measures, a higher value indicates a better result.

5.4.2 Benchmark algorithms

To evaluate the enhancement performance of the proposed VMNMF method, we implemented several benchmark algorithms, which are summarized below. Basic settings, such as the STFT analysis and synthesis, the number of basis vectors, the mini-batch size L_b and the reconstruction method, were kept identical when applicable (except the diagonal loading factor λ in (5.44) where we used 0.1 since it provided better enhancement performance). Moreover, the essential steps for the MVDR beamforming, i.e., the source localization and noise correlation matrix computation, were performed using the proposed method introduced in Subsections 5.3.2 and 5.3.3.

1) MMSE-STSA estimator: We implemented the MMSE-STSA estimator [5] and applied it to each channel of the noisy speech to obtain the initial estimates of the clean speech and noise spectra for the source localization and MVDR beamforming. The noise PSD was estimated based on [93]. Regarding the post-processing for the WGM-based clean speech estimation, we applied the MMSE-STSA estimator to the MVDR output. We used a smoothing factor of 0.85 in the decision-directed method for computing the *a priori* SNR [5], and used 0.9 for the noise PSD estimation [93].

2) MNMF: We implemented the MNMF algorithm based on the IS-divergence [33]. Specifically, we considered the MU rules, which were applied to the power spectral coefficients. During the training stage, the basis and activation matrices were initialized by first generating positive random numbers and subsequently applying the KL-based MU rules in (2.5) to $\bar{\mathbf{V}} = [\sum_{j=1}^{J} v_{kl}^{j}/J]$, where v_{kl}^{j} is the power spectral coefficient of the training data. The mixing coefficients were initialized to 1.

3) NTF: We implemented the NTF algorithms based on the KL and IS divergences (i.e., the MU rules introduced in [69]), which will be referred to as NTF-KL and NTF-IS, respectively. The NTF-KL method was applied to the magnitude spectral coefficients, whereas the NTF-IS was applied to the power spectral coefficients. During the training stage, the basis and activation matrices were initialized by positive random numbers for the NTF-KL method, whereas we followed the same procedure as in the MNMF method for the NTF-IS method. The mixing coefficients were initialized to 1, for both the NTF-KL and NTF-IS methods.

5.4.3 Results

The average results over all utterances are shown in Tables 5.2 to 5.5, where the values in bold indicate the best performance along the corresponding row. We can observe that the proposed VMNMF method provided better enhancement performance than the benchmark algorithms for all types of noises and input SNRs (except in some cases where the best Cov results at -5 dB input SNR were found in the MNMF method for the Bus and Street noises). Comparing between the IS-based methods (i.e., MNMF and NTF-IS methods), the MNMF method provided better results in general. This validates that the MNMF model, which employs the frequency-dependent mixing coefficients, can better handle the convolutive effects specified by the ATFs. Also comparing between NTF-KL and NTF-IS methods, the former method showed better enhancement performance in general. Such patterns that the KL-based method can provide better performance than the IS-based method, have been widely reported e.g., in [73, 74]. Moreover, we found that the computational cost of the VMNMF method was comparable to that of the efficient MU-based MNMF, NTF-KL and NTF-IS methods.

Figure 5.3 illustrates the magnitude spectra of the clean speech (point source signal), noisy speech (recorded at microphone $j = 1$) and the enhanced speech using the proposed VMNMF method. In this particular example, a female speech is degraded by the Street noise at 0 dB input SNR. As it can be observed, the background noise has been significantly reduced.

5.5 Summary

We introduced a supervised multi-channel speech enhancement algorithm based on a Bayesian MNMF model. In the proposed framework, we considered the PGM of MNMF, specified by Poisson-distributed latent variables and gamma-distributed priors. During the proposed training stage, the MNMF parameters of different speech and noise sources are estimated from the tensor-based training data via the VBEM algorithm. During the enhancement stage, the clean speech point source signal is estimated via the MNMF-based MVDR beamforming technique, whose realization involves two main steps. First, the speech source location is determined by observing the output powers of a DS beamformer applied to the MNMF-based pre-processed noisy speech signal. Second, the noise correlation matrix is

Table 5.2 Average results for Bus noise

Input SNR	Eval.	STSA	MNMF	NTF -KL	NTF -IS	VMNMF
-5 dB	PESQ	2.12	**2.36**	2.26	2.24	**2.36**
	SDR	-0.60	7.36	7.14	6.25	**8.77**
	Cov	2.45	**2.60**	2.47	2.47	2.50
0 dB	PESQ	2.43	2.63	2.57	2.54	**2.65**
	SDR	3.81	10.21	10.50	9.64	**11.72**
	Cov	2.84	2.91	2.91	2.86	**2.94**
5 dB	PESQ	2.74	2.91	2.87	2.84	**2.93**
	SDR	8.08	12.92	13.18	12.35	**14.05**
	Cov	3.21	3.21	3.30	3.22	**3.32**
10 dB	PESQ	2.99	3.18	3.16	3.11	**3.19**
	SDR	11.87	15.77	15.38	14.33	**16.13**
	Cov	3.49	3.48	3.62	3.52	**3.64**

Table 5.3 Average results for Cafe noise

Input SNR	Eval.	STSA	MNMF	NTF -KL	NTF -IS	VMNMF
-5 dB	PESQ	1.60	1.75	1.71	1.73	**1.83**
	SDR	-1.83	0.17	0.66	-0.55	**1.46**
	Cov	1.73	1.78	1.76	1.64	**1.92**
0 dB	PESQ	1.91	2.04	2.04	2.05	**2.17**
	SDR	3.58	5.36	5.49	5.02	**6.84**
	Cov	2.15	2.19	2.23	2.06	**2.41**
5 dB	PESQ	2.24	2.38	2.39	2.39	**2.50**
	SDR	8.21	9.80	9.49	9.25	**11.09**
	Cov	2.56	2.58	2.66	2.46	**2.83**
10 dB	PESQ	2.55	2.72	2.72	2.72	**2.80**
	SDR	12.03	13.65	12.41	12.38	**14.45**
	Cov	2.93	2.93	3.02	2.81	**3.17**

computed using the MNMF parameters for the magnitude components, and a combination of the noisy speech phase and steering vector for the phase components. Experimental results showed that the proposed Bayesian algorithm could provide better speech enhancement performance than the benchmark algorithms for different combinations of speaker and noise types under various input SNRs.

Table 5.4 Average results for Pedestrian noise

Input	Eval.	STSA	MNMF	NTF	NTF	VMNMF
SNR				-KL	-IS	
-5 dB	PESQ	1.64	1.78	1.78	1.86	**1.93**
	SDR	-1.81	-0.24	0.12	0.70	**2.58**
	Cov	1.82	1.91	1.86	1.91	**2.07**
0 dB	PESQ	1.99	2.12	2.12	2.18	**2.28**
	SDR	3.32	4.90	4.98	5.59	**7.66**
	Cov	2.29	2.35	2.36	2.37	**2.57**
5 dB	PESQ	2.32	2.47	2.47	2.48	**2.59**
	SDR	7.98	9.48	9.32	9.48	**11.61**
	Cov	2.69	2.76	2.78	2.77	**2.96**
10 dB	PESQ	2.63	2.80	2.76	2.77	**2.87**
	SDR	11.91	13.51	12.54	12.34	**14.79**
	Cov	3.04	3.10	3.11	3.11	**3.29**

Table 5.5 Average results for Street noise

Input	Eval.	STSA	MNMF	NTF	NTF	VMNMF
SNR				-KL	-IS	
-5 dB	PESQ	1.81	2.03	2.01	1.92	**2.15**
	SDR	-1.54	3.63	3.71	3.95	**5.61**
	Cov	2.06	**2.23**	2.12	1.90	2.21
0 dB	PESQ	2.17	2.34	2.34	2.26	**2.47**
	SDR	3.32	7.34	7.45	7.82	**9.59**
	Cov	2.52	2.63	2.61	2.42	**2.69**
5 dB	PESQ	2.50	2.66	2.65	2.58	**2.75**
	SDR	7.84	10.90	10.99	10.83	**12.72**
	Cov	2.92	3.00	3.03	2.86	**3.09**
10 dB	PESQ	2.79	2.97	2.95	2.87	**3.00**
	SDR	11.72	14.36	14.32	12.74	**15.23**
	Cov	3.25	3.32	3.38	3.23	**3.43**

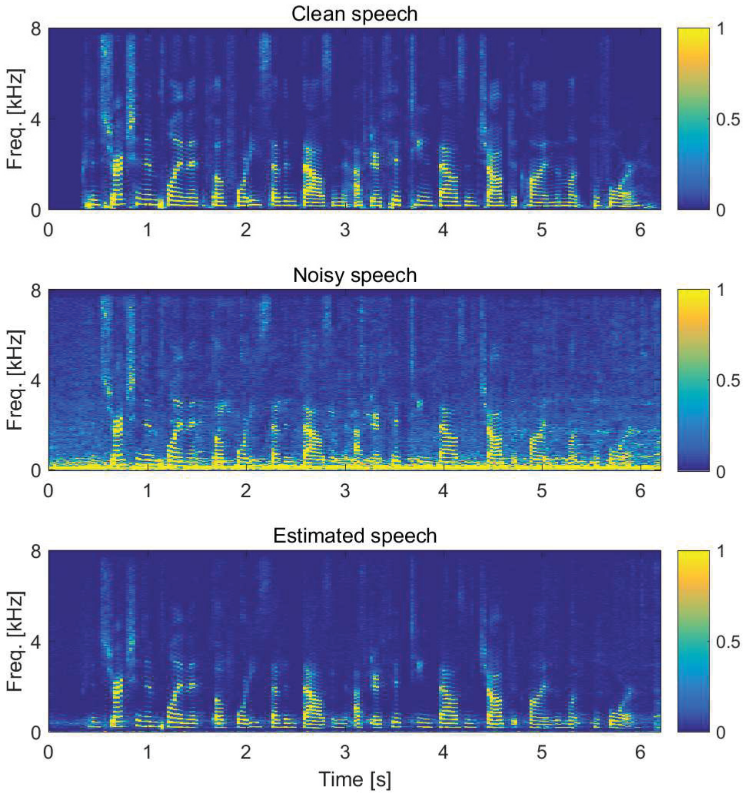

Fig. 5.3 Examples of magnitude spectrograms of the clean speech, noisy speech ($j = 1$) and estimated clean speech using the VMNMF method. A female speech is degraded by the Street noise at 0 dB input SNR.

Chapter 6

Conclusion and Future Works

In this thesis, we introduced and studied on training-based NMF algorithms for single and multi-channel speech enhancement.

After introducing the problem and reviewing background material in Chapter 2, we first presented a regularized NMF algorithm with Gaussian mixtures and masking model for single-channel speech enhancement in Chapter 3. In the proposed framework, *a priori* knowledge about the magnitude spectra of the clean speech and noise is captured by distinct GMMs, where normalized spectra are employed to handle the magnitude difference between the training and test data. The corresponding LLFs were included as regularization terms in the NMF cost function during the enhancement stage. Further improvement of the enhanced speech quality was obtained by exploiting the masking effects of the human auditory system. Specifically, we constructed a WWF where the weighting factor is selected based on the masking threshold calculated from the estimated clean speech PSD.

Second, we introduced a training and compensation algorithm of the class-conditioned basis vectors in the NMF model for single-channel speech enhancement in Chapter 4. We considered the PGM for both the NMF and classification models. The former is specified by a Poisson observation model, whereas the latter is specified by gamma class-conditional densities, which are used as *a priori* distribution for the basis vectors. During the training stage, the basis matrices for the clean speech and noises were estimated jointly by constraining them to belong to different classes. The parameters of the NMF model and PGM of classification were obtained by using the VBEM algorithm, which guarantees convergence to a stationary point. During the enhancement stage, we performed a noise classification followed by a basis compensation. The latter was implemented by using extra free basis vectors to capture features which are not included in the training data. The PGM parameters for classification were employed while estimating the free basis vectors as well as during the noise classification.

Third, we presented a novel algorithm for multi-channel speech enhancement in Chapter 5. Specifically, we considered the Poisson-distributed latent variables for MNMF, which corresponds to the KL-divergence within a statistical framework. During the training stage, the MNMF parameters were estimated from the tensor-based training data, by using the VBEM algorithm. During the enhancement stage, the clean speech point source signal was estimated via the NMF-based MVDR beamforming technique. Specifically, the clean speech and noise locations were determined by observing the output powers of the DS beamformer applied to the NMF-based pre-processed noisy speech signal. The noise correlation matrix was computed using the NMF parameters for the magnitude components, and the phase components were derived from the combination of the noisy speech phase and steering vector.

For each one of the above algorithm, objective experiments have been carried out for different combinations of speaker and noise types. The results showed that the proposed methods provide better speech enhancement performance than the selected benchmark algorithms under various conditions.

Finally, we comment on some interesting research avenues for further improving the enhancement performance regarding either single or multi-channel application. Firstly, we can consider modeling the basis vectors using a more accurate multimodal distribution, e.g., the gamma mixture model [55]. This extended prior modeling may also offer the potential of a noise-independent application (i.e., one universal basis matrix covering all noise types), by handling highly correlated noise sources. Secondly, to better deal with a highly reverberant environment, we can consider the ARMA processes to model the latent variables [43]. However, the computational complexity of such an approach makes it difficult to apply especially when training the basis vectors. To overcome this issue, we can consider different cost functions for the training and test stages. For example, we can train the basis vectors based on the Poisson-distributed latent variables during the training stage, and consider the ARMA-based latent variables during the enhancement stage.

Appendix A

Variational Lower Bound

In this appendix, we derive the variatoinal lower bound $\mathcal{L}_B(q(\boldsymbol{\theta}_L);\boldsymbol{\theta}_R)$ on the marginal LLF of the proposed class-conditioned NMF models: the bound regarding the single-channel NMF model given by (4.9) is described in Appendix A.1, while the bound regarding the multi-channel NMF model given by (5.12) is presented in Appendix A.2.

A.1 Single-channel Class-conditioned NMF Model

Based on (4.1), (4.2), (4.6) and (4.8), the logarithm of the full joint distribution is given by

$$\begin{aligned}
&\ln p(\mathbf{V},\mathbf{C},\mathbf{W},\mathbf{H};\boldsymbol{\theta}_R) \qquad \text{(A.1)}\\
&= \ln p(\mathbf{V}\,|\,\mathbf{C}) + \ln p(\mathbf{C}\,|\,\mathbf{W},\mathbf{H}) + \ln p(\mathbf{W};\boldsymbol{\theta}_C) + \ln p(\mathbf{H};\boldsymbol{\alpha}_h,\boldsymbol{\beta}_h)\\
&= \sum_{i=0}^{I_C-1}\sum_{k=1}^{K}\sum_{l=1}^{L_i} \ln\delta\Big(v_{kl}^i - \sum_{m=1}^{M_i} c_{kl}^{m,i}\Big)\\
&\quad + \sum_{i=0}^{I_C-1}\sum_{k=1}^{K}\sum_{m=1}^{M_i}\sum_{l=1}^{L_i}\Big(c_{kl}^{m,i}\ln(w_{km}^i h_{ml}^i) - w_{km}^i h_{ml}^i - \ln(c_{kl}^{m,i}!)\Big)
\end{aligned}$$

$$+\sum_{i=0}^{I_C-1}\sum_{m=1}^{M_i}\sum_{k=1}^{K}\Big((\alpha_{w,k}^i-1)\ln w_{km}^i-\frac{w_{km}^i}{\beta_{w,k}}-\ln\Gamma(\alpha_{w,k}^i)-\alpha_{w,k}^i\ln\beta_{w,k}\Big)$$
$$+\sum_{i=0}^{I_C-1}\sum_{m=1}^{M_i}\sum_{l=1}^{L_i}\Big((\alpha_h^i-1)\ln h_{ml}^i-\frac{\alpha_h^i}{\beta_h^i}h_{ml}^i-\ln\Gamma(\alpha_h^i)-\alpha_h^i\ln\Big(\frac{\beta_h^i}{\alpha_h^i}\Big)\Big).$$

The energy $\mathcal{L}_V(q(\boldsymbol{\theta}_L);\boldsymbol{\theta}_R)$ in (4.9) is simply found by evaluating the expectations of (A.1) with respect to $q(\mathbf{C},\mathbf{W},\mathbf{H})$ in (4.11)-(4.13), where the sufficient statistics are given by (4.19)-(4.21).

Based on (4.14), (4.16) and (4.18), and using the sufficient statistics in (4.19)-(4.21), the entropy $\mathcal{L}_E(q(\boldsymbol{\theta}_L))=-\,\mathbb{E}_q[\ln q(\boldsymbol{\theta}_L)]$ can be written as

$$\begin{aligned}
&\mathcal{L}_E(q(\boldsymbol{\theta}_L)) \qquad\qquad (A.2)\\
&=\sum_{i=0}^{I_C-1}\sum_{k=1}^{K}\sum_{l=1}^{L_i}\Big(-\ln(v_{kl}^i!)-\sum_{m=1}^{M_i}v_{kl}^i\bar{p}_{kl}^{m,i}\ln\bar{p}_{kl}^{m,i}\\
&\qquad-\mathbb{E}_q\Big[\ln\delta\Big(v_{kl}^i-\sum_{m=1}^{M_i}c_{kl}^{m,i}\Big)\Big]+\sum_{m=1}^{M_i}\mathbb{E}_q[\ln(c_{kl}^{m,i}!)]\Big)\\
&\quad-\sum_{i=0}^{I_C-1}\sum_{k=1}^{K}\sum_{m=1}^{M_i}\Big((\bar{\alpha}_{w,km}^i-1)\Psi(\bar{\alpha}_{w,km}^i)-\ln\bar{\beta}_{w,km}^i-\bar{\alpha}_{w,km}^i-\ln\Gamma(\bar{\alpha}_{w,km}^i)\Big)\\
&\quad-\sum_{i=0}^{I_C-1}\sum_{m=1}^{M_i}\sum_{l=1}^{L_i}\Big((\bar{\alpha}_{h,ml}^i-1)\Psi(\bar{\alpha}_{h,ml}^i)-\ln\bar{\beta}_{h,ml}^i-\bar{\alpha}_{h,ml}^i-\ln\Gamma(\bar{\alpha}_{h,ml}^i)\Big).
\end{aligned}$$

The lower bound on the marginal LLF, $\ln p(\mathbf{V};\boldsymbol{\theta}_R)$, is obtained by summing the energy and entropy terms as given by (4.9). Note that the terms in $\mathbb{E}_q[\cdot]$ in the third line in (A.2), which are analytically intractable, are canceled by their corresponding terms in the energy $\mathcal{L}_V(q(\boldsymbol{\theta}_L);\boldsymbol{\theta}_R)$ [41].

A.2 Multi-channel Bayesian NMF Model

Based on (5.2), (5.3), (5.9), (5.10) and (5.11), the logarithm of the full joint distribution is given by

$$
\begin{aligned}
&\ln p(\mathbf{V}, \mathbf{C}, \mathbf{A}, \mathbf{W}, \mathbf{H}; \boldsymbol{\theta}_R) \qquad\qquad (A.3)\\
&= \ln p(\mathbf{V} \,|\, \mathbf{C}) + \ln p(\mathbf{C} \,|\, \mathbf{A}, \mathbf{W}, \mathbf{H}) + \ln p(\mathbf{A}; \alpha_a, \beta_a) + \ln p(\mathbf{W}; \alpha_w, \beta_w) + \ln p(\mathbf{H}; \alpha_h, \beta_h)\\
&= \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L} \ln \delta\Big(v_{kl}^{j} - \sum_{m=1}^{M} c_{kl}^{m,j}\Big)\\
&\quad + \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{l=1}^{L} \Big(c_{kl}^{m,j} \ln(a_k^j w_{km} h_{ml}) - a_k^j w_{km} h_{ml} - \ln(c_{kl}^{m,j}!)\Big)\\
&\quad + \sum_{j=1}^{J}\sum_{k=1}^{K} \Big((\alpha_a - 1)\ln a_k^j - \frac{\alpha_a}{\beta_a} a_k^j - \ln\Gamma(\alpha_a) - \alpha_a \ln\Big(\frac{\beta_a}{\alpha_a}\Big)\Big)\\
&\quad + \sum_{k=1}^{K}\sum_{m=1}^{M} \Big((\alpha_w - 1)\ln w_{km} - \frac{\alpha_w}{\beta_w} w_{km} - \ln\Gamma(\alpha_w) - \alpha_w \ln\Big(\frac{\beta_w}{\alpha_w}\Big)\Big)\\
&\quad + \sum_{m=1}^{M}\sum_{l=1}^{L} \Big((\alpha_h - 1)\ln h_{ml} - \frac{\alpha_h}{\beta_h} h_{ml} - \ln\Gamma(\alpha_h) - \alpha_h \ln\Big(\frac{\beta_h}{\alpha_h}\Big)\Big).
\end{aligned}
$$

The energy $\mathcal{L}_V(q(\boldsymbol{\theta}_L); \boldsymbol{\theta}_R)$ in (5.12) is simply found by evaluating the expectations of (A.3) with respect to $q(\mathbf{C}, \mathbf{A}, \mathbf{W}, \mathbf{H})$ in (5.14)-(5.17), where the sufficient statistics are given by (5.23)-(5.26).

Based on (5.18), (5.20), (5.21) and (5.22), and using the sufficient statistics in (5.23)-(5.26), the entropy $\mathcal{L}_E(q(\boldsymbol{\theta}_L)) = -\mathbb{E}_q[\ln q(\boldsymbol{\theta}_L)]$ can be written as

$$\mathcal{L}_E(q(\boldsymbol{\theta}_L)) \qquad\qquad (A.4)$$

$$
\begin{aligned}
= \quad & \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}\Big(-\ln(v_{kl}^{j}!) - \sum_{m=1}^{M} v_{kl}^{j}\bar{p}_{kl}^{m,j}\ln\bar{p}_{kl}^{m,j} \\
& \qquad - \mathbb{E}_q\Big[\ln\delta\Big(v_{kl}^{j} - \sum_{m=1}^{M} c_{kl}^{m,j}\Big)\Big] + \sum_{m=1}^{M}\mathbb{E}_q[\ln(c_{kl}^{m,j}!)]\Big) \\
& - \sum_{j=1}^{J}\sum_{k=1}^{K}\Big((\bar{\alpha}_{a,k}^{j} - 1)\Psi(\bar{\alpha}_{a,k}^{j}) - \ln\bar{\beta}_{a,k}^{j} - \bar{\alpha}_{a,k}^{j} - \ln\Gamma(\bar{\alpha}_{a,k}^{j})\Big) \\
& - \sum_{k=1}^{K}\sum_{m=1}^{M}\Big((\bar{\alpha}_{w,km} - 1)\Psi(\bar{\alpha}_{w,km}) - \ln\bar{\beta}_{w,km} - \bar{\alpha}_{w,km} - \ln\Gamma(\bar{\alpha}_{w,km})\Big) \\
& - \sum_{m=1}^{M}\sum_{l=1}^{L}\Big((\bar{\alpha}_{h,ml} - 1)\Psi(\bar{\alpha}_{h,ml}) - \ln\bar{\beta}_{h,ml} - \bar{\alpha}_{h,ml} - \ln\Gamma(\bar{\alpha}_{h,ml})\Big).
\end{aligned}
$$

The lower bound on the marginal LLF, $\ln p(\mathbf{V};\boldsymbol{\theta}_R)$, is obtained by summing the energy and entropy terms as given by (5.12). As mentioned in Appendix A.1, the analytically intractable terms in $\mathbb{E}_q[\cdot]$ in the third line in (A.4) are canceled by their corresponding terms in the energy $\mathcal{L}_V(q(\boldsymbol{\theta}_L);\boldsymbol{\theta}_R)$.

Appendix B

A Brief Summary of NCP Method

In this appendix, we present a brief summary of the NCP method, which is our earlier work on training distinct basis vectors [101].

For a given training data set $\mathbf{V} = \{\mathbf{V}^i\}_{i=0}^{I_C-1}$, where $i \in \{0, ..., I_C - 1\}$ is the class index, the goal is to estimate $\mathbf{W} = \{\mathbf{W}^i\}_{i=0}^{I_C-1}$ and $\mathbf{H} = \{\mathbf{H}^i\}_{i=0}^{I_C-1}$ ($\mathbf{W}^i \in \mathbb{R}_+^{K\times M_i}$ and $\mathbf{H}^i \in \mathbb{R}_+^{M_i\times L_i}$). For the basis prior, we use the Gaussian-distributed PGM of classification, which is shown as

$$p(\mathbf{W};\boldsymbol{\theta}_C) = \prod_{i=0}^{I_C-1}\prod_{m=1}^{M_i} p_i\,\mathcal{N}(w_{km}^i;\mu_{ik},\sigma_k^2) \tag{B.1}$$

where p_i is the prior class probability, and $\boldsymbol{\theta}_C = \{p_i, \{\mu_{ik}, \sigma_k^2\}_{k=1}^K\}_{i=0}^{I_C-1}$ is the PGM parameter set for classification. For the activations, we employ sparse NMF regularization, which can be implemented by modeling the entries of $\mathbf{H}$ by an exponential distribution within a statistical framework [122]. Assuming that the entries are independent and identically distributed, the prior of $\mathbf{H}$ is shown as

$$p(\mathbf{H}) = \prod_{i=0}^{I_C-1} \lambda^{M_iL_i}\exp\left(-\lambda\sum_{m=1}^{M_i}\sum_{l=1}^{L_i} h_{ml}^i\right) \tag{B.2}$$

where the parameter λ controls the degree of sparsity.

The basis and activation matrices are obtained through a MAP estimator via the EM algorithm, which leads to maximizing the following criterion in the maximization step:

$$\mathcal{L}_C(\mathbf{V} \,|\, \mathbf{W}, \mathbf{H}; \boldsymbol{\theta}_C) = \sum_{i=0}^{I_C - 1} \mathcal{L}_C(\mathbf{V}^i \,|\, \mathbf{W}^i, \mathbf{H}^i) + \ln p(\mathbf{W}; \boldsymbol{\theta}_C) + \ln p(\mathbf{H}) \tag{B.3}$$

where $\mathcal{L}_C(\mathbf{V}^i \,|\, \mathbf{W}^i, \mathbf{H}^i)$ is given by (2.18). By setting the partial derivative of (B.3) with respect to w^i_{km} to zero, the update rule of w^i_{km} is found to be

$$(w^i_{km})^{(r+1)} = \frac{-q_{i1} + \sqrt{q_{i1}^2 + 4 q_0 q_{i2}}}{2 q_0} \tag{B.4}$$

where $q_0 = (\sigma_k^{-2})^{(r)}$, $q_{i1} = \sum_l (h^i_{ml})^{(r)} - \mu^{(r)}_{ik} (\sigma_k^{-2})^{(r)}$, $q_{i2} = \sum_l (\bar{c}^{m,i}_{kl})^{(r)}$, and the superscript (r) denotes the r-th iteration. Following a similar approach as for the basis estimation, the update rule of h^i_{ml} is obtained as

$$(h^i_{ml})^{(r+1)} = \frac{\sum_k (\bar{c}^{m,i}_{kl})^{(r)}}{\sum_k (w^i_{km})^{(r+1)} + \lambda}. \tag{B.5}$$

The set $\boldsymbol{\theta}_C$ is estimated by maximizing the marginal likelihood $p(\mathbf{V} \,|\, \mathbf{H}, \boldsymbol{\theta}_C)$, which becomes equivalent to maximizing (B.3) when we assume that $\mathbf{W}$ is well-determined [67]. The set $\boldsymbol{\theta}_C$ is then simply obtained by applying the ML criterion to $p(\mathbf{W} \,|\, \boldsymbol{\theta}_C)$ given by (B.1), where the resulting estimate in a closed form is interleaved with the EM update, as

$$p_i = \frac{M_i}{\sum_i M_i}, \qquad (\mu_{ik})^{(r+1)} = \frac{1}{M_i} \sum_{m=1}^{M_i} (w^i_{km})^{(r+1)} \tag{B.6}$$

$$(\sigma_k^2)^{(r+1)} = \frac{1}{M} \sum_{i=0}^{I_C - 1} \sum_{m=1}^{M_i} \left[(w^i_{km})^{(r+1)} - (\mu_{ik})^{(r+1)} \right]^2. \tag{B.7}$$

The enhancement stage is identical as presented in Subsection 2.2.1.

References

[1] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586-1604, Dec. 1979.

[2] P. Scalart and J. V. Filho, “Speech enhancement based on *a priori* signal to noise estimation,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 629-632, May 1996.

[3] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. 27, no. 2, pp. 113-120, Apr. 1979.

[4] N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system,” *IEEE Trans. Speech and Audio Process.*, vol. 7, no. 2, pp. 126-137, Mar. 1999.

[5] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

[6] C. H. You, S. N. Koh and S. Rahardja, “β-order MMSE spectral amplitude estimation for speech enhancement,” *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 4, pp. 475-486, July 2005.

[7] E. Plourde and B. Champagne, “Auditory-based spectral amplitude estimators for speech enhancement,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 8, pp. 1614-1623, Nov. 2008.

[8] Y. Ephraim and H. L. Van Trees, “A signal subspace approach for speech enhancement,” *IEEE Trans. Speech and Audio Process.*, vol. 3, no. 4, pp. 251-266, July 1995.

[9] S. H. Jensen, P. C. Hansen, S. D. Hansen and J. A. Sorensen, “Reduction of broadband noise in speech by truncated QSVD,” *IEEE Trans. Speech and Audio Process.*, vol. 3, no. 6, pp. 439-448, Nov. 1995.

[10] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 6, pp. 700-708, Nov. 2003.

[11] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 177-180, Apr. 1987.

[12] S. Gannot, D. Burshtein and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech and Audio Process.*, vol. 6, no. 4, pp. 373-385, July 1998.

[13] Z. Goh, K. -C. Tan and B. T. G. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Trans. Speech and Audio Process.*, vol. 7, no. 5, pp. 510-524, Sep. 1999.

[14] X. Anguera, C. Wooters and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 7, pp. 2011-2022, Sep. 2007.

[15] O. L. Frost, "An algorihtm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926-935, Aug. 1972.

[16] M. Souden, J. Benesty and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4925-4935, Sep. 2010.

[17] A. Krueger, E. Warsitz and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 1, pp. 206-219, Jan. 2011.

[18] S. Markovich, S. Gannot and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 6, pp. 1071-1086, June 2009.

[19] R. C. Hendriks, R. Heusdens, U. Kjems and J. Jensen, "On optimal multichannel mean-squared error estimators for speech enhancement," *IEEE Signal Process. Letters*, vol. 16, no. 10, pp. 885-888, Oct. 2009.

[20] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukic, T. Gerkmann, S. Doclo and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP J. Advances in Signal Process.*, no. 1, p. 61, Dec. 2015.

[21] J. S. Erkelens, J. Jensen and R. Heusdens, “Speech enhancement based on Rayleigh mixture modeling of speech spectral amplitude distributions,” in *Proc. European Signal Process. Conf. (EUSIPCO)*, pp. 65-69, Sep. 2007.

[22] G. -H. Ding, X. Wang, Y. Cao, F. Ding and Y. Tang, “Speech Enhancement Based on Speech Spectral Complex Gaussian Mixture Model,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 165-168, Mar. 2005.

[23] J. Hao, H. Attias, S. Nagarajan, T. -W. Lee and T. G. Sejnowski, “Speech enhancement, gain, and noise spectrum adaptation using approximate Bayesian estimation,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 1, pp. 24-37, Jan. 2009.

[24] J. Hao, T. -W. Lee and T. J. Sejnowski, “Speech enhancement using Gaussian scale mixture models,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 6, pp. 1127-1136, Aug. 2010.

[25] S. Chehrehsa and T. J. Moir, “Speech enhancement using maximum a-posteriori and Gaussian mixture models for speech and noise periodogram estimation,” *Computer Speech and Language*, vol. 36, pp. 58-71, Mar. 2016.

[26] Y. Ephraim, D. Malah and B. H. Juang, “On the application of hidden Markov models for enhancing noisy speech,” *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. 37, no. 12, pp. 1846-1856, Dec. 1989.

[27] A. P. Varga and R. K. Moore, “Hidden Markov Model decomposition of speech and noise,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 845-848, Apr. 1990.

[28] Y. Ephraim, “A Bayesian estimation approach for speech enhancement using hidden Markov models,” *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725-735, Apr. 1992.

[29] D. Y. Zhao and W. B. Kleijn, “HMM-based gain modeling for enhancement of speech in noise,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 882-892, Mar. 2007.

[30] S. Zafeiriou, A. Tefas, I. Buciu and I. Pitas, “Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification,” *IEEE Trans. Neural Networks*, vol. 17, no. 3, pp. 683-695, May 2006.

[31] N. Bertin, R. Badeau and E. Vincent, “Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 3, pp. 538-549, Mar. 2010.

[32] T. Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 1066-1074, Mar. 2007.

[33] A. Ozerov and C. Févotte, “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 3, pp. 550-563, Mar. 2010.

[34] N. Mohammadiha, P. Smaragdis and A. Leijon, “Supervised and unsupervised speech enhancement using nonnegative matrix factorization,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 10, pp. 2140-2151, Oct. 2013.

[35] J. Traa, P. Smaragdis, N. D. Stein and D. Wingate, “Directional NMF for joint source localization and separation,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1-5, Oct. 2015.

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

[37] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Proc. Advances in Neural Information Process. Systems (NIPS)*, pp. 556-562, 2001.

[38] C. Févotte, N. Bertin and J. -L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis,” *Neural Computation*, vol. 21, no. 3, pp. 793-830, Mar. 2009.

[39] C. Févotte and J. Idier, “Algorithms for nonnegative matrix factorization with the β-divergence,” *Neural Computation*, vol. 23, no. 9, pp. 2421-2456, Mar. 2011.

[40] E. M Grais and H. Erdogan, “Regularized nonnegative matrix factorization using Gaussian mixture priors for supervised single channel source separation,” *Computer Speech and Language*, vol. 27, no. 3, pp. 746-762, May 2013.

[41] A. T. Cemgil, “Bayesian inference for nonnegative matrix factorisation models,” *Computational Intelligence and Neuroscience*, no. 4, Article ID 785152, pp. 1-17, 2009.

[42] R. Badeau and A. Drémeau, “Variational Bayesian EM algorithm for modeling mixtures of non-stationary signals in the time-frequency domain (HR-NMF),” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 6171-6175, May 2013.

[43] R. Badeau and M. D. Plumbley, “Multichannel high-resolution NMF for modeling convolutive mixtures of non-stationary signals in the time-frequency domain,” *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 22, no. 11, pp. 1670-1680, Nov. 2014.

[44] Y. Bengio, A. Courville and P. Vincent,“Representation learning: a review and new perspectives,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798-1828, Aug. 2013.

[45] D. Ciregan, U. Meier and J. Schmidhuber, “Multi-column deep neural networks for image classification,” in *Proc. Computer Vision and Pattern Recognition (CVPR)*, pp. 3642-3649, June 2012.

[46] L. Deng, G. Hinton and B. Kingsbury, “New types of deep neural network learning for speech recognition and related applications: an overview,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 8599-8603, May 2013.

[47] Y. Xu, J. Du, L. -R. Dai and C. -H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 23, no. 1, pp. 7-19, Jan. 2015.

[48] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks and T. Zhang, “Learning spectral mapping for speech dereverberation and denoising,” *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 23, no. 6, pp. 982-992, June 2015.

[49] A. A. Nugraha, A. Liutkus and E. Vincent, “Multichannel audio source separation with deep neural networks,” *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 24, no. 9, pp. 1652-1664, Sep. 2016.

[50] P. -S. Huang, M. Kim, M. Hasegawa-Johnson and P. Smaragdis, “Joint optimization of masks and deep recurrent neural networks for monaural source separation,” *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 23, no. 12, pp. 2136-2147, Dec. 2015.

[51] F. Weninger, F. Eyben and B. Schuller, “Single-channel speech separation with memory-enhanced recurrent neural networks,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 3737-3741, May 2014.

[52] T. G. Kang, K. Kwon, J. W. Shin and N. S. Kim, “NMF-based target source separation using deep neural network,” *IEEE Signal Process. Letters*, vol. 22, no. 2, pp. 229-233, Feb. 2015.

[53] S. Nie, S. Linag, H. Li, X. Zhang, Z. Yang, W. Liu and L. Dong, “Exploiting spectro-temporal structures using NMF for DNN-bsaed supervised speech separation,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 469-473, Mar. 2016.

[54] G. J. Mysore and P. Smaragdis, “A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 17-20, May 2011.

[55] T. Virtanen and A. T. Cemgil, “Mixtures of gamma priors for non-negative matrix factorization based speech separation,” in *Proc. Independent Component Analysis and Signal Separation*, pp. 646-653, Mar. 2009.

[56] K. Kwon, J. W. Shin and N. S. Kim, “NMF-based speech enhancement using bases update,” *IEEE Signal Process. Letters*, vol. 22, no. 4, pp. 450-454, Apr. 2015.

[57] Z. Wang and F. Sha, “Discriminative non-negative matrix factorization for single-channel speech separation,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 3749-3753, May, 2014.

[58] P. Sprechmann, A. M. Bronstein and G. Sapiro, “Supervised non-Euclidean sparse NMF via bilevel optimization with applications to speech enhancement,” in *Proc. 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, pp. 1-5, May, 2014.

[59] F. Weninger, J. L. Roux, J. R. Hershy and S. Watanabe, “Discriminative NMF and its application to single-channel source separation,” in *Proc. Interspeech*, pp. 865-869, Sep. 2014.

[60] R. Badeau, N. Bertin and E. Vincent, “Stability analysis of multiplicative update algorithms and application to nonnegative matrix factorization,” *IEEE Trans. Neural Networks*, vol. 21, no. 12, pp. 1869-1881, Dec. 2010.

[61] H. Sawada, H. Kameoka, S. Araki and N. Ueda, “Multichannel extensions of non-negative matrix factorization with complex-valued data,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 971-982, May 2013.

[62] J. Nikunen and T. Virtanen, “Direction of arrival based spatial covariance model for blind sound source separation,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 22, no. 3, pp. 727-739, Mar. 2014.

[63] M. Nakano, H. Kameoka, J. L. Roux, Y. Kitano, N. Ono and S. Sagayama, “Convergence-guaranteed multiplicative algorihtms for nonnegative matrix factorization with β-divergence,” in *Proc. IEEE Int. Workshop on Machine Learning for Signal Process. (MLSP)*, pp. 283-288, Aug. 2010.

[64] J. Eggert and E. Korner, “Sparse coding and NMF,” in *Proc. Int. Joint Conf. Neural Networks*, pp. 486-491, Oct. 2008.

[65] C. Févotte and A. T. Cemgil, “Nonnegative matrix factorizations as probabilistic inference in composite models,” in *Proc. European Signal Process. Conf. (EUSIPCO)*, pp. 1913-1917, Aug. 2009.

[66] A. Dempster, N. M. Laird and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. Royal Statistical Soc. Series B (Methodological).*, vol. 39, no. 1, pp. 1-38, Jan. 1977.

[67] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[68] E. M. Grais and H. Erdogan, “Source separation using regularized NMF with MMSE estimates under GMM priors with online learning for the uncertainties,“ *Digital Signal Process.*, vol. 29, pp. 20-34, Mar. 2014.

[69] C. Févotte and A. Ozerov, “Notes on nonnegative tensor factorization of the spectrogram for audio source separation: Statistical insights and towards self-clustering of the spatial cues,” in *Proc. Int. Symposium on Computer Music Modeling and Retrieval*, pp. 102-115, June 2010.

[70] D. O’Shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, 1987.

[71] T. F. Quatieri, *Discrete-time Speech Signal Processing: Principles and Practice*, Pearson Education, 2008.

[72] N. Mohammadiha, T. Gerkmann and A. Leijon, “A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization,” in *Proc. IEEE Workshop on Applications of Signal Process. to Audio and Acoustics (WASPAA)*, pp. 45-48, Oct. 2011.

[73] D. FitzGerald, M. Cranitch and E. Coyle, “On the use of the beta divergence for musical source separation,” in *Proc. Irish Signals and Systems Conference*, 2009.

[74] H. Kameoka, H. Kagami and M. Yukawa, “Complex NMF with the generalized Kullback-Leibler divergence,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 56-60, Mar. 2017.

[75] J. B. Allen, “Short term spectral analysis, synthesis, and modification by discrete Fourier transform,” *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. 25, no. 3, pp. 235-238, June 1977.

[76] P. Kabal, *Windows for Transform Processing.* Tech. Rep., McGill University, 2005.

[77] T. T. Vu, B. Bigot and E. S. Chng, “Speech enhancement using beamforming and non negative matrix factorization for robust speech recognition in the CHiME-3 challenge,” in *Proc. Automatic Speech Recognition and Understanding (ASRU)*, pp. 423-429, Dec. 2015.

[78] R. C. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 1, pp. 223-233, Jan. 2012.

[79] H. Chung, E. Plourde and B. Champagne, "Regularized NMF-based speech enhancement with spectral components modeled by Gaussian mixtures," in *Proc. IEEE Int. Workshop on Machine Learning for Signal Process. (MLSP)*, six pages, Sep. 2014.

[80] H. Chung, E. Plourde and B. Champagne, "Regularized non-negative matrix factorization with Gaussian mixtures and masking model for speech enhancement," *Speech Communication*, vol. 87, pp. 18-30, Mar. 2017.

[81] E. M. Grais and H. Erdogan, "Hidden Markov models as priors for regularized non-negative matrix factorization in single-channel source separation," in *Proc. Annual Int. Symposium on Computer Architecture (ISCA)*, pp. 1536-1539, June 2012.

[82] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, Springer Science & Business Media, vol. 22, 2007.

[83] Y. Hu and P. C. Loizou, "Incorporating a psychoacoustical model in frequency domain speech enhancement," *IEEE Signal Process. Letters*, vol. 11, no. 2, pp. 270-273, Feb. 2004.

[84] A. Natarajan, J. H. L. Hansen, K. H. Arehart and J. Rossi-Katz, "An auditory-masking-threshold-based noise suppression algorithm GMMSE-AMT [ERB] for listeners with sensorineural hearing loss," *EURASIP Journal on Applied Signal Process.*, no. 18, pp. 2938-2953, Dec. 2005.

[85] J. H. Hansen, V. Radhakrishnan and K. H. Arehart, "Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 6, pp. 2049-2063, Nov. 2006.

[86] S. Kırbız and B. Günsel, "Perceptually enhanced blind single-channel music source separation by non-negative matrix factorization," *Digital Signal Process.*, vol. 23, no. 2, pp. 646-658, Mar. 2013.

[87] T. Virtanen, *Monaural sound source separation by perceptually weighted non-negative matrix factorization.* Tech. Rep., Tampere University of Technology, 2007.

[88] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451-515, Apr. 2000.

[89] A. Spriet, M. Moonen and J. Wouters, “Stochastic gradient-based implementation of spatially preprocessed speech distortion weighted multichannel Wiener filtering for noise reduction in hearing aids,” *IEEE Trans. Signal Process.*, vol. 53, np. 3, pp. 911-925, Mar. 2005.

[90] S. Gustafsson, P. Jaz and P. Vary, “A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 397-400, May 1998.

[91] B. Defraene, K. Ngo, T. V. Waterschoot, M. Diehl and M. Moonen, “A psychoacoustically motivated speech distortion weighted multi-channel Wiener filter for noise reduction,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 4637-4640, Mar. 2012.

[92] I. Kodrasi, D. Marquardt and S. Doclo, “Curvature-based optimization of the trade-off parameter in the speech distortion weighted multichannel Wiener filter,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 315-319, Apr. 2015.

[93] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 4, pp. 1383-1393, May 2012.

[94] J. Nikunen and T. Virtanen, “Noise-to-mask ratio minimization by weighted non-negative matrix factorization,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 25-28, Mar. 2010.

[95] P. Kabal, *TSP Speech Database.* Tech. Rep., McGill University, Montreal, Canada, 09-02, 2002.

[96] M. Cooke, J. Barker, S. Cunningham and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *J. Acoust. Soc. America*, vol. 120, no. 5, pp. 2421-2424, Nov. 2008.

[97] A. Varga and H. J. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, no. 3, pp. 247-251, July 1993.

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

[99] E. Vincent, R. Gribonval and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech and Language Process.*, vol. 14, no. 4, pp. 1462-1469, July 2006.

[100] H. Chung, E. Plourde and B. Champagne, "Basis compensation in non-negative matrix factorization model for speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 2249-2253, Mar. 2016.

[101] H. Chung, E. Plourde and B. Champagne, "Discriminative training of NMF model based on class probabilities for speech enhancement," *IEEE Signal Process. Letters,* vol. 23, no. 4, pp. 502-506, Feb. 2016.

[102] H. Chung, E. Plourde and B. Champagne, "Single-channel enhancement of convolutive noisy speech based on a discriminative NMF algorithm," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 2302-2306, Mar. 2017.

[103] H. Chung, R. Badeau, E. Plourde and B. Champagne, "Training and compensation of class-conditioned NMF bases for speech enhancement," *Neurocomputing*, vol. 284, pp. 107-118, Apr. 2018.

[104] E. M. Grais and H. Erdogan, "Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation," in *Proc. Interspeech*, pp. 808-812, Aug. 2013.

[105] K. Kwon, J. W. Shin and N. S. Kim, "Target source separation based on discriminative nonnegative matrix factorization incorporating cross-reconstruction error," *IEICE Trans. Information and Systems*, vol. 98, no. 11, pp. 2017-2020, Nov. 2015.

[106] E. M. Grais and H. Erdogan, "Adaptation of speaker-specific bases in non-negative matrix factorization for single channel speech-music separation," in *Proc. Interspeech*, pp. 569-572, Aug. 2011.

[107] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183-233, Nov. 1999.

[108] J. -T. Chien and P. -K. Yang, "Bayesian factorization and learning for monaural source separation," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 24, no. 1, pp. 185-195, Jan. 2016.

[109] M. D. Hoffman, "Poisson-Uniform nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. 5361-5364, Mar. 2012.

[110] I. Ulusoy and C. M. Bishop, "Generative versus discriminative models for object recognition," in *Proc. Computer Vision and Pattern Recognition*, pp. 258-265, June 2005.

[111] L. K. Saul and D. D. Lee, "Multiplicative updates for classification by mixture models," in *Proc. Advances in Neural Information Process. Systems (NIPS)*, vol. 14, pp. 897-904, Dec. 2001.

[112] S.M. Kim, J. H. Park, H. K. Kim, S. J. Lee and Y. K. Lee, "Non-negative matrix factorization based noise reduction for noise robust automatic speech recognition," *Lecture Notes in Computer Science*, vol. 7191, pp. 338-346, 2012.

[113] M. Sun, Y. Li, F. F. Gemmeke and X. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback-Leibler divergence," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 23, no. 7, pp. 1233-1242, July 2015.

[114] Z. Duan, A. J. Mysore and P. Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environment, in *Proc. Interspeech*, pp. 595-598, Sep. 2012.

[115] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, vol. 46, pp. 535-557, Dec. 2016.

[116] A. Ozerov, C. Févotte, R. Blouet and J. -J. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pp. , May 2011.

[117] S. Mirzaei, Y. Norouzi and H. V. Hamme, "Two-stage blind audio source counting and separation of stereo instantaneous mixtures using Bayesian tensor factorisation," *IET Signal Process.*, vol. 9, no. 8, pp. 587-595, Sep. 2015.

[118] C. Blandin, A. Ozerov and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Process.*, vol. 92, no. 8, pp. 1950-1960, Aug. 2012.

[119] J. Benesty, J. Chen and Y. Huang, *Topics in Signal Processing: Microphone Array Signal Processing, vol. 1.* Springer, 2008.

[120] X. Mestre and M. A. Launas, "On diagonal loading for minimum variance beamformers," in *Proc. IEEE Int. Symposium on Signal Process. and Information Technology (ISSPIT)*, pp. 459-462, Dec. 2003.

[121] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 1, pp. 229-238, Jan. 2008.

[122] M. N. Schmidt and J. Larsen, “Reduction of non-stationary noise using a non-negative latent variable decomposition,” in *Proc. IEEE Int. Workshop on Machine Learning for Signal Process. (MLSP)*, pp. 486-491, Oct. 2008.