Thesis for the degree of Doctor of Philosophy

Speech Enhancement Using Nonnegative Matrix Factorization and Hidden Markov Models

Nasser Mohammadiha

KTH Electrical Engineering

Communication Theory Laboratory
School of Electrical Engineering
KTH Royal Institute of Technology

Stockholm 2013

Mohammadiha, Nasser
Speech Enhancement Using Nonnegative Matrix Factorization and Hidden Markov Models

ISBN 978-91-7501-833-1
TRITA-EE 2013:030
ISSN 1653-5146

Communication Theory Laboratory
School of Electrical Engineering
KTH Royal Institute of Technology
SE-100 44 Stockholm, Sweden

Abstract

Reducing interference noise in a noisy speech recording has been a challenging task for many years yet has a variety of applications, for example, in handsfree mobile communications, in speech recognition, and in hearing aids. Traditional single-channel noise reduction schemes, such as Wiener filtering, do not work satisfactorily in the presence of non-stationary background noise. Alternatively, supervised approaches, where the noise type is known in advance, lead to higher-quality enhanced speech signals. This dissertation proposes supervised and unsupervised single-channel noise reduction algorithms. We consider two classes of methods for this purpose: approaches based on nonnegative matrix factorization (NMF) and methods based on hidden Markov models (HMM).

The contributions of this dissertation can be divided into three main (overlapping) parts. First, we propose NMF-based enhancement approaches that use temporal dependencies of the speech signals. In a standard NMF, the important temporal correlations between consecutive short-time frames are ignored. We propose both continuous and discrete state-space nonnegative dynamical models. These approaches are used to describe the dynamics of the NMF coefficients or activations. We derive optimal minimum mean squared error (MMSE) or linear MMSE estimates of the speech signal using the probabilistic formulations of NMF. Our experiments show that using temporal dynamics in the NMF-based denoising systems improves the performance greatly. Additionally, this dissertation proposes an approach to learn the noise basis matrix online from the noisy observations. This relaxes the assumption of an a-priori specified noise type and enables us to use the NMF-based denoising method in an unsupervised manner. Our experiments show that the proposed approach with online noise basis learning considerably outperforms state-of-the-art methods in different noise conditions.

Second, this thesis proposes two methods for NMF-based separation of sources with similar dictionaries. We suggest a nonnegative HMM (NHMM) for babble noise that is derived from a speech HMM. In this approach, speech and babble signals share the same basis vectors, whereas the activation of the basis vectors are different for the two signals over time. We derive an MMSE estimator for the clean speech signal using the proposed NHMM. The objective evaluations and performed subjective listening test show that the

proposed babble model and the final noise reduction algorithm outperform the conventional methods noticeably. Moreover, the dissertation proposes another solution to separate a desired source from a mixture with arbitrarily low artifacts.

Third, an HMM-based algorithm to enhance the speech spectra using super-Gaussian priors is proposed . Our experiments show that speech discrete Fourier transform (DFT) coefficients have super-Gaussian rather than Gaussian distributions even if we limit the speech data to come from a specific phoneme. We derive a new MMSE estimator for the speech spectra that uses super-Gaussian priors. The results of our evaluations using the developed noise reduction algorithm support the super-Gaussianity hypothesis.

Keywords: Speech enhancement, noise reduction, nonnegative matrix factorization, hidden Markov model, probabilistic latent component analysis, online dictionary learning, super-Gaussian distribution, MMSE estimator, temporal dependencies, dynamic NMF.

List of Papers

The thesis is based on the following papers:

[A] 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 Applications of Signal Process. Audio Acoustics (WASPAA)*, oct. 2011, pp. 45–48.

[B] 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.

[C] N. Mohammadiha and A. Leijon, "Nonnegative HMM for Babble Noise Derived From Speech HMM: Application to Speech Enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 998–1011, may 2013.

[D] N. Mohammadiha, R. Martin, and A. Leijon, "Spectral Domain Speech Enhancement Using HMM State-dependent Super-Gaussian Priors," *IEEE Signal Process. Letters*, vol. 20, no. 3, pp. 253–256, mar. 2013.

[E] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Prediction Based Filtering and Smoothing to Exploit Temporal Dependencies in NMF," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, may 2013, pp. 873–877.

[F] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Low-artifact Source Separation Using Probabilistic Latent Component Analysis," in *Proc. IEEE Workshop Applications of Signal Process. Audio Acoustics (WASPAA)*, oct. 2013.

In addition to papers A-F, the following papers have also been produced in part by the author of the thesis:

[1] P. Smaragdis, C. Févotte, N. Mohammadiha, G. J. Mysore, M. Hoffman , "A Unified View of Static and Dynamic Source Separation Using Non-Negative Factorizations", *IEEE Signal Process. Magazine: Special Issue on Source Separation and Applications*, to be submitted.

[2] G. Panahandeh, N. Mohammadiha, A. Leijon, P. Händel, "Continuous Hidden Markov Model for Pedestrian Activity Classification and Gait Analysis," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 5, pp. 1073–1083, may 2013.

[3] N. Mohammadiha, P. Smaragdis, A. Leijon, "Simultaneous Noise Classification and Reduction Using a Priori Learned Models ," *IEEE Int. Workshop on Machine Learning for Signal Process. (MLSP)*, sep. 2013.

[4] N. Mohammadiha, W. B. Kleijn, A. Leijon, "Gamma Hidden Markov Model as a Probabilistic Nonnegative Matrix Factorization ," in *Proc. European Signal Process. Conf. (EUSIPCO)*, sep. 2013.

[5] N. Mohammadiha, J. Taghia, and A. Leijon, "Single Channel Speech Enhancement Using Bayesian NMF With Recursive Temporal Updates of Prior Distributions," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, mar. 2012, pp. 4561–4564.

[6] J. Taghia, N. Mohammadiha, and A. Leijon, "A Variational Bayes Approach to the Underdetermined Blind Source Separation with Automatic Determination of the Number of Sources," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, mar. 2012, pp. 253–256.

[7] H. Hu, N. Mohammadiha, J. Tahgia, A. Leijon, M. E Lutman, S. Wang, "Sparsity Level Selection of a Non-Negative Matrix Factorization Based Speech Processing Strategy in Cochlear Implants," in *Proc. European Signal Process. Conf. (EUSIPCO)*, aug. 2012, pp. 2432–2436.

[8] G. Panahandeh, N. Mohammadiha, and M. Jansson, "Ground Floor Feature Detection for Mobile Vision-Aided Inertial Navigation," in *Proc. Int. Conf. on Intelligent Robots and Systems (IROS)*, oct. 2012, pp. 3607–3611.

[9] G. Panahandeh, N. Mohammadiha, A. Leijon, and P. Händel, "Chest-Mounted Inertial Measurement Unit for Human Motion Classification Using Continuous Hidden Markov Model," in *Proc. IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, may 2012, pp. 991–995.

[10] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A New Approach for Speech Enhancement Based on a Constrained Nonnegative Matrix Factorization," in *Proc. IEEE Int. Symposium on. Intelligent Signal Process. and Communication Systems (ISPACS)*, dec. 2011, pp. 1–5.

[11] N. Mohammadiha and A. Leijon, "Model Order Selection for Non-Negative Matrix Factorization with Application to Speech Enhancement," *KTH Royal Institute of Technology, Tech. Rep.*, 2011.

[12] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An Evaluation of Noise Power Spectral Density Estimation Algorithms in Adverse Acoustic Environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, may 2011, pp. 4640–4643.

[13] H. Hu, J. Taghial, J. Sang, J. Taghia, N. Mohammadiha, M. Azarpour, R. Dokku, S.Wang, M. Lutman, and S. Bleeck, "Speech Enhancement Via Combination of Wiener Filter and Blind Source Separation," in *Proc. Springer Int. Conf. on Intelligent Systems and Knowledge Engineering (ISKE)*, dec. 2011, pp. 485–494.

[14] N. Mohammadiha and A. Leijon, "Nonnegative Matrix Factorization Using Projected Gradient Algorithms with Sparseness Constraints," in *Proc. IEEE Int. Symposium on Signal Process. and Information Technology (ISSPIT)*, dec. 2009, pp. 418–423.

Acknowledgements

The years have passed quickly and it has become time to write and defend my PhD dissertation. When I look back on my time as a student, I see the support of many individuals who have assisted me during these years. I would like to take this opportunity to acknowledge all those who have encouraged and helped me.

First and foremost, I would like to sincerely thank my supervisor, Prof. Arne Leijon. Your deep knowledge of the field, honesty, and your sense of responsibility brought me a very effective supervision. I am highly grateful that you gave me the freedom to explore different ideas and enhanced them with your valuable suggestions. Throughout my PhD, I always felt at ease discussing my problems with you, and I know from experience that you always were ready to help me in different aspects. I would also like to express my great appreciation to my principal supervisor Prof. W. Bastiaan Kleijn, who gave me the opportunity to begin my doctoral studies at KTH Royal Institute of Technology. Your professionalism has influenced me a lot and I highly value your suggestions.

I would like to thank Prof. Paris Smaragdis for giving me the opportunity to visit his group at University of Illinois at Urbana-Champaign (UIUC). Your creativity, friendly discussions, and on-the-spot feedback were always of excellent quality. I am also grateful to my colleagues at UIUC, especially Minje Kim and Johannes Traa for the interesting discussions.

Three years of my research was funded by the AUDIS project. I would like to extend my thanks to everyone involved in the project, especially the board members. In particular, I wish to express my gratitude to the project coordinator Prof. Rainer Martin. I benefited a lot from your fruitful suggestions, both during and after the project. Special thanks go to Dr. Stefan Bleeck for his great support during my visits to University of Southampton.

I wish to thank all my current and past colleagues at Osquldas väg 10, including the always supportive Associate Prof. Markus Flierl and Prof. Peter Händel. Special thanks go to Assistant Prof. Timo Gerkmann, Dr. Saikat Chatterjee, Dr. Cees Taal, Jalil Taghia, Gustav Eje Henter, Dr. Zhanyu Ma, and Petko Petkov for constructive discussions regarding my research. I also enjoyed the experience of teaching with Prof. Kleijn, Associate Prof. Flierl, Dr. Minyue Li, Pravin Kumar Rana, and Haopeng Li. I am grateful to Dora

Söderberg for her support in various administrative matters.

I am indebted to Petko Petkov, Obada Alhaj Moussa, Jalal Taghia, Du Liu, and Jalil Taghia for proofreading the summary part of my thesis.

Moving to a new country can be a challenging experience. Obada, I greatly appreciate your generous support when my wife and I moved to Sweden. I would like to thank Alla, Farshad, Sadegh, and Nima for helping me to relocate when I was in USA. I would also like to thank my friends at UIUC. Negin, Mohammad, Vahid, and Mostafa, our game evenings at Urbana-Champaign and the fun we have had together are unforgettable.

Finally, I would like to thank my parents and parents-in-law. Without your tremendous support and love, pursuing a PhD would have been impossible. Most importantly, I would like to thank my dear wife Ghazaleh who has been hugely supportive in my life. Your love and belief in me has been an unparalleled source of strength for me throughout these years.

Nasser Mohammadiha
Stockholm, October 2013

Contents

Abstract i

List of Papers iii

Acknowledgements vii

Contents ix

Acronyms xiii

I Summary 1

1 Theoretical Background 1
1.1 Speech Enhancement Background 1
1.2 Single-channel Speech Enhancement 3
1.2.1 Wiener Filter 3
1.2.2 Kalman Filter 6
1.2.3 Estimators Using Super-Gaussian Priors . 6
1.3 Hidden Markov Model 10
1.3.1 HMM-based Speech Enhancement 11
1.4 Nonnegative Matrix Factorization 14
1.4.1 Probabilistic NMF 15
1.4.2 Source Separation and Speech Enhancement Using NMF 18
2 Methods and Results 21
2.1 Speech Enhancement Using Dynamic NMF 22
2.1.1 Speech Denoising Using Bayesian NMF . . 24
2.1.2 Nonnegative Linear Dynamical Systems . 28
2.1.3 Nonnegative Hidden Markov Model 30
2.2 NMF-based Separation of Sources with Similar Dictionaries 30

2.3 Super-Gaussian Priors in HMM-based Enhancement Systems . . . 32
2.4 Discussion . . . 35
3 Conclusions . . . 37
References . . . 38

II Included papers 53

A A New Linear MMSE Filter for Single Channel Speech Enhancement Based on Nonnegative Matrix Factorization A1
1 Introduction . . . A1
2 Notation and Basic Concepts . . . A2
3 Noise PSD estimation Using NMF . . . A3
4 Linear MMSE Filter Based on NMF . . . A4
5 Evaluation . . . A6
5.1 Results and Discussion . . . A7
6 Conclusions . . . A9
References . . . A10

B Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization B1
1 Introduction . . . B1
2 Review of State-of-the-art NMF-Based Speech Enhancement B4
3 Speech Enhancement Using Bayesian NMF . . . B8
3.1 BNMF-HMM for Simultaneous Noise Classification and Reduction . . . B9
3.2 Online Noise Basis Learning for BNMF . . . B13
3.3 Informative Priors for NMF Coefficients . . . B16
4 Experiments and Results . . . B17
4.1 Noise Reduction Using a-Priori Learned NMF Models B19
4.2 Experiments with Unsupervised Noise Reduction . B23
5 Conclusions . . . B25
References . . . B27

C Nonnegative HMM for Babble Noise Derived From Speech HMM: Application to Speech Enhancement C1
1 Introduction . . . C1
2 Speech Signal Model . . . C4
2.1 Single-voice Gamma HMM . . . C4
2.2 Gamma-HMM as a Probabilistic NMF . . . C6
3 Probabilistic Model of Babble Noise . . . C7
4 Speech Enhancement Method . . . C10
4.1 Clean Speech Mixed with Babble . . . C10

4.2 Clean Speech Estimator C11
5 Parameter Estimation C13
5.1 Speech Model Training C13
5.2 Babble Model Training C15
5.3 Updating Time-varying Parameters C16
6 Experiments and Results C19
6.1 System Implementation C20
6.2 Evaluations C21
6.2.1 Objective Evaluation of the Noise Reduction C21
6.2.2 Effect of Systems on Speech and Noise Separately C21
6.2.3 Effect of the Number of Speakers in Babble C23
6.2.4 Cross-predictive Test for Model Fitting . . C24
6.2.5 Subjective Evaluation of the Noise Reduction C26
7 Conclusion C28
8 Appendix C28
8.1 MAP Estimate of the Gain Variables C28
8.2 Posterior Distribution of the Gain Variables C30
8.3 Gradient and Hessian for Babble States C31
References C31

D Spectral Domain Speech Enhancement Using HMM State-dependent Super-Gaussian Priors D1
1 Introduction D1
2 Conditional Distribution of the Speech Power Spectral Coefficients D2
2.1 Experimental Data D3
3 HMM-based Speech Enhancement D4
3.1 Speech Model D4
3.2 Noise Model D5
3.3 Speech Estimation: Complex Gaussian Case D6
3.4 Speech Estimation: Erlang-Gamma Case D7
4 Experiments and Results D9
5 Conclusion D10
References D10

E Prediction Based Filtering and Smoothing to Exploit Temporal Dependencies in NMF E1
1 Introduction E1
2 Proposed Method E2
2.1 Background E3
2.2 Filtering E4
2.3 Smoothing E5
2.4 Source Separation Using the Proposed Method . . . E5

3 Experiments and Results E6
3.1 Separation of Speech and Its Time-reversed Version E6
3.2 Speech Denoising E7
3.3 Speech Source Separation E9
4 Conclusion E10
References E11

F Low-artifact Source Separation Using Probabilistic Latent Component Analysis F1
1 Introduction F1
2 Proposed Solution F3
2.1 PLCA: A Review F3
2.2 PLCA with Exponential Priors F4
2.3 Example: Separation of Sources with One Common Basis Vector F5
2.4 Identifying Common Bases F7
3 Experiments Using Speech Data F8
3.1 Source Separation F8
3.2 Reducing Babble Noise F9
4 Conclusions F10
References F10

Acronyms

AMAP	Approximate Maximum A-Posteriori
ASR	Automatic Speech Recognizer
BNMF	Bayesian NMF
CCCP	Concave-Convex Procedure
DFT	Discrete Fourier Transform
EM	Expectation Maximization
ETSI	European Telecommunications Standards Institute
GARCH	Generalized Autoregressive Conditional Heteroscedasticity
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
i.i.d.	Independent and Identically Distributed
IS	Itakura-Saito
KL	Kullback-Leibler
LDA	Latent Dirichlet Allocation
LMMSE	Linear Minimum Mean Squared Error
MAP	Maximum A-Posteriori
MFCC	Mel-Frequency Cepstral Coefficient
MIXMAX	Mixture-Maximization
ML	Maximum Likelihood
NMF	Nonnegative Matrix Factorization
NHMM	Nonnegative Hidden Markov Model

MSE	Mean Square Error
MTD	Mixture Transition Distribution
PESQ	Perceptual Evaluation of Speech Quality
PLCA	Probabilistic Latent Component Analysis
PLSI	Probabilistic Latent Semantic Indexing
PSD	Power Spectral Density
SAR	Source to Artifact Ratio
SDR	Source to Distortion Ratio
SIR	Source to Interference Ratio
SNR	Signal to Noise Ratios
STSA-GenGamma	Speech Short-time Spectral Amplitude Estimator Using Generalized Gamma Prior Distributions
VAD	Voice Activity Detector
VAR	Vector Autoregressive

Part I

Summary

1 Theoretical Background

1.1 Speech Enhancement Background

Speech enhancement or noise reduction has been one of the main investigated problems in the speech community for a long time. The problem arises whenever a desired speech signal is degraded by some disturbing noise. The noise can be additive or convolutive. In practice, a convolutive noise should be rather considered due to the reverberation. However, it is usually assumed that the noise is additive since it makes the problem simpler and also the developed algorithms based on this assumption lead to satisfactory results in practice [1, 2]. Even this additive noise can reduce the quality and intelligibility of the speech signal considerably. Therefore, the aim of the noise reduction algorithms is to estimate the clean speech signal from the noisy recordings in order to improve the quality and intelligibility of the enhanced signal [1, 2]. Figure 1 shows a simplified diagram of a speech enhancement system in which the noise is assumed to be additive.

There are various applications of speech enhancement in our daily life. For example, consider a mobile communication where you are located in a noisy environment, e.g., a street or inside a car. Here, a noise reduction approach can be used to make the communication easier by reducing the interfering noise. A similar approach can be used in communications over internet, such as Skype or Google Talk. Speech enhancement algorithms can be also used to design robust speech/speaker recognition systems by reducing the mismatch between the training and testing stages. In this case, a speech enhancement approach is applied to reduce the noise before extracting a set of features.

Another very important application of the noise reduction is for users of hearing aids or cochlear implants. Since speech signals are highly redundant, normal hearing people can understand a target speech signal even at low signal to noise ratios (SNR). For instance, normal hearing people can understand up to 50% of the words from a babble-corrupted noisy speech signal at a 0 dB SNR [3]. For a hearing impaired person, however, some part of the speech signal will be totally inaudible or heavily distorted due to the hearing loss. Therefore, the perceived signal has less redundancy. As a result, hearing impaired people will have a greater problem in the presence of an interfering noise [4]. Recently, there has been a growing interest to design noise reduction algorithms for hearing aids to reduce listening effort and increase the intelligibility [5–7]. Such an algorithm can be combined with the other digital signal processing techniques that are implemented in current hearing aids.

In a real speech communication system, the target speech signal (from a speaker in the *far end*) can be degraded with both the far-end noise and also the noise from the near-end environment (the listener side). Figure 2

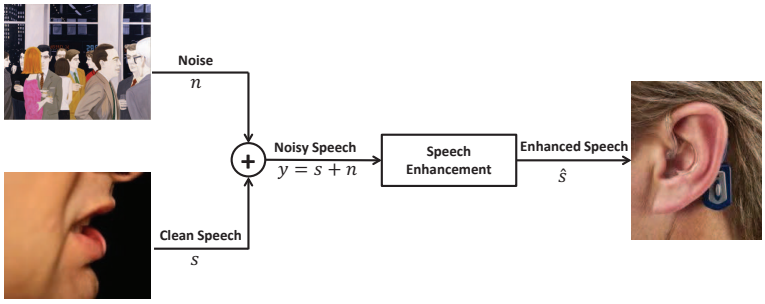

Figure 1: A simplified diagram of a speech enhancement system: the corrupting noise is assumed to be additive.

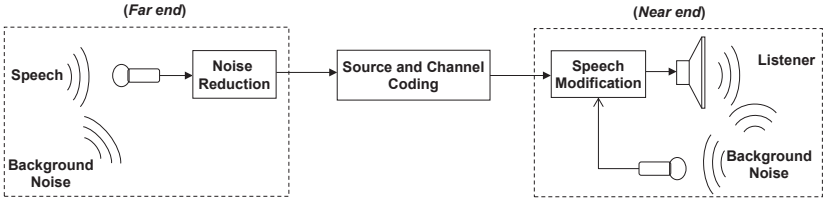

Figure 2: A schematic diagram of a speech communication system. Interfering background noise is present in both the speaker side (*far end*) and the listener side (*near end*).

shows a diagram of such a system. A background noise might be present in both the speaker and the listener sides. The speech enhancement algorithms have been traditionally applied to reduce the far-end noise. This block is named "Noise Reduction" in Figure 2 and is located in the *far end*. In these approaches, the clean speech signal is not known and the goal of the system is to estimate the speech signal by reducing the interfering noise. The enhanced signal will be then transmitted to the listener side. Another class of algorithms have been recently considered to suppress the effect of the near-end background noise. The corresponding block is named "Speech Modification" in Figure 2. In these systems, the speech signal is assumed to be known and the goal is to modify the known speech signal (given some noise statistics) such that the intelligibility of the played-back speech is maximized [8–10]. The design of such systems is a constrained-optimization problem in which the speech energy is usually constrained to be unchanged after the processing.

Estimation of a clean speech signal from a noisy recording is a typical signal estimation task. But due to the non-stationarity of the speech and most of the practical noise signals, and also due to the importance of the

problem, significant amount of research has been devoted to this challenging task. Single-channel speech enhancement algorithms, e.g., [11–18], use the temporal and spectral information of speech and noise to design the estimator. In this case, only the noisy recording obtained from a single microphone is given while the noise type, speaker identity or speaker gender are usually not known.

Multichannel or multimicrophone noise reduction systems, on the other hand, utilize the temporal and spectral information as well as the spatial information to estimate a desired speech signal from the given noisy recordings, see e.g., [19–22]. In this thesis, we focus on single-channel speech enhancement algorithms.

1.2 Single-channel Speech Enhancement

In general, speech enhancement methods can be categorized into two broad classes: unsupervised and supervised. In unsupervised methods such as Wiener and Kalman filters and estimators of the speech DFT coefficients using super-Gaussian priors [12–15, 17], a statistical model is assumed for each of the speech and noise signals, and the clean speech is estimated from the noisy observations without any prior information on the noise type or speaker identity. Hence, no supervision and labeling of signals as speech or a specific noise type is required in these algorithms. For the supervised methods, e.g., [23–28], on the other hand, a model is considered for both the speech and noise signals and the model parameters are learned using the training samples of that signal. Then, an interaction model is defined by combining speech and noise models and the noise reduction task is carried out. In this section, we first discuss the unsupervised noise reduction algorithms. At the end of the section, we will consider the supervised approaches.

1.2.1 Wiener Filter

Wiener filtering is one of the oldest approaches that is used for noise reduction. In the following, we review the Wiener filter in the discrete Fourier transform (DFT) domain, in order to introduce the notation and because it is a baseline for later work. Let us denote the quantized, time domain noisy speech, clean speech, and noise signals by y, s, and n, respectively. Also, denote the sample index by m. For an additive noise, the signal model is written as:

$$y_m = s_m + n_m. \tag{1}$$

To transfer the noisy signal into the frequency domain, data is first segmented into overlapped frames, and then each frame is multiplied by a tapered window (such as Hamming window) to reduce the spectral leakage, and then DFT is applied to the windowed data. The signal is then

processed in the DFT domain and the enhanced signal is reconstructed by using the overlap-add framework [29]. The frame length is usually between 10 and 30 ms, and the speech signal within each frame is assumed to be stationary. Let k and t represent the frequency bin and short-time frame indices, respectively. We denote the vector of the complex DFT coefficients corresponding to frame t of the noisy signal by $\mathbf{y}_t = \{y_{kt}\}$, where y_{kt} is the k-th element of $\mathbf{y}_t$. The vector of the DFT coefficients of the clean speech and noise signals are shown by $\mathbf{s}_t$ and $\mathbf{n}_t$, respectively.

Wiener filtering is a linear minimum mean squared error (LMMSE) estimator that is a special case of the Bayesian Gauss–Markov theorem [30,31]. Using the Wiener filter, the clean speech DFT coefficients are estimated by an element-wise product of the noisy signal $\mathbf{y}_t$ and a weight vector $\mathbf{h}_t$ [1][1]:

$$\hat{\mathbf{s}}_t = \mathbf{h}_t \odot \mathbf{y}_t, \tag{2}$$

where $\odot$ denotes an element-wise product. To obtain the weight vector $\mathbf{h}_t$, the mean square error (MSE) between the clean and estimated speech signals is minimized. Assuming that different frequency bins are independent[2], we can minimize the MSE for each individual frequency bin k separately:

$$h_{kt} = \underset{h_{kt}}{\operatorname{argmin}}\, E\left(|s_{kt} - \hat{s}_{kt}|^2\right), \tag{3}$$

where the expectation is computed with respect to (w.r.t.) the joint distribution $f(s_{kt}, y_{kt})$. Setting the partial derivative w.r.t. the real and imaginary parts of h_{kt} to zero, and assuming that the speech and noise signals are zero-mean and uncorrelated, the optimal weights are obtained as:

$$h_{kt} = \frac{E\left(|s_{kt}|^2\right)}{E\left(|s_{kt}|^2\right) + E\left(|n_{kt}|^2\right)}. \tag{4}$$

To implement Eq. (4), statistics of noise and speech are usually adapted over time to obtain a time-varying gain function. This helps to take into account the non-stationarity of the signals. Eq. (4) is typically implemented

[1] For an optimal linear filter in the time-domain, we assume that the desired estimate is linear in the input. Thus, the parameters of interest are obtained as a convolution of the impulse response of the filter and the observed data. Eq. (2) is then obtained considering the relation of the time-domain convolution and Fourier domain multiplication. In our notations, $\mathbf{h}_t$ denotes the DFT of the filter impulse response.

[2] In a Gaussian model, the independency assumption is equivalent to the assumption that the complex Fourier coefficients are uncorrelated. This has usually been justified by the observation that the correlation between different frequency bins approaches zero as the frame length approaches infinity [13]. In practice, use of the tapered windows will also help to reduce the correlation between widely separated DFT coefficients, at the cost of increasing the correlation between close-by DFT coefficients.

as a function of the *a priori* and *a posteriori* SNRs. For this purpose, the *a priori* SNR (ξ_{kt}) and *a posteriori* SNR (η_{kt}) are defined as:

$$\xi_{kt} = \frac{E\left(|s_{kt}|^2\right)}{E\left(|n_{kt}|^2\right)}, \tag{5}$$

$$\eta_{kt} = \frac{|y_{kt}|^2}{E\left(|n_{kt}|^2\right)}. \tag{6}$$

The optimal weight vector can now be written as:

$$h_{kt} = \frac{\xi_{kt}}{\xi_{kt}+1}. \tag{7}$$

To implement the Wiener filter, we need to have an estimate of the *a priori* SNR ξ_{kt}. One of the commonly used approaches to estimate ξ_{kt} is known as the decision-directed method [13, 32] in which the *a priori* SNR is estimated as:

$$\xi_{kt} = \max\left\{\xi_{\min}, \alpha\frac{|\hat{s}_{k,t-1}|^2}{E\left(|n_{k,t-1}|^2\right)} + (1-\alpha)\max\left\{\eta_{kt}-1, 0\right\}\right\}, \tag{8}$$

where $\xi_{min} \approx 0.003$ is used to lower-limit the amount of noise reduction. Other approaches to estimate ξ_{kt} have also been proposed. For example, in [16], a method is introduced that is based on a generalized autoregressive conditional heteroscedasticity (GARCH) method.

As can be seen in (8), to estimate ξ_{kt} we need to estimate the noise power spectral density (PSD), $E\left(|n_{kt}|^2\right)$. Estimation of the noise PSD is the main difficulty of most of the unsupervised speech enhancement methods, including the Wiener filtering. The simplest approach for this purpose is to use a voice activity detector (VAD)[3]. In this approach, the noise PSD is updated during the speech pauses. These methods can be very sensitive to the performance of the VAD and cannot perform very well at the presence of a non-stationary noise. The alternative methods use the statistical properties of the speech and noise signals to continuously track the noise PSD [34–37]. A recent comparative study was performed in [38] in which the MMSE approach from [36] was found to be the most robust noise estimator among the considered algorithms. A good introduction to the noise estimation algorithms can be found in [2, Chapter 9].

[3]For a review of VAD algorithms see [33, Chapter 10].

1.2.2 Kalman Filter

Although the time-varying Wiener filter (4) is optimal in the sense of mean square error for a given short-time frame, it does not use the prior knowledge about the speech production. For example, the temporal dependencies are not optimally used in the Wiener filtering. Therefore, Kalman filtering and smoothing have been proposed in the literature to improve the performance of the noise reduction algorithms [15, 39–42]. In these methods, the time-domain speech signal is modeled as an autoregressive signal:

$$s_m = \sum_{j=1}^{J} a_j s_{m-j} + w_m, \tag{9}$$

where w_m is a white noise excitation signal, and J is the speech model order which is usually set to 10 in systems with 8 kHz sampling rate [1]. Storing J consecutive samples of s in a vector $\mathbf{s}_m = [s_m, s_{m-1}, \dots s_{m-J+1}]^\top$ with $\top$ denoting the matrix transpose, Eq. (1) and (9) can be written in a state-space formulation as:

$$\mathbf{s}_m = \mathbf{F}\mathbf{s}_{m-1} + \mathbf{G}w_m \tag{10}$$

$$y_m = \mathbf{H}^\top \mathbf{s}_m + n_m. \tag{11}$$

See, e.g., [15] for the definition of the matrices $\mathbf{F}, \mathbf{G}$ and $\mathbf{H}$. Now we can apply Kalman filtering (if we have access to only past data) or smoothing (if we have access to both past and future data) approaches to estimate the clean speech signal [31]. Gannot *et al.* [40] showed that for a white noise and at an input SNR above 0 dB, a fixed-lag variant of the Kalman smoothing can outperform the Wiener filtering by up to 2.5 dB in overall SNR. If we additionally model the noise signal with an autoregressive model, the measurement equation will turn into a noise-free or perfect measurement problem, which is also addressed in the literature, e.g., [41, 43]. In this case, we may introduce a coordinate transformation in order to remove the singularity in the error covariance recursion [44, Chapter 5.10], [41, 45].

1.2.3 Estimators Using Super-Gaussian Priors

The Wiener filter is the optimal linear MMSE filter in which the joint distribution $f(s_{kt}, y_{kt})$ (and hence the distribution of speech $f(s_{kt})$ and distribution of noise $f(n_{kt})$) is not necessarily Gaussian. However, if $f(s_{kt})$ and noise $f(n_{kt})$ are indeed Gaussian, then the Wiener filter will be the optimal MMSE estimator. The assumption of Gaussian distribution for the DFT coefficients was first motivated by the central limit theorem since each DFT coefficient is a weighted sum of many random variables [46,47]. However, for speech signals and the typical frame lengths less than 30 ms, the Gaussian assumption does not agree well with the statistics of data. Figure 3 shows

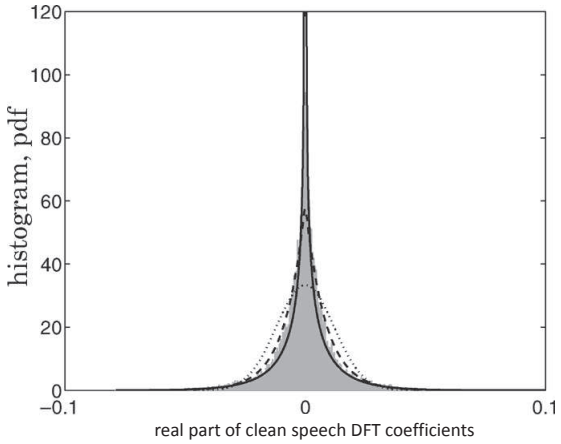

Figure 3: (Dotted) Gaussian, (dashed) Laplace, and (solid) gamma densities fitted to a histogram (shaded) of the real part of clean speech DFT coefficients. Frame length is 32 ms, the sampling rate is 8000 Hz (Source: [14]).

the result of an experiment from [14]. This experiment shows that the real parts of the DFT coefficients of speech have a super-Gaussian (e.g., Laplace or gamma densities that have a sharper peaks and fatter tails) rather than a Gaussian distribution. Experiments performed in [48] also verify this observation (also see [49]). As a result, there has been an increasing interest on obtaining MMSE estimates of the speech DFT coefficients under a given super-Gaussian model [14, 16, 17, 48, 50–53]

In the following, we briefly explain an estimator of the clean speech DFT magnitudes under a one-sided generalized gamma prior density of the form [17]

$$x_{kt} \sim \frac{\gamma \beta^{\nu}}{\Gamma(\nu)} x_{kt}^{\gamma\nu-1} \exp\left(-\beta x_{kt}^{\gamma}\right), \quad \beta > 0, \gamma > 0, \nu > 0, x_{kt} \geq 0, \tag{12}$$

where $x_{kt} = |s_{kt}|$ is the speech DFT magnitude at frequency bin k, and short-time frame t. As discussed in [17], EQ. (12) includes some other distributions as special cases. For example, the Rayleigh distribution (which corresponds to the assumption of Gaussian distribution for the real and imaginary parts of the DFT coefficients) occurs when $\gamma = 2$ and $\nu = 1$. The MMSE estimator is identical to the mean of the posterior distribution of the considered variable, which is given by [30]:

$$\hat{x}_{kt} = E\left(x_{kt} \mid v_{kt}\right) = \frac{\int_0^{\infty} x_{kt} f\left(x_{kt}\right) f\left(v_{kt} \mid x_{kt}\right) dx_{kt}}{\int_0^{\infty} f\left(x_{kt}\right) f\left(v_{kt} \mid x_{kt}\right) dx_{kt}}, \tag{13}$$

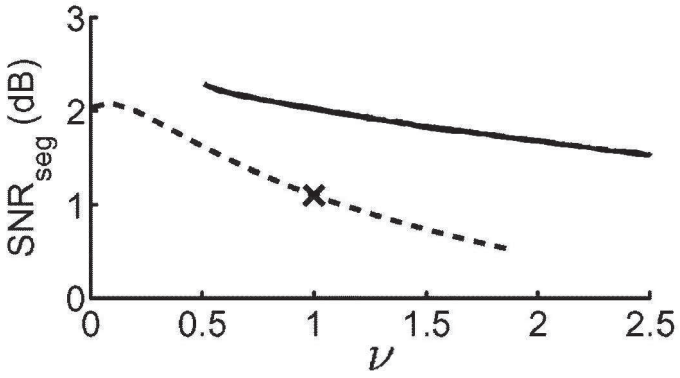

Figure 4: Comparing the performance of Gaussian and super-Gaussian models in terms of segmental SNR as a function of the parameter ν defined in (12). The "cross" corresponds to the Gaussian assumption, the solid and dashed lines correspond to setting $\gamma = 1$ and $\gamma = 2$ in (12), respectively (taken from [17, Fig. 10] and modified for clarity).

where $v_{kt} = |y_{kt}|$ represents a noisy DFT magnitude. Since, in general, this integral cannot be computed in closed form, different approximations are proposed in [17], which usually involve the use of the parabolic cylinder functions [54, Chapter 19].

Figure 4 compares the performance of the Gaussian and super-Gaussian models for street noise and an input SNR of 5 dB. The figure shows the segmental SNR [2, Chapter 10] which is a commonly used objective measure to evaluate a noise reduction algorithm. Other objective measure are also used in [17] that are in line with the results presented in this figure. As it can be seen, a super-Gaussian prior distribution has improved the performance more than 1 dB, compared to the Gaussian prior in this experiment[4]. Later in Section 2, we will use this algorithm with $\gamma = \nu = 1$ to compare the performance of the proposed algorithms.

Many other speech enhancement methods may be classified as unsupervised noise reduction algorithms. Some examples include: iterative Wiener filtering [12], spectral subtraction [11, 55], MMSE log spectral amplitude estimator [56], subspace algorithms [57], and schemes based on periodic models of the speech signal [18].

[4]In general, the performance of a noise reduction algorithm depends on different factors such as the considered prior distributions and the approach used to estimate the *a priori* SNR. In this experiment, the *a priori* SNR is estimated using a decision-directed approach. Different results might be obtained if a different estimator is used for this purpose (see [16] for discussion).

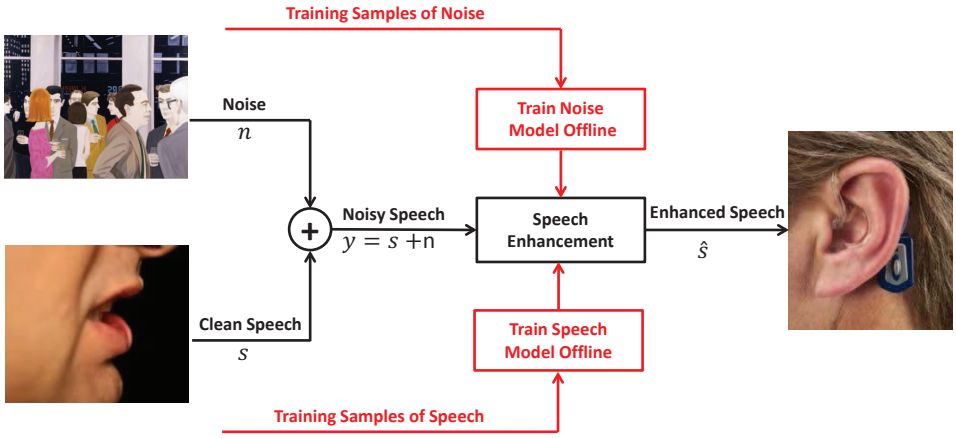

Figure 5: Schematic diagram of a typical supervised speech enhancement system (compare to Figure 1).

As it was mentioned in the beginning of this section, supervised speech enhancement algorithms use some additional information such as noise type or speaker identity to deliver a better enhancement system. In these methods, the speech and noise models are usually trained offline using some training samples (see Figure 5). Some examples of this class of algorithms include the codebook-based approaches [23,58], hidden Markov model (HMM) based systems [24–26,59–61], and methods based on the nonnegative matrix factorization (NMF) [27,62–64]. We will explain the HMM and NMF based denoising methods in greater details in later sections. In this thesis, we propose HMM and NMF based supervised noise reduction schemes. As we will see later, some of the proposed methods can be used in an unsupervised fashion where the algorithm does not require any information that is not available in practice.

One of the main advantages of the supervised methods is that there is no need to estimate the noise PSD using a separate algorithm. Therefore, the algorithms can perform well even at the presence of a non-stationary noise, given that we know the noise type and we train a model for that. The supervised approaches have been shown to produce better quality enhanced speech signals compared to the unsupervised methods [23, 25], which can be expected as more information is fed to the system in these cases and the considered models are trained for each specific type of signals. The required prior information on noise type (and speaker identity in some cases) can be given by the user, or can be obtained using a built-in classification scheme [23, 25, 27], or can be provided by a separate acoustic environment classification algorithm [65–67].

1.3 Hidden Markov Model

Hidden Markov models (HMM) are one of the simple and yet often used dynamical models to describe a correlated sequence of data [68]. An HMM can be seen as a generalization of a mixture model in which the hidden variables, corresponding to the mixture weights, are related through a Markov process. HMM is characterized by a set of hidden states and a set of state-dependent output probability distributions. Let us denote the (multidimensional) data at time t by $\mathbf{s}_t$, and represent the scalar hidden variable by z_t. An HMM consists of a discrete Markov chain and a set of state-conditional probability distributions shown by $P\left(s_t \mid z_t = j\right), j \in \{1 \dots J\}$ where J is the number of states in the HMM. The Markov chain itself is characterized by an initial probability vector over the hidden states, denoted by $\mathbf{q}$ with $q_j = P\left(z_1 = j\right)$ and a transition matrix between the states, denoted by $\mathbf{A}$ with elements $a_{ij} = P\left(z_t = j \mid z_{t-1} = i\right)$.

The model parameters in HMM are usually estimated by maximizing the marginalized likelihood. Due to the presence of hidden states, the maximum likelihood (ML) estimate of the HMM parameters are obtained using the expectation maximization (EM) algorithm [68–70]. In fact, the Baum-Welch training algorithm was proposed years earlier than the EM algorithm [71], and in [69], it was observed that Baum-Welch approach is an example of the EM algorithm.

Let us denote the HMM parameters by $\lambda = \{\mathbf{q}, \mathbf{A}, \boldsymbol{\theta}\}$ where $\boldsymbol{\theta}$ represents all the parameters of the output distributions. Assume that we want to estimate the HMM parameters given a sequence of the observed data $\mathbf{s}_1^T = \{\mathbf{s}_1, \dots \mathbf{s}_T\}$. Denote the corresponding sequence of the hidden variables by $z_1^T = \{z_1, \dots z_T\}$, i.e., z_j shows which state is used to generate $\mathbf{s}_j$. The main assumption in using EM is that the maximization of $f\left(\mathbf{s}_1^T, z_1^T; \lambda\right)$ is much easier than directly maximizing $f\left(\mathbf{s}_1^T; \lambda\right)$. In the E step of the EM algorithm, a lower bound is obtained on $\log\left(f\left(\mathbf{s}_1^T; \lambda\right)\right)$, and in the M step, this lower bound is maximized [72, Chapter 9]. The EM lower bound takes the form

$$\begin{aligned} \mathcal{L}\left(f\left(z_1^T \mid \mathbf{s}_1^T; \lambda\right), \hat{\lambda}\right) &= Q\left(\hat{\lambda}, \lambda\right) + \text{const.}, \quad \text{where} \\ Q\left(\hat{\lambda}, \lambda\right) &= \sum_{z_1, \dots z_T} f\left(z_1^T \mid \mathbf{s}_1^T; \lambda\right) \log\left(f\left(\mathbf{s}_1^T, z_1^T; \hat{\lambda}\right)\right), \end{aligned} \quad (14)$$

where λ includes the estimated parameters from the previous iteration of the EM, and $\hat{\lambda}$ contains the new estimates to be obtained. In words, the E step of EM (or computing $Q\left(\hat{\lambda}, \lambda\right)$) is equivalent to computing the expected value of the log-likelihood of the complete data (i.e., both $\mathbf{s}_1^T$ and z_1^T) w.r.t. the posterior distribution of the hidden variables z_1^T. In the M step, the derivative of (14) is computed and set to zero to obtain $\hat{\lambda}$. The E and M

steps are iteratively performed until a stationary point of the log-likelihood is achieved. It can be proved that the EM algorithm always converges and a locally optimal solution can be obtained [72].

The presented HMM can be seen as the discrete counterpart of the Kalman filter where the state-space is discretized [73]. From the application perspective, Kalman filters have been usually used to characterize the time-evolution of a source (tracking) while HMMs are used for classification purposes, e.g., [68, 74, 75]. HMMs with a continuous state-space or infinite number of states are also addressed in literature [73, 76–78]. However, an exact implementation of the EM algorithm for these methods is generally not possible, except for some very few specific cases, e.g., Gaussian linear state-space models, and simulation-based methods have to be used instead [76].

1.3.1 HMM-based Speech Enhancement

HMM-based speech enhancement was first addressed in [24,59,79]. For this purpose, an additive noise model was considered as in (1):

$$y_m = s_m + n_m. \tag{15}$$

L consecutive samples (one frame) of the noisy signal are stored in the vector $\mathbf{y}_t = [y_m, y_{m-1}, \dots y_{m-L+1}]$ with t denoting the frame index. The vectors $\mathbf{s}_t$ and $\mathbf{n}_t$ are similarly defined for the clean speech and noise signals, respectively. The speech and noise time-domain signals are modeled with a first order HMM where each output density function is given by a zero-mean Gaussian mixture model (GMM) [24, 59, 79]. Furthermore, it is assumed that, given a hidden state, the speech and noise processes are autoregressive (similar to (9)). In this model, the probability of a sequence of clean speech vectors $\mathbf{s}_1^T = \{\mathbf{s}_1, \dots \mathbf{s}_T\}$ is given by

$$f\left(\mathbf{s}_1^T; \lambda\right) = \sum_{z_1} \cdots \sum_{z_T} \prod_{t=1}^{T} a_{z_{t-1} z_t} f\left(\mathbf{s}_t \mid z_t; \boldsymbol{\theta}\right), \tag{16}$$

where $a_{z_0 z_1} \triangleq P(z_1)$ is the probability of the initial state z_1 and $f(\mathbf{s}_t \mid z_t; \boldsymbol{\theta})$ is given by a state-dependent GMM:

$$f\left(\mathbf{s}_t \mid z_t; \boldsymbol{\theta}\right) = \sum_{i=1}^{I} w_{i,z_t} \mathcal{G}\left(\mathbf{s}_t; 0, \mathbf{C}_{i,z_t}\right), \tag{17}$$

where w_{i,z_t} is the mixture weight for the i-th component of state z_t, and $\mathbf{C}_{i,z_t}$ is the covariance matrix of the i-th component of state z_t. The model parameters can be estimated using the EM algorithm.

To model the noisy signal, the speech and noise HMMs are combined to obtain a bigger HMM (later known as a factorial HMM [80]) in which

the Markov chain of each source evolves over time independently. Both the maximum a posterior (MAP) and the MMSE estimates have been investigated for HMM-based speech enhancement [24, 59, 79]. An iterative MAP and an iterative approximate MAP (AMAP) approaches were proposed in [79]. In these approaches, the noise HMM consists of a single state and a single Gaussian component. For the MAP approach, given the estimate of the clean speech signal in the current iteration, the probability of being at a specific state is found using the forward-backward algorithm [68]. These weights are then used to update the speech estimate using a sum of weighted Wiener filters. The enhanced speech signal is used to start the next iteration. This approach was further developed in [24] by adding a speech gain adaption scheme. The gain adaption plays an important role to make the algorithm practical since for different levels of the signal (for instance at a different loudness level), the covariance matrix of the Gaussian components will change [81, Section 6].

For the AMAP approach, which is a simplified approximation of MAP, a single state and mixture pair from speech HMM is assumed to dominantly explain the estimated speech signal at the current iteration, at each time frame t. As a result, the clean speech signal is estimated using a single Wiener filter that corresponds to the dominant state and mixture pair, and it is used in the next iteration. This approach is based on the most likely sequence of states and mixture components obtained by applying the Viterbi algorithm [79].

The MMSE estimators for HMM-based enhancement systems are addressed in [24, 25, 60]. It can be shown that the optimal MMSE estimator is the sum of the weighted state-dependent MMSE estimators where the weights are given by the posterior probability of the states. An important issue of the supervised approaches is addressed in [25] in which the noise type is not known a priori and is selected based on the noisy observations. For this purpose, different noise models are trained offline, and then during intervals of speech pauses (longer than 100 ms), a Viterbi algorithm is performed using different noise models. The noise HMM generating the best score is selected and a gain adjustment is carried out to adapt to the noise level using another Viterbi algorithm. This can be seen as a heuristic noise gain adaptation using VAD.

In the evaluations using a multitalker babble noise in [25], the HMM-based MMSE estimator outperformed a spectral subtraction algorithm by at least 2.5 dB in overall SNR for all the input SNRs above 0 dB. It was also observed that at input SNRs above 15 dB, the implemented spectral subtraction method actually deteriorates the output signal (where output SNR is lower than the input SNR) while the HMM-based system keeps improving the SNR.

As mentioned earlier, gain modeling is an important issue in the HMM-based systems. While HMMs can model the spectral shape of different

speech sounds, they usually do not model the variations of the speech energy levels within a state. Also, they do not adapt to different long-term noise levels, which can happen, e.g., due to movement of a noise source or a change of SNR. Zhao *et al.* [60, 82] proposed an approach in which log-normal prior distributions are considered over the speech and noise gains to explicitly model these level changes. The time-invariant model parameters are learned using an EM algorithm offline. The time-variant parameters, the mean value of the gain distributions denoted by μ_s and μ_n, are updated online (given only the noisy signal) using a recursive EM algorithm [83–85]. The recursive EM algorithm is a stochastic approximation in which the parameters of interest are updated sequentially. To do so, the EM help function is defined as the conditional expectation of the log-likelihood of the complete data until the current time w.r.t. the posterior distribution of the hidden variables. Then, this help function is maximized over the parameters by a single-iteration stochastic approximation in each time instance. Based on the online estimation of the HMM parameters [85], an online gain adaption is proposed in [60] in which μ_s and μ_n are updated in a recursive manner after the estimation of the clean speech signal is done for the current frame t. Therefore, given the noisy signal, a correction term is calculated and is added to the current estimates to obtain a new estimate of the parameters to be used in the next frame.

In the algorithms that we have discussed so far [24, 25, 59, 60, 79], the speech and noise signals are assumed to be independent and distributed according to a GMM. Therefore, the state-conditional distribution of the noisy signal is a GMM. Here, each Gaussian component of a given state has a mean equal to zero and a covariance matrix equal to the sum of the covariance matrices of the clean speech and noise signals at the given mixtures and states (see e.g. [24, Eq. (5) and the following paragraph]). Hence, the forward-backward and Viterbi algorithms can be carried out easily. In general, obtaining the conditional distribution of the noisy observation might be very difficult. Also, if there are many states in the HMMs (which is usual in HMM-based speech source separation), the exact implementation of the forward-backward algorithm might not be feasible. In these situations, different approximations may be used to simplify the calculations [86, 87].

For example, in an early effort to use HMM-based noisy speech decomposition in [86], the log energy levels of a 27-channel filter bank was used as the observation and was modeled by the multivariate Gaussian distributions. Because of the filter bank and the log operator, the distribution of the noisy speech is difficult to obtain and hence an approximation is required. For this purpose, it is assumed that in each channel, the observation can be approximated by the maximum of the log energies of the clean speech and the noise signals. This is known as the mixture-maximization (MIXMAX) approximation, and Radfar *et al.* [88] have proved that the MIXMAX approximation is a nonlinear MMSE estimator. For a similar

observation setup and using a very large state space for the HMMs, another approximation approach is proposed in [87] to facilitate obtaining the most probable state in each time frame. Other approximation methods have been discussed in [89–91].

In [92] speech recognition using Mel-frequency cepstral coefficients (MFCC) is studied where a factorial HMM [80] is used to model noisy features. Assuming that the MFCC features have Gaussian distribution and using the properties of the MFCCs, a Gaussian distribution is obtained to model the noisy MFCC features. The noise and speech signals are assumed to have different levels and a greedy algorithm is proposed to obtain the best state sequence and the speech and noise gains. Hence, given the gains, a 2D Viterbi [86] algorithm is applied to find the best composite state, which is then used to update the speech and noise gains using a greedy optimization algorithm. The use of MFCCs for speech enhancement is further developed in [61] in which a parallel cepstral and spectral modeling is proposed. The work in [61] is motivated by the observation that the estimation of the filter weights (to weight state-dependent filters), i.e., the filter selection, is actually a pattern recognition problem in which a higher recognition rate results in a better speech enhancement algorithm. Accordingly, the proposed noise reduction system uses MFCCs to obtain the filter weights while the state-dependent filters are constructed in a high resolution spectral domain.

1.4 Nonnegative Matrix Factorization

Nonnegative matrix factorization (NMF) is a technique to project a nonnegative matrix $\mathbf{V} = \{v_{kt}\} \in \mathbb{R}_+^{K\times T}$ onto a space spanned by a linear combination of a set of basis vectors, i.e., $\mathbf{V} \approx \mathbf{WH}$, where $\mathbf{W} \in \mathbb{R}_+^{K\times I}$ and $\mathbf{H} \in \mathbb{R}_+^{I\times T}$ [93]. Here, $\mathbb{R}_+$ is used to denote the nonnegative real vector space. In a usual setup, $K > I$, and hence, $\mathbf{H}$ provides a low-dimensional representation of data in terms of a set of basis vectors. Assume that the complex DFT coefficients of a signal is given by $\mathbf{Y} = \{y_{kt}\}$, where k and t are the frequency bin and time indices. The input to NMF, $\mathbf{V}$, is a nonnegative transformation of $\mathbf{Y}$. One of the popular choices is $v_{kt} = |y_{kt}|$, i.e., the input to NMF is the magnitude spectrogram of the speech signal with spectral vectors stored by column. In this notation, $\mathbf{W}$ is the basis matrix or dictionary, and $\mathbf{H}$ is referred to as the NMF coefficient or the activation matrix.

To obtain a nonnegative decomposition of a given matrix, a cost function is usually defined and minimized:

$$\begin{aligned} (\mathbf{W},\mathbf{H}) &= \underset{\mathbf{W},\mathbf{H}}{\operatorname{argmin}}\ D(\mathbf{V}\|\mathbf{WH}) \\ \text{s. t.} &\quad w_{ki} \geq 0, h_{it} \geq 0,\ \forall k,i,t \end{aligned} \tag{18}$$

where $D(\mathbf{V}\|\hat{\mathbf{V}})$ is a cost function [94,95]. The NMF problem in not convex

in general, and it is usually solved by iteratively optimizing (18) w.r.t. $\mathbf{W}$ and $\mathbf{H}$. One of the simple algorithms that has been used frequently is the one with the multiplicative update rule. For a Euclidean cost function ($D(\mathbf{V}\|\hat{\mathbf{V}}) = \sum_{k,t}(v_{kt} - \hat{v}_{kt})^2$), these updates are given by [95]:

$$w_{ki} \leftarrow w_{ki}\frac{\left[\mathbf{V}\mathbf{H}^\top\right]_{ki}}{\left[\mathbf{W}\mathbf{H}\mathbf{H}^\top\right]_{ki}},\ \forall k,i, \tag{19}$$

$$h_{it} \leftarrow h_{it}\frac{\left[\mathbf{W}^\top\mathbf{V}\right]_{it}}{\left[\mathbf{W}^\top\mathbf{W}\mathbf{H}\right]_{it}},\ \forall i,t. \tag{20}$$

Starting from nonnegative initializations for the factors, (19) and (20) lead to a locally optimal solution for (18). These update rules can be motivated by investigating the Karush-Kuhn-Tucker conditions [96,97]. Another derivation for this algorithm can be given using the split gradient methods (SGM) [97]. In the SGM approach, gradient of the error function is assumed to have a decomposition as $\nabla\mathcal{E} = [\nabla\mathcal{E}]^+ - [\nabla\mathcal{E}]^-$ with $[\nabla\mathcal{E}]^+ > 0$ and $[\nabla\mathcal{E}]^- > 0$, and the update rule is given by

$$\theta \leftarrow \theta\frac{[\nabla\mathcal{E}]^-}{[\nabla\mathcal{E}]^+}. \tag{21}$$

The multiplicative update rules arise as a special case of the gradient-descent algorithms [98]. More efficient projected gradient approaches have been also used to obtain NMF representations [99, 100], which may also improve the performance in a specific application [100].

For most of the practical applications such as blind source separation (BSS) and speech enhancement, the performance might be improved by imposing constraints, e.g., sparsity and temporal dependencies. In these scenarios, a regularized cost function is minimized to obtain the NMF representation:

$$\begin{aligned}(\mathbf{W},\mathbf{H}) &= \underset{\mathbf{W},\mathbf{H}}{\operatorname{argmin}}\ D(\mathbf{V}\|\mathbf{W}\mathbf{H}) + \mu g(\mathbf{W},\mathbf{H}), \\ \text{s. t.}\quad & w_{ki} \geq 0, h_{it} \geq 0,\ \forall k,i,t\end{aligned} \tag{22}$$

where $g(\cdot)$ is the regularization term, and μ is the regularization weight [100–103]. A proper choice of μ gives a good trade-off between the fidelity and satisfying the imposed regularization.

1.4.1 Probabilistic NMF

For stochastic signals like speech, it is beneficial to formulate the NMF decomposition in a probabilistic framework. In these approaches, the EM algorithm is usually used to maximize the log-likelihood of data and to obtain an NMF representation, e.g., [104–106]. The discussed Euclidean

NMF (EUC-NMF) can be seen as a probabilistic NMF in which each observation v_{kt} is derived from a Gaussian distribution with a mean value $\hat{v}_{kt} = \sum_i w_{ki} h_{it}$ and a constant variance, see [105, 107].

Another frequently used NMF is based on minimizing the Kullback-Leibler divergence [95]. The KL-NMF can be also seen as a probabilistic NMF in which v_{kt} is assumed to be drawn from a Poisson distribution with a parameter given by $\hat{v}_{kt}$. As a result, the observed data has to be scaled to be integer. It has been shown that this scaling is usually practical [27, 106], however, it might imply theoretical problems since the scaling level directly affects the assumed noise level in the model [108]. Using this Poisson model, Cemgil [106] has proposed an EM algorithm in which the update rules are identical to the multiplicative update rules for the NMF with the KL divergence [95].

Févotte *et al.* [107] have proposed a probabilistic NMF that minimizes the Itakura-Saito divergence (IS-NMF). IS divergence exhibits a scale-invariant property (i.e, $D\left(v_{kt} \| \hat{v}_{kt}\right) = D\left(\gamma v_{kt} \| \gamma \hat{v}_{kt}\right)$. This means that a bad approximation for low-power coefficients has a similar effect in the cost function as a bad approximation for higher power coefficients, i.e., the relative errors are important rather than the absolute error values. This is relevant to speech signals in which the higher frequency bins have low power but are very important to perceive the sound [109]. Authors in [107] propose a statistical model for the IS-NMF in which the complex variables y_{kt} are assumed to be sum of complex Gaussian components (with parameters specified with NMF factors). Another statistical model is also proposed in [107] that gives rise to the gamma multiplicative noise. The ML estimate of the parameters in both of these models is shown to be equivalent to performing an IS-NMF on $\mathbf{V}$ with $v_{kt} = |y_{kt}|^2$ [107]. Other probabilistic NMF approaches have been also developed in the literature that correspond to different statistical models, e.g., [110, 111].

In the following, we describe one probabilistic NMF that is called probabilistic latent component analysis (PLCA) [112]. Since this approach has been used in some of the proposed methods, we provide some more details about that. PLCA is a derivation of the probabilistic latent semantic indexing (PLSI) [113, 114], which has mainly been applied to document indexing. In document models such as PLSI or latent Dirichlet allocation (LDA) [115–117], the term-frequency representation is usually used to represent a text corpus as count data [118]. Hence, each element v_{kt} is the number of repetitions of word k in document t. In PLSI, the distribution of words within a document is approximated by a convex combination of some weighted marginal distributions. Each marginal distribution corresponds to a "topic" and shows how frequently the words are used within this topic. The popularity of a topic within a document is reflected in its corresponding weight. To generate a word for a document, first a topic is chosen from the document-specific topic distribution. Then, a word is chosen according to

the topic-dependent word distribution. This procedure is repeated continuously to produce a complete document. In a speech processing application, a word is replaced by a frequency bin, and a document is replaced by a short-time spectrum.

In PLCA, the distribution of an input vector is assumed to be a mixture of some marginal distributions. A latent variable is defined to refer to the index of the underlying mixture component, which has generated an observation, and the probabilities of different outcomes of this latent variable determine the weights in the mixture. In this model, each vector of the observation matrix, $\mathbf{v}_t = |\mathbf{y}_t|$[5], is assumed to be distributed according to a multinomial distribution [119] with a parameter vector denoted by $\boldsymbol{\theta}_t$, and an expected value given by $E(\mathbf{v}_t) = \gamma_t \boldsymbol{\theta}_t$. Here, $\gamma_t = \sum_k v_{kt}$ is the total number of draws from the distribution at time t. The k-th element of $\boldsymbol{\theta}_t$ (θ_{kt}) indicates the probability that the k-th row of $\mathbf{v}_t$ will be chosen in a particular draw from the multinomial distribution.

Let us define the scalar random variable Φ_t that can take one of the K possible frequency indices $k = 1, \dots K$ as its outcome. The k-th element of $\boldsymbol{\theta}_t$ is now given by: $\theta_{kt} = P(\Phi_t = k)$. Also, let $\mathbb{H}_t$ denote a scalar random latent variable that can take one of the I possible discrete values $i = 1, \dots I$. Using the conditional probabilities, $P(\Phi_t = k)$ is given by

$$\theta_{kt} = P\left(\Phi_t = k\right) = \sum_{i=1}^{I} P\left(\Phi_t = k \mid \mathbb{H}_t = i\right) P\left(\mathbb{H}_t = i\right). \tag{23}$$

Using the terminology of document models, each outcome of $\mathbb{H}_t$ corresponds to a specific topic. We define a coefficient matrix $\mathbf{H}$ with elements $h_{it} = P(\mathbb{H}_t = i)$, and a basis matrix $\mathbf{W}$ with elements $w_{ki} = P(\Phi_t = k \mid \mathbb{H}_t = i)$. In principle, $\mathbf{W}$ is time-invariant and includes the possible spectral structures of the speech signals. Eq. (23) is now equivalently written as: $\boldsymbol{\theta}_t = \mathbf{W}\mathbf{h}_t$. An observed magnitude spectrum $\mathbf{v}_t$ can be approximated by the expected value of the underlying multinomial distribution as $\mathbf{v}_t \approx \gamma_t \boldsymbol{\theta}_t = \gamma_t \left(\mathbf{W}\mathbf{h}_t\right)$. The basis and coefficient matrices ($\mathbf{W}$ and $\mathbf{H}$) can be estimated using the EM algorithm [119]. The iterative update rules are given by:

$$h_{it} \leftarrow \frac{h_{it} \sum_k w_{ki} \left(v_{kt} / \hat{v}_{kt}\right)}{\sum_i h_{it} \sum_k w_{ki} \left(v_{kt} / \hat{v}_{kt}\right)}\ , \tag{24}$$

$$w_{ki} \leftarrow \frac{w_{ki} \sum_t h_{it} \left(v_{kt} / \hat{v}_{kt}\right)}{\sum_k w_{ki} \sum_t h_{it} \left(v_{kt} / \hat{v}_{kt}\right)}\ , \tag{25}$$

where $\hat{\mathbf{v}}_t = \gamma_t \mathbf{W}\mathbf{h}_t$ is the model approximation that is updated after each iteration. It can be shown that the PLCA minimizes a weighted KL divergence as $D_{\text{PLCA}} = \sum_t \gamma_t D_{\text{KL}}\left(\boldsymbol{\lambda}_t \| \hat{\boldsymbol{\lambda}}_t\right)$ where $\boldsymbol{\lambda}_t = \mathbf{v}_t / \gamma_t$, $\hat{\boldsymbol{\lambda}}_t = \mathbf{W}\mathbf{h}_t$, and

[5] All the operations are element-wise, unless otherwise mentioned.

$D_{\mathrm{KL}}\left(\boldsymbol{\lambda}_t \| \hat{\boldsymbol{\lambda}}_t\right) = \sum_k \lambda_{kt} \log \frac{\lambda_{kt}}{\hat{\lambda}_{kt}}$ corresponds to the KL divergence between the normalized data and its approximation at time t [119, supplementary document]. Various other versions of PLCA, e.g., sparse overcomplete, have been proposed in the literature [119–121].

1.4.2 Source Separation and Speech Enhancement Using NMF

NMF has been widely used as a source separation technique applied to monaural mixtures, e.g., [93, 101, 107, 122–129]. More recently, NMF has also been used to estimate the clean speech from a noisy observation [27, 62–64, 130–135]. As before, we denote the matrix of complex DFT coefficients of noisy speech, clean speech, and noise signals by $\mathbf{Y}$, $\mathbf{S}$, and $\mathbf{N}$, respectively. To apply NMF, we first obtain a nonnegative transformation of these matrices, which are denoted by $\mathbf{V}$, $\mathbf{X}$, and $\mathbf{U}$, such that $v_{kt} = |y_{kt}|^p$, $x_{kt} = |s_{kt}|^p$, and $u_{kt} = |n_{kt}|^p$ where $p = 1$ for magnitude spectrogram and $p = 2$ for power spectrogram.

Let us consider a supervised denoising approach where the basis matrix of speech $\mathbf{W}^{(s)}$ and the basis matrix of noise $\mathbf{W}^{(n)}$ are learned using some appropriate training data ($\mathbf{X}_{\mathrm{tr}}$ and $\mathbf{U}_{\mathrm{tr}}$) prior to the enhancement. The commonly used assumption to model the noisy speech signal is the additivity of speech and noise spectrograms, i.e., $\mathbf{v_t} = \mathbf{x}_t + \mathbf{u}_t$. Although in real world problems this assumption is not justified completely, the developed algorithms have shown to produce satisfactory results, e.g., [122]. The basis matrix of the noisy signal is obtained by concatenating the speech and noise basis matrices as $\mathbf{W} = \left[\mathbf{W}^{(s)}\ \mathbf{W}^{(n)}\right]$ (see Figure 6). Given $\mathbf{v}_t$, the NMF problem (18) is now solved (with fixed $\mathbf{W}$) to obtain the noisy NMF coefficients $\mathbf{h}_t$, i.e., $\mathbf{v}_t \approx \mathbf{W}\mathbf{h}_t = \left[\mathbf{W}^{(s)}\ \mathbf{W}^{(n)}\right]\left[\mathbf{h}_t^{(s)\top}\ \mathbf{h}_t^{(n)\top}\right]^\top$. Finally, an estimate of the clean speech spectrum is obtained by a Wiener-type filtering as:

$$\hat{\mathbf{x}}_t = \frac{\mathbf{W}^{(s)}\mathbf{h}_t^{(s)}}{\mathbf{W}^{(s)}\mathbf{h}_t^{(s)} + \mathbf{W}^{(n)}\mathbf{h}_t^{(n)}} \odot \mathbf{v}_t, \tag{26}$$

where the division is performed element-wise, and $\odot$ denotes an element-wise multiplication. The clean waveform is estimated by using $|\hat{\mathbf{x}}_t|^{1/p}$ and the noisy phase, and by applying the inverse DFT. One advantage of the NMF-based approaches over the HMM-based [25, 60] or codebook-driven [23] methods is that NMF automatically captures the long-term levels of the signals, and no additional gain modeling is necessary.

When NMF algorithms are used for speech source separation, a good separation can be expected only when speaker-dependent basis matrices are learned. In contrast, for noise reduction, even if a general speaker-independent basis matrix of speech is learned, a good enhancement can be achieved [133, 135]. Since the basic NMF allows a large degree of free-

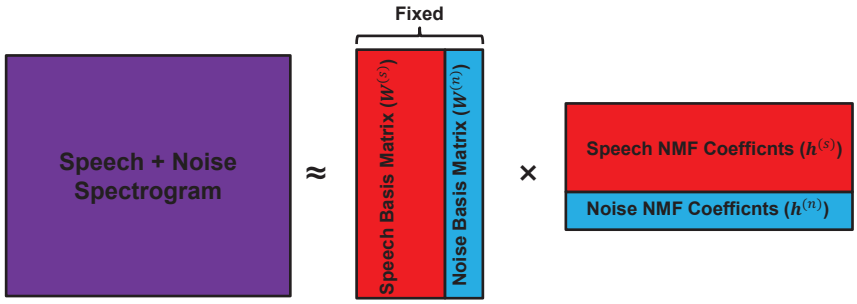

Figure 6: Applying NMF on noisy speech.

dom, the performance of the source separation algorithms can be improved by imposing extra constraints and regularizations, motivated by the sparsity of the basis vectors and NMF coefficients or smoothness of the NMF coefficients. In probabilistic NMFs, these constraints can be applied in the form of prior distributions. Among different priors, a significant attention has been paid to model the temporal dependencies in the signals because this important aspect of audio signals is ignored in a basic NMF approach [27, 63, 64, 122, 136–141]. This issue will be discussed in more details in Section 2.

Schmidt *et al.* [130] presented an NMF-based unsupervised batch algorithm for noise reduction. In this approach, it is assumed that the entire noisy signal is observed, then the noise basis vectors are learned during the speech pauses. In the intervals of speech activity, the noise basis matrix is kept fixed and the rest of the parameters (including speech basis and speech and noise NMF coefficients) are learned by minimizing the Euclidean distance with an additional regularization term to impose sparsity on the NMF coefficients. The enhanced signal is then obtained similarly to (26). The reported results show that this method outperforms a spectral subtraction algorithm, especially for highly non-stationary noises. However, the NMF approach is sensitive to the performance of the voice activity detector (VAD). Moreover, the proposed algorithm in [130] is applicable only in the batch mode, which is not practical in many real-world problems.

In [62], a supervised NMF-based denoising scheme is proposed in which a heuristic regularization term is added to the cost function. By doing so, the factorization is enforced to follow the pre-obtained statistics. In this method, the basis matrices of speech and noise are learned from training data offline. Also, as part of the training, the mean and covariance of the log of the NMF coefficients are computed. Using these statistics, the negative likelihood of a Gaussian distribution (with the calculated mean and covariance) is used to regularize the cost function during the enhancement.

The clean speech signal is then estimated as $\hat{\mathbf{x}}_t = \mathbf{W}^{(s)}\mathbf{h}_t^{(s)}$. Although it is not explicitly mentioned in [62], to make regularization meaningful, the statistics of the speech and noise NMF coefficients have to be adjusted according to the long-term levels of speech and noise signals.

The above NMF-based enhancement system was evaluated and compared to the ETSI two-stage Wiener filter [142] in [62]. For the NMF approach, two alternatives were tried in which the speech basis matrix was either speaker-dependent (NMF-self) or gender-dependent (NMF-group). The simulation was done for different noises and at an input SNR of 0 dB. Considering the bus/street noise and male speakers, evaluations showed that the NMF-self approach leads to 0.45 MOS higher Perceptual Evaluation of Speech Quality (PESQ) [143] and around 1.8 dB higher segmental SNR compared to the Wiener filter. The NMF-group was also found to outperform the Wiener filter by more than 0.2 MOS in PESQ while the improvement in segmental SNR was negligible.

A semi-supervised approach is proposed in [131] to denoise a noisy signal using NMF. In this method, a nonnegative hidden Markov model (NHMM) is used to model speech magnitude spectrogram. Here, the output density function of each state is assumed to be a mixture of multinomial distributions, and thus, the model is closely related to probabilistic latent component analysis (PLCA) [112]. An NHMM is described by a set of basis matrices and a Markovian transition matrix that captures the temporal dynamics of the underlying data. To describe a mixture signal, the corresponding NHMMs are used to construct a factorial HMM. When applied for noise reduction, a speaker-dependent NHMM is trained on a speech signal. Then, assuming that the whole noisy signal is available (batch mode), the EM algorithm is run to simultaneously estimate a single-state NHMM for noise and to estimate the NMF coefficients of the speech and noise signals. The proposed algorithm does not use a VAD to update the noise dictionary, as was done in [130], but the algorithm requires the entire spectrogram of the noisy signal, which makes it difficult for practical applications. Moreover, the employed speech model is speaker-dependent, and requires a separate speaker identification algorithm in practice. Finally, similar to the other approaches based on the factorial models, the method in [131] suffers from high computational complexity.

Raj *et al.* [144] proposed a phoneme-dependent approach to use NMF for speech enhancement in which a set of basis vectors is learned for each phoneme a priori. Given the noisy recording, an iterative NMF-based speech enhancer combined with an automatic speech recognizer (ASR) is pursued to estimate the clean speech signal. In the experiments, a mixture of speech and music is considered and the estimation of the clean speech is carried out using a set of speaker-dependent basis matrices.

The approaches mentioned here do not model the temporal dependencies

in an optimal way or the speech estimation in not optimal in a statistical sense. Additionally, none of these methods address the problem where the underlying sources have similar basis matrices. Moreover, some of these algorithms are only applicable in a batch mode, and hence, cannot be applied for online speech enhancement. This dissertation proposes solutions and improvements regarding these problems.

2 Methods and Results

This section summarizes the main contributions of this thesis. The summary includes only the papers that are included in Part II of this dissertation. We have proposed and evaluated single-channel NMF and HMM-based speech enhancement systems. The proposed methods can be divided into three categories in general:

1. Speech Enhancement Using Dynamic NMF
2. NMF-based Separation of Sources with Similar Dictionaries
3. Super-Gaussian Priors in HMM-based Enhancement Systems

Two important shortcomings of the standard NMF approaches have been addressed in our NMF-based speech enhancement algorithms:

1) We have developed NMF-based enhancement approaches that use temporal dynamics. As mentioned earlier, the correlation between consecutive time-frames is not directly used in a standard NMF. However, the time dependencies are an important aspect of the audio signals. As we will show, using this information in an NMF-based denoising system can improve the performance significantly. We will discuss both continuous and discrete dynamical systems in Section 2.1. Using these systems, we have derived optimal estimators for the clean speech signal. Additionally, we present an approach to learn the noise basis matrix online from the noisy observations. Section 2.1 is mainly based on the following papers:

Paper A 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 Applications of Signal Process. Audio Acoustics (WASPAA)*, oct. 2011, pp. 45-48.

Paper B 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.

Paper C N. Mohammadiha and A. Leijon, "Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement,"

IEEE Trans. Audio, Speech, and Language Process., vol. 21, no. 5, pp. 998–1011, may 2013.

Paper E N. Mohammadiha, P. Smaragdis, and A. Leijon, "Prediction based filtering and smoothing to exploit temporal dependencies in NMF," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, may 2013, pp. 873–877.

2) We present our methods for NMF-based separation of sources with similar dictionaries in Section 2.2. For some applications, such as denoising a babble-contaminated speech signal or separation of sources with similar-gender speakers, the basis matrices of the underlying sources might be quite similar or at least may have some common set of basis vectors. As a result, the performance of the NMF-based algorithms is usually worse in these cases. Section 2.2 briefly explains our solutions which are mainly based on the following papers:

Paper C N. Mohammadiha and A. Leijon, "Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 998–1011, may 2013.

Paper F N. Mohammadiha, P. Smaragdis, and A. Leijon, "Low-artifact Source Separation Using Probabilistic Latent Component Analysis," in *Proc. IEEE Workshop Applications of Signal Process. Audio Acoustics (WASPAA)*, oct. 2013.

Finally, in Section 2.3, we present our experiments with the periodogram coefficients of speech signals conditioned on a given phone and show that even the phoneme-conditioned speech DFT coefficients are rather super-Gaussian distributed. We also review our HMM-based spectral enhancement approach with super-Gaussian priors. This section is based on the following paper:

Paper D N. Mohammadiha, R. Martin, and A. Leijon, "Spectral domain speech enhancement using HMM state-dependent super-Gaussian priors," *IEEE Signal Process. Letters*, vol. 20, no. 3, pp. 253–256, mar. 2013.

2.1 Speech Enhancement Using Dynamic NMF

One of the straightforward approaches to enhance the NMF decomposition to model time dependencies is to use regularizations in NMF. Motivated by this, we proposed an NMF-based noise PSD estimation algorithm in [134]. In this work, the speech and noise basis matrices are trained offline, after

which a constrained KL-NMF (similar to (22)) is applied to the noisy spectrogram in a frame by frame basis. The added penalty term encourages consecutive speech and noise NMF coefficients to take similar values, and hence, to model the signals' time dependencies. After performing NMF by minimizing the regularized cost function, the instantaneous noise periodogram is obtained as in (26) by switching the role of the speech and noise approximates. This approach and other regularized NMFs, e.g., [122] provide an ad hoc way to use the temporal dependencies, and hence, finding a systematic method to model the temporal dynamics has been investigated in this thesis. Moreover, the Wiener-type estimator in (26) is not optimal in a statistical sense. In the following, we first introduce an approach to obtain an optimal estimator for the speech signal, and then we explain the proposed methods to model the temporal dynamics.

We proposed a linear MMSE estimator for NMF-based speech enhancement in Paper A [133]. In this work, NMF is applied on $\mathbf{v}_t = |\mathbf{y}_t|^p$ for both options of $p = 1$ and $p = 2$ in a frame by frame routine. Let $\mathbf{x}_t = |\mathbf{s}_t|^p$ and $\mathbf{u}_t = |\mathbf{n}_t|^p$ denote the nonnegative transformations of the speech and noise DFT coefficients, respectively. Similar to Section 1.4, we assume that $\mathbf{v}_t = \mathbf{x}_t + \mathbf{u}_t$. Here, a gain variable $\mathbf{g}_t$ is obtained to filter the noisy signal and to estimate the speech signal: $\hat{\mathbf{x}}_t = \mathbf{g}_t \odot \mathbf{v}_t$. Assuming that the basis matrices of speech and noise are obtained during the training stage, and that the NMF coefficients $\mathbf{h}_t$ are random variables, $\mathbf{g}_t$ is derived such that the mean square error between $\hat{\mathbf{x}}_t$ and $\mathbf{x}_t$ is minimized. The optimal gain is shown to be:

$$\mathbf{g}_t = \frac{\boldsymbol{\xi}_t + c^2\sqrt{\boldsymbol{\xi}_t}}{\boldsymbol{\xi}_t + 1 + 2c^2\sqrt{\boldsymbol{\xi}_t}}, \tag{27}$$

where $c = \sqrt{\pi}/2$ for $p = 1$, and $c = \sqrt{2}/2$ for $p = 2$, and $\boldsymbol{\xi}_t$ is called the smoothed speech to noise ratio, which is estimated using a decision-directed approach[6]:

$$\xi_{kt} = \alpha \frac{\hat{x}^2_{k,t-1}}{E\left(\left[\mathbf{W}^{(n)}\mathbf{h}^{(n)}_{t-1}\right]^2_k\right)} + (1-\alpha)\frac{\left[\mathbf{W}^{(s)}\mathbf{h}^{(s)}_{t}\right]^2_k}{E\left(\left[\mathbf{W}^{(n)}\mathbf{h}^{(n)}_{t}\right]^2_k\right)}. \tag{28}$$

The conducted simulations in Paper A [133] using Perceptual Evaluation of Speech Quality (PESQ) [143] and source to distortion ratio (SDR) [145, 146] show that the results using $p = 1$ are superior to those using $p = 2$ (which is in line with previously reported observations, e.g., [122]) and that both of them are better than the results of a state-of-the-art Wiener filter.

In the linear MMSE approach, Paper A [133], we use the speech and noise temporal dependencies to obtain a smooth estimate for ξ_{kt} (28). However,

[6] In Paper A [133], the basis matrices are shown by T and the NMF coefficients of noisy speech, clean speech and noise signals are shown by $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$, respectively.

we do not consider any explicit prior density function to model the temporal dynamics. In a general framework, we can think of some state variables that evolve over time. In NMF, these variables correspond to the NMF coefficients. We can have continuous state-space or discrete state-space formulations. The main underlying assumption for the following approaches is that the NMF coefficients (or activations) are modeled using an autoregressive model such that:

$$E\left(\mathbf{h}_t\right) \quad = \quad \sum_{j=1}^{J} \mathbf{A}_j \mathbf{h}_{t-j}, \tag{29}$$

where each $\mathbf{A}_j$ is the $I \times I$ autoregressive coefficient matrix associated with j-th lag. First, let us assume that the state-space is continuous. The discrete state-space that is referred to as nonnegative HMM will be discussed later. A unified view of different dynamic NMF approaches is provided in [147].

2.1.1 Speech Denoising Using Bayesian NMF

Our proposed approaches in Paper B [27] and [135] assume that different elements of $\mathbf{h}_t$ are independent. This implies that matrices $\mathbf{A}_j, \forall j$ are assumed to be diagonal in (29). Also, each element of $\mathbf{h}_t$ is distributed according to a gamma distribution with a mean value given by (29):

$$f\left(h_{it}\right) = \text{Gamma}\left(h_{it}; \phi_{it}, \tfrac{E(h_{it})}{\phi_{it}}\right), \tag{30}$$

in which $\text{Gamma}\left(h; \phi, \theta\right) = \exp((\phi - 1)\log h - h/\theta - \log\Gamma\left(\phi\right) - \phi \log\theta)$ denotes the gamma density function with ϕ as the shape parameter and θ as the scale parameter, and $\Gamma\left(\phi\right)$ is the gamma function. In Paper B [27], the mean of $\mathbf{h}_t$ is recursively updated using (29) where the diagonal elements of $\mathbf{A}_j$ are exponentially decaying as j increases. Then, the obtained prior distributions are used in a Bayesian formulation of NMF to obtain an MMSE estimator for the clean speech signal in a noise reduction application. In Bayesian terminology, the posterior distribution of the NMF coefficients at the previous time frames are widened and are used as the prior distribution for the current time frame, as shown in Figure 7.

For the speech enhancement in Paper B [27] and [135], the probabilistic NMF from [106] was used. This approach assumes that an input matrix $\mathbf{V}$ is stochastic, and to perform NMF as $\mathbf{V} \approx \mathbf{WH}$ the following model is

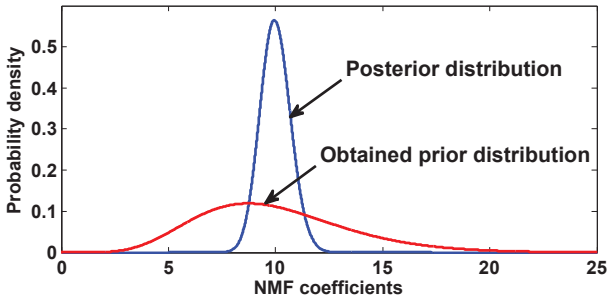

Figure 7: Using the posterior distribution of $\mathbf{h}_{t-1}$ as a prior distribution for $\mathbf{h}_t$.

considered[7]:

$$v_{kt} = \sum_i q_{kit}, \tag{31}$$

$$f(q_{kit}) = \mathcal{PO}(q_{kit}; w_{ki}h_{it}) = (w_{ki}h_{it})^{q_{kit}} e^{-w_{ki}h_{it}} / (q_{kit}!), \tag{32}$$

where $\mathbf{Q} = \{q_{kit}\} \in \mathbb{Z}_+^{K\times I\times T}$ are integer-valued latent variables, $\mathcal{PO}(q;\lambda)$ denotes the Poisson distribution, and $q!$ is the factorial of q. A schematic representation of this model is shown in Figure 8.

In the Bayesian formulation, in addition to the NMF coefficients h_{it}, the basis elements w_{ki} are also assumed to be distributed according to a gamma distribution. As the exact Bayesian inference for (31) and (32) is analytically intractable, a variational Bayes (VB) approach [72] has been proposed in [106] to obtain the approximate posterior distributions of $\mathbf{W}$ and $\mathbf{H}$. In this approximate inference, it is assumed that the posterior distribution of the parameters are independent, and these uncoupled posteriors are inferred iteratively by maximizing a lower bound on the marginal log-likelihood of data (known as the model evidence). This procedure is guaranteed to converge [72].

More specifically for this Bayesian NMF, in an iterative scheme, the current estimates of the posterior distributions of $\mathbf{Q}$ are used to update the posterior distributions of $\mathbf{W}$ and $\mathbf{H}$, and these new posteriors are used to update the posteriors of $\mathbf{Q}$ in the next iteration. The iterations are carried on until convergence. The posterior distributions for $\mathbf{q}_{k,:,t}$ are shown to be multinomial density functions (: denotes 'all the indices'), while for w_{ki} and h_{it} they are gamma density functions.

[7]The latent variables are shown by Z in Paper B [27]. Also, the factorization of the noisy spectrogram is shown as $\mathbf{y} \approx \mathbf{bv}$.

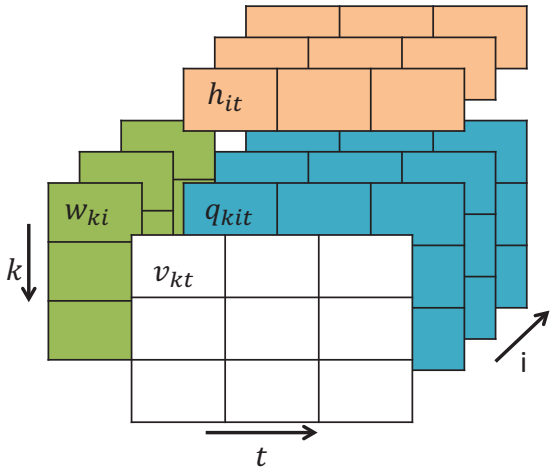

Figure 8: A schematic representation of (31), redrawn from [106].

We use the Bayesian framework from [106] and extend it to devise a noise reduction approach. Our extension mainly includes (1) deriving an MMSE estimator for the speech signal, (2) proposing a method to recursively update the prior distributions of the activations to model the temporal dependencies, and (3) evaluating the developed speech enhancement algorithm. To enhance a given noisy speech signal, the prior distributions (30) are applied in the VB framework to obtain the posterior distributions of $\mathbf{h}_t$. During enhancement, the posterior distributions of the speech and noise basis matrices are held fixed[8]. Assuming that speech and noise spectrograms are additive, the MMSE estimate of the clean speech signal is shown to be:

$$\hat{x}_{kt} = \frac{\sum_{i=1}^{I^{(s)}} e^{E(\log w_{ki} + \log h_{it} | \mathbf{v}_t)}}{\sum_{i=1}^{I^{(s)}+I^{(n)}} e^{E(\log w_{ki} + \log h_{it} | \mathbf{v}_t)}} v_{kt}. \tag{33}$$

We further developed this approach in Paper B [27] and [148] to use it in an unsupervised fashion. For this purpose, two solutions are proposed. In the first one, the BNMF is combined with an HMM, denoted by BNMF-HMM . In this method, each state of the HMM corresponds to one specific noise type whose NMF model is learned offline (See Figure 9). Also, a universal BNMF model is learned for speech that does not introduce any limitation since we do not use any assumption on the identity or gender of the speakers.

[8]These distributions can be obtained offline using some training data. Later, we will shortly mention how the noise basis matrix can be learned online.

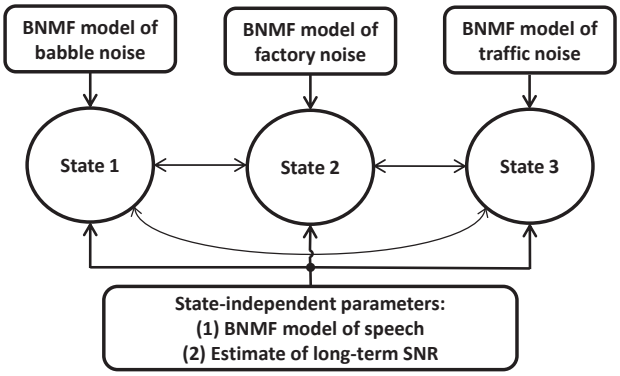

Figure 9: A schematic diagram of BNMF-HMM approach. Source: Paper B [27].

As the second solution, we developed an online noise basis learning algorithm in Paper B [27] and combined it with BNMF to design an unsupervised NMF-based speech enhancement system. The noise adaption scheme is based on a sliding window concept in which the past noisy frames are stored into buffers and are used to update the posterior distribution of the noise basis matrix.

Figure 10 presents a comparison of NMF-based systems. In this experiment (for details, see Paper B [27]), the DFT was applied to frames of 32 ms length. The experiment is performed using the core test set of the TIMIT database (192 sentences) [149]. Moreover, the results are averaged over three noise types of babble, factory and city traffic noises. In this figure, the BNMF-HMM approach is compared with a General-model BNMF in which a single noise dictionary is learned for all the noises. Also, an oracle BNMF is considered in which the noise type is known a priori. Hence, this approach is an ideal case of BNMF-HMM. Similarly, an oracle maximum likelihood implementation of NMF (ML-NMF) and the oracle NHMM [131] are considered for the comparison. Finally, the performance of the NMF-based methods is compared to the speech short-time spectral amplitude estimator using super-Gaussian priors (STSA-GenGamma) [17] with $\gamma = \nu = 1$.

Figure 10 shows the SDR, source to interference ratio (SIR), and source to artifact ratio (SAR) from the BSS-Eval toolbox [145, 146]. The simulations show that the Oracle BNMF has led to the best performance, which is closely followed by BNMF-HMM. For instance, at a 0 dB input SNR, the BNMF-HMM outperforms the STSA-GenGamma by 2 dB in SDR. This shows the superiority of the BNMF approach, and also, it indicates that the HMM-based classification scheme is working successfully. Another in-

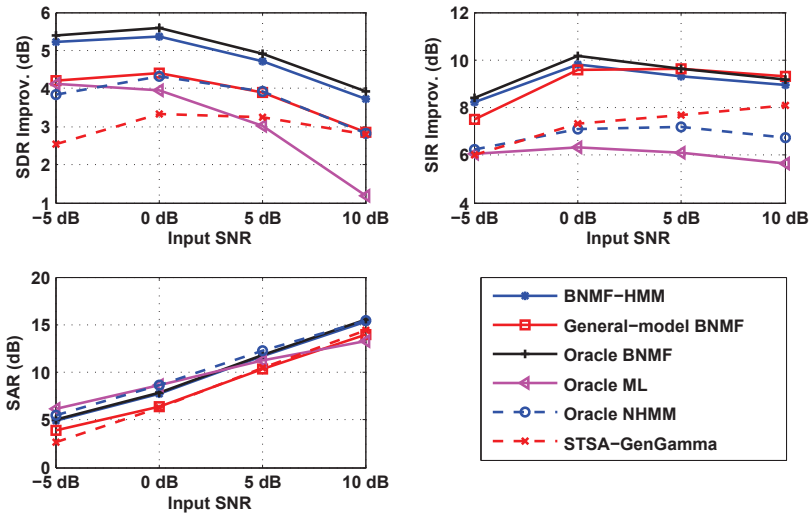

Figure 10: A comparison of NMF-based approaches with a state-of-the-art speech spectral amplitude estimator using super-Gaussian priors (Paper B [27]).

teresting result is that except for the Oracle ML, the other NMF-based techniques outperform STSA-GenGamma. The ML-NMF approach gives a poor noise reduction particularly at high input SNRs. However, after modeling temporal dependencies and using optimal MMSE estimators, the performance of the NMF-based algorithms is improved considerably.

2.1.2 Nonnegative Linear Dynamical Systems

We can write Eq. (29) and nonnegative factorization in a state-space form as

$$\mathbf{h}_t = \sum_{j=1}^{J} \mathbf{A}_j \mathbf{h}_{t-j} + \boldsymbol{\epsilon}_t, \tag{34}$$

$$\mathbf{v}_t = \gamma_t \mathbf{W}\mathbf{h}_t + \boldsymbol{\zeta}_t\,, \tag{35}$$

in which we have considered PLCA for decomposition, hence, $\gamma_t = \sum_k v_{kt}$, $\boldsymbol{\epsilon}_t$ is the process noise, and $\boldsymbol{\zeta}_t$ is the observation noise in the model, Paper E [63][9]. Multiplicative process and measurement noises with $J = 1$ are

[9]These equations are identical to (4) and (5) of Paper E [63] where the NMF coefficients are shown by $\mathbf{v}_t$, basis matrix is represented by $\mathbf{b}$, and observations are denoted by $\mathbf{x}_t$.

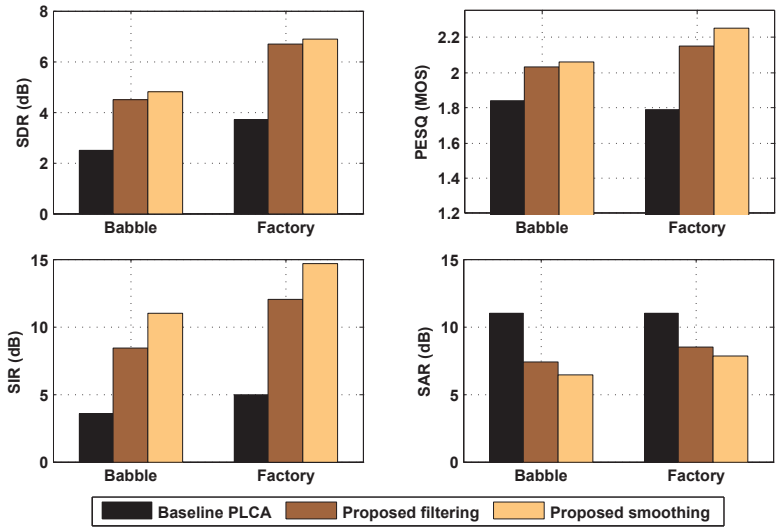

Figure 11: Performance of denoising algorithms for a noisy signal at a 0 dB input SNR (Paper E [63])

considered in [139]. Eq. (34) represents the J-th order vector autoregressive (VAR) model. Moreover, we normalized columns of $\mathbf{A}_j$, and hence, the model can be compared to a multimatrix mixture transition distribution (MTD) model [150]. In contrast to the model used in Section 2.1.1, Eq. (34) does not use the independence assumption on different basis vectors' activities. We have proposed causal filtering and fixed-lag smoothing algorithms in Paper E [63] that use Kalman-like prediction in NMF and PLCA. An important advantage of this method over the factorial NHMM approaches (to be explained in the next section) is that the computational complexity is significantly less for this approach.

Figure 11 presents some results in which (34) and (35) are used for denoising. The results show a significant improvement in SDR, which results in a better-quality denoised speech, as compared to the baseline PLCA. Moreover, the evaluation shows that applying the temporal dynamics has increased the SIR whereas the SAR was reduced compared to the baseline. In fact, the algorithms have led to a fair trade-off between removing noise and introducing artifacts in the enhanced signal. The PESQ values also confirm a very good quality improvement using the proposed algorithms. Additionally, the figure illustrates that the smoothing algorithm has produced slightly better SDR and PESQ values than the filtering approach.

2.1.3 Nonnegative Hidden Markov Model

We consider a discrete state-space in this section. Let $\mathbb{H}_t$ denote a one-of-I random indicator column vector. In the generative model, we first use (29) with $J = 1$ to compute $E(\mathbb{H}_t)$[10] which is then modified using a *winner-take-all* nonlinearity [151] to obtain $\mathbf{h}_t$. Hence, $\mathbf{h}_t$ is a sparse vector whose l_0-norm is one. The non-zero element of $\mathbf{h}_t$ indicates which state of HMM is used to generate data. Then, an observation is obtained by sampling from a desired distribution $f(\mathbf{v}_t; \mathbf{W}, \mathbf{h}_t, \gamma_t)$. In this nonnegative discrete model, the matrix $\mathbf{A}_1$ is in fact the transition matrix of an HMM, and hence the model is referred to as NHMM.

We have proposed an NHMM (Paper C [64] and [111]) in which the HMM state-conditional output distributions ($f(\mathbf{v}_t; \mathbf{W}, \mathbf{h}_t, \gamma_t)$) are assumed to be gamma distributions. The choice of a gamma distribution provides a great flexibility to model audio signals. Since the NMF coefficients $\mathbf{h}_t$ are normalized, gain modeling is required in this approach. In the algorithm developed in Paper C [64], we used a gamma distribution to govern the gain variable γ_t. The mean value of this distribution is time-variant and is updated over time.

Considering the above explanation, NHMM is a sparse NMF approach in which only one basis vector is used at each time to generate an observation. Compared to the continuous dynamical systems in Section 2.1.2, this implies less flexibility. In the next section, we present an extension, which relaxes this limitation, and we use it to enhance a babble-contaminated noisy speech signal, Paper C [64]. Some evaluation results will be also explained in the next section.

2.2 NMF-based Separation of Sources with Similar Dictionaries

In some denoising or source separation applications, the basis matrices of the signals are quite similar. In practice this happens, e.g., when we try to separate speech signals from a mixture in which two speakers have the same gender, or when a speech signal is mixed with a multitalker babble noise. In these cases, the performance of the separation algorithms degrades and the estimated signal might have a high level of artifacts.

Let us first introduce a relaxation of the NHMM explained in Section 2.1.3. A straightforward extension of the sparse NMF can be derived by letting $\mathbf{h}_t$ be non-sparse. For this purpose, we define a fixed weighting matrix $\overline{\mathbf{H}} \in \mathbb{R}_+^{I \times J}$ where I and J can be set to different values. In the generative model, we first obtain a sparse $\mathbf{h}_t$ as before. But to generate an observation, we consider $\overline{\mathbf{H}}\mathbf{h}_t$ instead of $\mathbf{h}_t$, i.e., we sample an observation

[10] Note that, $\mathbb{H}_t$ is an indicator vector but $E(\mathbb{H}_t)$ is a normalized continuous-valued vector.

from $f\left(\mathbf{v}_t; \mathbf{W}, \overline{\mathbf{H}}\mathbf{h}_t, \gamma_t\right)$. This can also be seen as a two-layer NMF [152]. In this view, $\mathbf{h}_t$ acts as an indicator vector and chooses one set of activities, i.e., a column of $\overline{\mathbf{H}}$. The weighting matrix $\overline{\mathbf{H}}$ can be also considered to be time-varying, for more details see [131, 147].

We proposed a probabilistic model for multitalker babble noise in Paper C [64] that is based on NHMM. We modeled the waveform of the babble noise as a weighted sum of M i.i.d. clean speech sources. Therefore, the expected value of the short-time power spectrum vector (periodogram) of babble at time t, $\mathbf{u}_t = |\mathbf{n}_t|^2$, is given by:

$$E\left(\mathbf{u}_t\right) = \sum_{m=1}^{M} E\left(\mathbf{x}_{mt}\right), \tag{36}$$

where $\mathbf{x}_{mt} = |\mathbf{s}_{mt}|^2$ is the power spectrum vector corresponding to the speaker m at time t, while each $\mathbf{x}_{mt}$ is independently generated by an instance of the NHMM described in Section 2.1.3. We first train an NHMM for the clean speech signal and obtain the speech basis matrix $\mathbf{W}^{(s)}$[11]. It is worth to mention again that in the NMF representation obtained using this algorithm most of the elements of $\mathbf{h}_t$ are close to zero[12]. Note that in (36) different weights are used for different speakers as a consequence of the gain modeling which is hidden in (36). Eq. (36) suggests that the basis matrix of babble should be kept the same as that of the speech signal. To describe the babble noise, we use $\mathbf{W}^{(s)}$, and we relax the sparsity of NHMM by learning a weighting matrix $\overline{\mathbf{H}}$. In Paper C [64], we suggested an approach based on the concave-convex procedure (CCCP) [153, 154] to learn $\overline{\mathbf{H}}$ given some training samples of babble noise. The i-th column of this matrix is referred to as a *babble state value vector* and is denoted by $\widehat{\ddot{\mathbf{s}}}'_i$ in Paper C [64]. In this model, which is referred to as gamma NHMM, the babble basis matrix is the same as the speech basis matrix, and only the activation factors (weights) of the basis vectors are different for the two signals over time.

To enhance a babble-contaminated speech signal, the speech and babble HMMs are combined to obtain a factorial HMM. Then, assuming that speech and babble DFT coefficients are complex Gaussian distributed, the MMSE estimate of the speech signal is derived in Paper C [64] that is shown to be a weighted sum of state-dependent Wiener filters. In Section 2.3, we present an extension of this approach that uses super-Gaussian distributions to enhance a noisy signal. The parameters of the gain distributions are time-varying (to adjust to the signal level) in this method. We used a recursive EM algorithm to estimate them over time.

[11] $\mathbf{W}^{(s)}$ is identical to $\hat{\mathbf{b}}$ that is defined after Eq. (7) in Paper C [64].

[12] In performing NMF, when we approximate each observation $\mathbf{x}_t$ as $\gamma_t \mathbf{W}^{(s)} \mathbf{h}_t$, l_0-norm of $\mathbf{h}_t$ is not required to be one, see definition of $\mathbf{u}'$ in the paragraph following Eq. (8) in Paper C [64].

To assess the subjective quality of the estimated speech signal, a subjective MUSHRA listening test [155] was carried out in Paper C [64]. Ten listeners participated in the test. The subjective evaluation was performed for three input SNRs (0 dB, 5 dB, and 10 dB), and for each SNR seven sentences from the core test set of the TIMIT database were presented to the listeners. In each listening session, 5 signals were compared by the listeners: (1) reference clean speech signal, (2) noisy speech signal, (3,4) estimated speech signals using the gamma-NHMM and BNMF [135], and (5) a hidden anchor signal that was chosen to be the noisy signal at a 5 dB lower SNR than the noisy signal processed by the systems. The listeners were asked to rate the signals based on the global speech quality. The results of this listening test, averaged over all of the participants, with a 95% confidence interval are presented in Figure 12. At all of the three SNRs the gamma-NHMM was preferred over the BNMF algorithm. For 0 dB, the difference is 9.5 MOS units, whereas for 5 dB and 10 dB, the preference is around 5 on a scale of 1 to 100. According to the spontaneous comments by the listeners, the remaining noise and artifacts in the enhanced signal by the gamma-NHMM is more like a natural babble-noise while the artifacts introduced by the BNMF are more artificially modulated.

We conclude this section by introducing an alternative approach to separate sources that share some common basis vectors, Paper F [156]. This problem can be seen as a source separation task where we would like to separate one of the sources with low artifacts. In the case of speech enhancement, our desired source is speech for which an undistorted estimate is preferred. One solution to separate a desired source with low artifacts is that we discourage the activation of the common basis vectors in the basis matrix of the interfering source. By doing so, we let the basis vectors of the desired source to take over and explain the mixture signal. In Paper F [156], we proposed a PLCA-based algorithm that can be used for this purpose. In this paper, we argued that the Dirichlet distribution is not suitable as a prior to estimate the nonnegative elements in PLCA, even though it is the conjugate distribution for this purpose. We instead proposed to use an exponential distribution as the prior and showed that it can be used to force some basis vectors to be inactive. Moreover, we derived a MAP approach to identify a set of the common basis vectors and use that to separate a desired source with an arbitrarily low artifacts. Our experiments showed that this approach can be used to obtain a higher quality for the estimated signal by reducing the artifacts.

2.3 Super-Gaussian Priors in HMM-based Enhancement Systems

As mentioned earlier in Section 1.2.3, the real and imaginary parts of the speech (and noise) DFT coefficients are better modeled with super-Gaussian

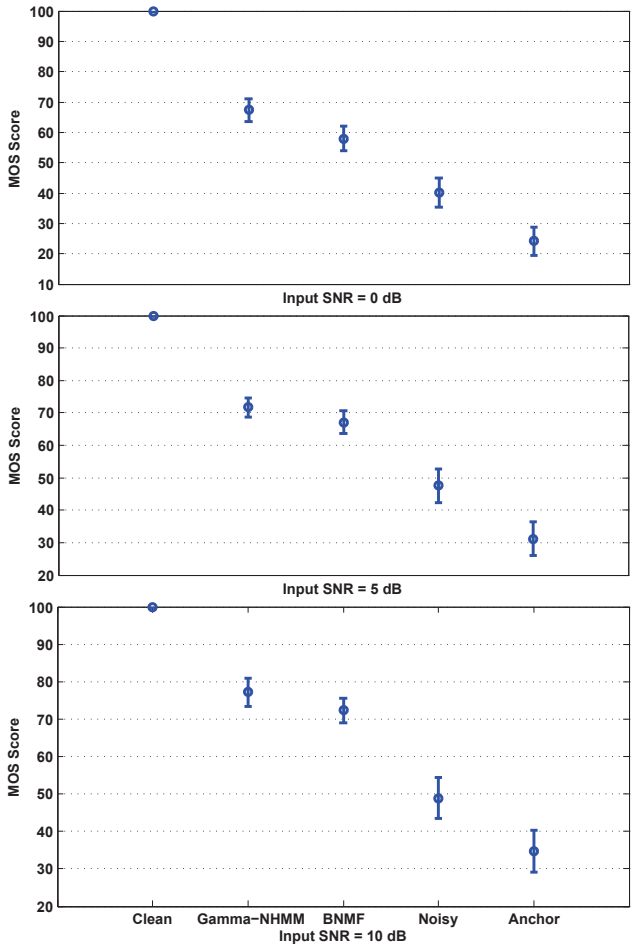

Figure 12: Results of MUSHRA test at 3 input SNRs: 0 dB, 5 dB, 10 dB (top to down) with 95% confidence interval. Source: Paper C [64].

distributions [14]. In the state-of-the-art approaches, the super-Gaussianity is considered for the long-term statistics of a speech signal and is not conditioned on the phoneme type. Hence, an interesting question is whether this phenomenon depends on the phoneme type. We performed an experiment in Paper D [26] aimed to answer this question.

Let us define the conditional gamma distribution as:

$$f\left(x_{kt} \mid z_t = i\right) = \text{Gamma}\left(x_{kt}; \alpha_{ki}, b_{ki}\right), \tag{37}$$

where x_{kt} represents speech magnitude-squared DFT coefficients, $z_t = i \in \{1, \dots I\}$ is the hidden variable, $\text{Gamma}\left(x_{kt}; \alpha_{ki}, b_{ki}\right)$ denotes a gamma density function (as defined in (30)) with α_{ki} and b_{ki} as the state-dependent shape and scale parameters[13]. If $I = 50 \sim 60$, each state is identified roughly by one phoneme. For $\alpha_{ki} = 1$, (37) reduces to an exponential distribution. This corresponds to assuming that real and imaginary parts of the DFT coefficients have a Gaussian distribution. For $\alpha_{ki} < 1$, however, the resulting distribution for DFT coefficients will be super-Gaussian, as shown in Paper D [26].

To obtain the experimental phoneme-conditioned distribution of the speech power spectral coefficients, we used 2000 realizations for each phoneme from the TIMIT database at a sampling rate of 16 kHz. The DFT was applied with a frame length of 20 ms and 50% overlap. The top panel of Figure 13 shows the shape parameters of the estimated gamma distributions for two phonemes, "ah" and "sh". The shape parameters for these two phonemes are less than one at all frequencies. In the bottom panel of Figure 13, the histogram of the power spectral coefficients of "ah" at frequency 2500 Hz (left) and of "sh" at frequency 6000 Hz (right) are shown. Also, the estimated gamma and exponential distributions are shown in this figure for comparison. As a result, we see that the power spectral coefficients have gamma rather than exponential distributions even if we limit the speech data to come from a specific phoneme. Therefore, real and imaginary parts of the phoneme-conditioned speech DFT coefficients have super-Gaussian distributions.

Using this knowledge, and considering that the AR-HMM does not model the spectral fine structure of the voiced speech sounds and may result in low-level resonant noise in some voiced segments [60, 61], we proposed an HMM-based speech spectral enhancement algorithm using super-Gaussian prior distributions in Paper D [26]. In this work, we extended the HMM-based speech enhancement method from Paper C [64] and derived a new MMSE estimator by assuming that the speech power spectral coefficients are gamma-distributed while noise power spectral coefficients are Erlang-distributed. Our simulations show that the performance of the proposed

[13] In Paper D [26], the hidden variables are shown by $\bar{S}_t$, $\ddot{S}_t$, and S_t for speech, noise, and noisy signals, respectively.

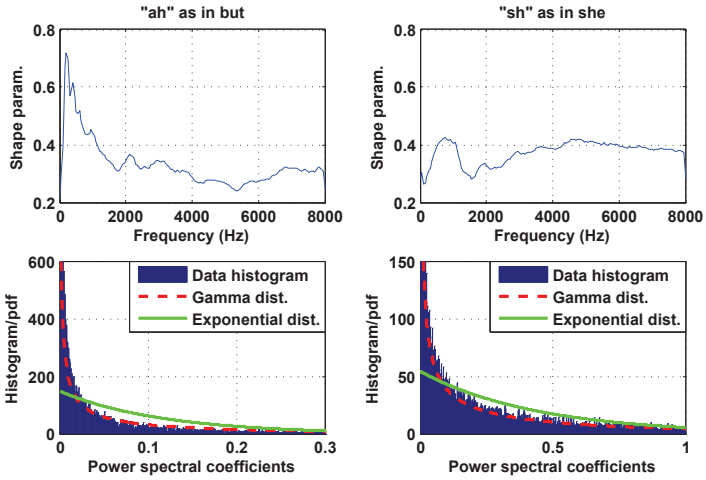

Figure 13: Experimental distribution of the speech power spectral coefficients. The bottom panel shows the fitted distributions and the histogram of the power spectral coefficients of "ah" at frequency 2500 Hz (left) and of "sh" at frequency 6000 Hz (right). Taken from Paper D [26].

denoising algorithm with super-Gaussian priors is superior to that of the algorithm with the Gaussian priors. Hence, the results support the super-Gaussianity hypothesis.

2.4 Discussion

We believe that future supervised speech enhancement algorithms should focus on two objectives: (1) developing algorithms to solve difficult problems, such as the cocktail party problem, for which unsupervised methods cannot provide a satisfactory solution, and (2) designing external or built-in classification techniques allowing supervised approaches to be used in an unsupervised fashion. This dissertation has proposed solutions in both of these directions.

In the proposed NHMM for babble noise, Paper C [64], we have derived a mathematical model for babble noise. The proposed approach provides a generative model for the babble noise that is based on a single-voice speech HMM. We have used this model to successfully derive and evaluate a noise reduction system. Our derivation uses some simplifying assumptions. For example, this work does not provide a systematic way to model the reverberation that might exist in the so-called cocktail party. Nevertheless, if reverberant babble noise is used for the training, the babble model will be

adapted to some effects of reverberation. Moreover, we derived a discrete state-space model for babble while a continuous state-space model might be preferred. However, our objective and subjective MUSHRA listening test indicates that even with these simplifications the proposed approach outperforms the competing algorithms. Therefore, we believe that the suggested method can provide a good basis to develop further signal processing algorithms to solve the cocktail party problem.

Paper B [27] proposes noise reduction algorithms using BNMF in which we have used either a built-in classifier or an online noise dictionary learning scheme to use the method in an unsupervised setting. Our objective evaluation of the proposed unsupervised BNMF-based enhancement system shows that it provides a considerable improvement over state-of-the-art. Our evaluation in this paper (and the other papers included in Part II) is mainly based on the segmental SNR (SegSNR), PESQ, and SDR. SegSNR and PESQ are two of commonly used measures to evaluate the quality of the enhanced signal [2]. SDR is another instrumental measure, recently proposed [145], which measures the overall quality of the speech signal. Since none of these measures can perfectly predict the actual subjective performance of a noise reduction algorithm, performing a formal listening test is therefore suggested as a future work to provide an additional evaluation of Paper B [27]. This can include a MUSHRA listening test to examine the quality of the enhanced speech signals, and word or sentence tests [2] to evaluate the intelligibility improvement provided by the enhancement systems. Moreover, additional study is recommended to formally evaluate the robustness of the algorithms from Paper B [27] in real-world applications.

To design a real-time noise reduction algorithm, we have to carefully select several options, such as filter type (causal or non-causal) and process delay. Causality is an important property of a real-time system, where we do not have access to future data. Therefore, an algorithm needs to only rely on the past data. We have considered this important constraint and most of our proposed approaches are causal. Latency or delay is another very important parameter in designing speech enhancement systems. For example, the total process delay (from input to output) is required to be less than 30 ms in many applications. This requirement implies that (1) we need to use shorter time frames in the DFT analysis, which might degrade the performance of some algorithms [132, 136], and (2) the computational complexity of the algorithm must be low enough to satisfy the application needs. We have evaluated our proposed noise reduction schemes considering these constraints. For example, we used a frame length of 20 ms in Paper C [64] and Paper D [26].

Continuous state-space nonnegative dynamical systems, as in Paper E [63] and [139], can provide a better way to model speech temporal dependencies and can lead to significantly less computational complexity compared to the discrete state-space formulations of dynamic NMF (or nonneg-

ative HMMs), Paper C [64], [111,131,157]. However, NHMMs are proposed earlier and might still be the preferred choice in some applications.

Finally, it is worth mentioning that our findings in Paper D [26] shows that the phoneme-conditioned speech DFT coefficients should be preferably modeled with super-Gaussian prior distributions. This can serve as a base to derive a variety of new estimators for the speech signal, similar to what has happened for the unsupervised methods.

3 Conclusions

This dissertation investigated the application of NMF and HMM in speech enhancement systems. We derived and evaluated speech enhancement algorithms in which an NMF model or HMM is trained for each of the noise and speech signals. We proposed both supervised and unsupervised noise reduction schemes.

The main achievements of this dissertation are summarized as:

- Developing and evaluating a noise reduction algorithm based on a Bayesian NMF with recursive temporal updates of prior distributions. We used temporal dynamics in the form of a prior distribution in a probabilistic formulation of NMF. Moreover, we derived optimal MMSE estimators to estimate the clean speech signal from a noisy recording. We evaluated the developed denoising schemes for different noise types and SNRs. Our experiments showed that a noise reduction system using a maximum likelihood implementation of NMF—with a universal speaker-independent speech model—does not outperform state-of-the-art approaches. However, by incorporating the temporal dependencies and using optimal MMSE filters, the performance of the NMF-based methods increased considerably. For example, our evaluations showed that at a 0 dB input SNR, the proposed BNMF-based speech enhancement method can outperform a speech short-time spectral amplitude estimator using super-Gaussian priors by up to 2 dB in SDR (Paper B).

- Proposing an algorithm to learn the noise NMF model online from the noisy signal. The method was validated through different experiments (Paper B).

- Proposing nonnegative dynamical systems to use the temporal dynamics in NMF. This method was used to develop and evaluate a noise reduction and source separation system. In the case of speech denoising with factory noise at 0 dB input SNR, the developed algorithm outperformed a baseline NMF by 3.2 dB in SDR and around 0.5 MOS in PESQ (Paper E).

- Derivation and evaluation of a linear MMSE estimator for NMF-based speech enhancement. Our experiments showed that using the speech magnitude spectrogram as the observation matrix in NMF leads to a better performance than using the power spectrogram (Paper A).
- Developing a nonnegative HMM for babble noise and using it to design and evaluate a noise reduction system to enhance a babble-contaminated speech signal. The babble model is derived from a single-voice HMM and its basis matrix is similar to that of the speech signal. Here, the main distinction of speech and babble signals is the activity pattern of the basis vectors over time. Objective evaluations and a subjective MUSHRA listening test indicated that the proposed method is capable of strong performance. In our listening test and at a 0 dB input SNR, the enhanced speech of this system was preferred by around 10 MOS units to the enhanced speech of the BNMF and by 27 to the input noisy signal in the scale of 1 to 100 (Paper C).
- Developing a low-artifact source separation scheme using PLCA. The method was used to enhance a babble-contaminated speech signal and to separate speech sources with similar-gender speakers. Our simulations showed that the proposed method not only reduces artifacts but also increases the overall quality of the estimated signal (Paper F).
- Derivation and evaluation of HMM-based speech spectral enhancement algorithms with super-Gaussian prior distributions. Our experiments with the empirical distributions together with the simulation results using the proposed MMSE-based denoising algorithm showed that the speech DFT coefficients rather have super-Gaussian distributions even at the scale of individual phones. Evaluations showed that the proposed speech enhancement system with super-Gaussian priors can outperform a counterpart system with Gaussian priors by up to 0.8 dB in SDR (Paper D).

References

[1] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. West Sussex, England: John Wiley & Sons, 2006.

[2] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 1st ed. Boca Raton, FL: CRC Press, 2007.

[3] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *Journal of Acoustical Society of America (JASA)*, vol. 122, no. 3, pp. 1777–1786, sep. 2007.

[4] H. Levitt, "Noise reduction in hearing aids: An overview," *Journal of Rehabilitation Research and Development*, vol. 38, no. 1, pp. 111–121, 2001.

[5] R. Bentler, Y.-H. Wu, J. Kettel, and R. Hurtig, "Digital noise reduction: Outcomes from laboratory and field studies," *Int. Journal of Audiology*, vol. 47, no. 8, pp. 447–460, 2008.

[6] H. Luts, K. Eneman, J. Wouters *et al.*, "Multicenter evaluation of signal enhancement algorithms for hearing aids," *Journal of Acoustical Society of America (JASA)*, vol. 127, no. 3, pp. 1491–1505, 2010.

[7] J. Sang, "Evaluation of the sparse coding shrinkage noise reduction algorithm for the hearing impaired," Ph.D. dissertation, University of Southampton, jul. 2012.

[8] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 1, 2006, pp. 493–496.

[9] C. H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Process. Letters*, vol. 20, no. 3, pp. 225–228, mar. 2013.

[10] P. N. Petkov, G. E. Henter, and W. B. Kleijn, "Maximizing phoneme recognition accuracy for enhanced speech intelligibility in noise," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 1035–1045, may 2013.

[11] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, apr. 1979.

[12] 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.

[13] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[14] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 13, no. 5, pp. 845–856, sep. 2005.

[15] V. Grancharov, J. Samuelsson, and B. Kleijn, "On causal algorithms for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 3, pp. 764–773, may 2006.

[16] I. Cohen, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *Signal Process.*, vol. 86, no. 4, pp. 698–709, apr. 2006.

[17] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 6, pp. 1741–1752, aug. 2007.

[18] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 7, pp. 1948–1963, sep. 2012.

[19] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, jan. 1982.

[20] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, sep. 2002.

[21] T. Lotter, C. Benien, and P. Vary, "Multichannel speech enhancement using Bayesian spectral amplitude estimation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 1, 2003, pp. 880–883.

[22] B. Cornelis, S. Doclo, T. V. dan Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 2, pp. 342–355, feb. 2010.

[23] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 163–176, jan. 2006.

[24] 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.

[25] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 445–455, sep. 1998.

[26] N. Mohammadiha, R. Martin, and A. Leijon, "Spectral domain speech enhancement using HMM state-dependent super-Gaussian priors," *IEEE Signal Process. Letters*, vol. 20, no. 3, pp. 253–256, mar. 2013.

[27] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using NMF," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 10, pp. 2140–2151, oct. 2013.

[28] R. Talmon, I. Cohen, S. Gannot, and R. R. Coifman, "Supervised graph-based processing for sequential transient interference suppression," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 9, pp. 2528–2538, 2012.

[29] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Prentice Hall, aug. 2009.

[30] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall, 1993.

[31] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Prentice Hall, 2000.

[32] Y. Ephraim and I. Cohen, *Recent Advancements in Speech Enhancement*. in The Electrical Engineering Handbook, CRC Press, 2005.

[33] A. M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communication Systems*, 2nd ed. West Sussex, England: John Wiley & Sons, 2004.

[34] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, jul. 2001.

[35] I. Cohen, "Noise spectrum estimation in adverse environments : Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, sep. 2003.

[36] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, mar. 2010, pp. 4266–4269.

[37] 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, 2012.

[38] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, may 2011, pp. 4640–4643.

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

[40] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 373–385, 1998.

[41] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Process.*, vol. 39, pp. 1732–1742, 1991.

[42] W.-R. Wu and P.-C. Chen, "Subband Kalman filtering for speech enhancement," *IEEE Trans. on Circuits and Systems II: Analog and Digital Signal Process.*, vol. 45, no. 8, pp. 1072–1083, 1998.

[43] D. C. Popescu and I. Zeljković, "Kalman filtering of colored noise for speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 2, 1998, pp. 997–1000.

[44] P. S. Maybeck, *Stochastic Models, Estimation, and Controlled, volume 1.* Academic, 1979.

[45] D. Simon, *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches.* John Wiley & Sons, 2006.

[46] W. A. Pearlman and R. M. Gray, "Source coding of the discrete Fourier transform," *IEEE Trans. on Information Theory*, vol. 24, no. 6, pp. 683–692, nov. 1978.

[47] D. Brillinger, *Time Series: Data Analysis and Theory.* San Francisco: CA: Holden-Day, 1981.

[48] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Applied Signal Process.*, vol. 2005, pp. 1110–1126, 2005.

[49] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Proc. Int. Workshop on Acoust. Echo and Noise Control (IWAENC)*, 2010.

[50] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 9, 1984, pp. 53–56.

[51] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 1, 2002, pp. 253–256.

[52] B. Chen and P. C. Loizou, "A Laplacian-based MMSE estimator for speech enhancement," *Speech Communication*, vol. 49, no. 2, pp. 134–143, feb. 2007.

[53] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, 2008, pp. 4037–4040.

[54] M. Abramowitz and I. A. Stegun, Eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover, 1964. [Online]. Available: http://people.math.sfu.ca/ cbm/aands/

[55] P. Händel, "Power spectral density error analysis of spectral subtraction type of speech enhancement methods," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007.

[56] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, apr. 1985.

[57] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 334–341, jul. 2003.

[58] T. V. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 383–389, sep. 1996.

[59] Y. Ephraim, "Gain-adapted hidden Markov models for recognition of clean and noisy speech," *IEEE Trans. Signal Process.*, vol. 40, no. 6, pp. 1303–316, jun. 1992.

[60] 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.

[61] H. Veisi and H. Sameti, "Speech enhancement using hidden Markov models in Mel-frequency domain," *Speech Communication*, vol. 55, no. 2, pp. 205–220, feb. 2013.

[62] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Proc. Int. Conf. Spoken Language Process. (Interspeech)*, 2008, pp. 411–414.

[63] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Prediction based filtering and smoothing to exploit temporal dependencies in NMF," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, may 2013, pp. 873–877.

[64] N. Mohammadiha and A. Leijon, "Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 998–1011, may 2013.

[65] P. Gaunard, C. G. Mubikangiey, C. Couvreur, and V. Fontaine, "Automatic classification of environmental noise events by hidden Markov models," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 6, may 1998, pp. 3609–3612.

[66] K. El-Maleh, A. Samouelian, and P. Kabal, "Frame level noise classification in mobile environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 1, mar. 1999, pp. 237–240.

[67] L. Ma, D. J. Smith, and B. P. Milner, "Context awareness using environmental noise classification," in *European Conf. on Speech Communication and Technology (ISCA)*, 2003, pp. 2237–2240.

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

[69] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. Ser. B. 39. l, pp. 1–38, 1977.

[70] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," U.C. Berkeley, Tech. Rep. ICSI-TR-97-021, 1997.

[71] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *The Annals of Mathematical Statistics*, vol. 41, pp. 164–171, 1970.

[72] C. M. Bishop, *Pattern Recognition and Machine Learning*. NewYork: Springer-Verlag, 2006.

[73] P. L. Ainsleigh, "Theory of continuous-state hidden Markov models and hidden Gauss-Markov models," Naval Undersea Warfare Center, Newport, Rhode Island, USA, Tech. Rep., mar. 2001.

[74] G. Panahandeh, N. Mohammadiha, A. Leijon, and P. Händel, "Continuous hidden Markov model for pedestrian activity classification and gait analysis," *IEEE Trans. on Instrumentation and Measurement*, vol. 62, no. 5, pp. 1073–1083, may 2013.

[75] M. R. Gahrooei and D. Work, "Estimating traffic signal phases from turning movement counters," in *IEEE Conf. on Intelligent Transportation Systems*, apr. 2013.

[76] O. Cappé, E. Moulines, and T. Ryden, *Inference in Hidden Markov Models*, ser. Springer Series in Statistics. New York, Inc. Secaucus, NJ, USA: Springer, 2005.

[77] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, "The infinite hidden Markov model," in *Advances in Neural Information Process. Systems (NIPS)*. MIT Press, 2002, pp. 29–245.

[78] J. V. Gael, Y. W. Teh, and Z. Ghahramani, "The infinite factorial hidden Markov model," in *Advances in Neural Information Process. Systems (NIPS)*. MIT Press, 2008, pp. 1017–1024.

[79] Y. Ephraim, D. Malah, and B. H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1846–1856, dec. 1989.

[80] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, pp. 245–273, nov. 1997.

[81] R. M. Gray, "Toeplitz and circulant matrices: A review," *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[82] D. Y. Zhao, W. B. Kleijn, A. Ypma, and B. de Vries, "Online noise estimation using stochastic-gain HMM for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 4, pp. 835–846, may 2008.

[83] D. M. Titterington, "Recursive parameter estimation using incomplete data," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 46, pp. 257–267, 1984. [Online]. Available: http://www.jstor.org/stable/2345509

[84] E. Weinstein, M. Feder, and A. V. Oppenheim, "Sequential algorithms for parameter estimation based on the Kullback-Leibler information measure," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 9, pp. 1652–1654, sep. 1990.

[85] V. Krishnamurthy and J. B. Moore, "On-line estimation of hidden Markov model parameters based on the Kullback-Leibler information measure," *IEEE Trans. Signal Process.*, vol. 41, no. 8, pp. 2557–2573, aug. 1993.

[86] 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)*, apr. 1990, pp. 845–848.

[87] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Process. Systems (NIPS)*. MIT Press, 2000, pp. 793–799.

[88] M. H. Radfar, A. H. Banihashemi, R. M. Dansereau, and A. Sayadiyan, "Nonlinear minimum mean square error estimator for mixture-maximisation approximation," *Electronics Letters*, vol. 42, no. 12, pp. 724–725, 2006.

[89] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 2, may 2004, pp. 817–820.

[90] J. R. Hershey, T. Kristjansson, S. Rennie, and P. A. Olsen, "Single channel speech separation using factorial dynamics," in *Advances in Neural Information Process. Systems (NIPS)*. MIT Press, 2007, pp. 593–600.

[91] S. Rennie, P. Olsen, J. Hershe, and T. Kristjansson, "The iroqouis model: Separating multiple speakers using temporal constraints," in *Workshop on Statistical and Perceptual Audition*, 2006.

[92] T. Virtanen, "Speech recognition using factorial hidden Markov models for separation in the feature space," in *Proc. Int. Conf. Spoken Language Process. (Interspeech)*, 2006.

[93] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation.* New York: John Wiley & Sons, 2009.

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

[95] ——, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Process. Systems (NIPS)*. MIT Press, 2000, pp. 556–562.

[96] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[97] S. Choi, "Algorithms for orthogonal nonnegative matrix factorization," in *IEEE Int. Joint Conf. on Neural Networks*, 2008.

[98] H. Lantéri, C. Theys, C. Richard, and C. Févotte, "Split gradient method for nonnegative matrix factorization," in *Proc. European Signal Process. Conf. (EUSIPCO)*, aug. 2010, pp. 1199–1203.

[99] R. Zdunek and A. Cichocki, "Fast nonnegative matrix factorization algorithms using projected gradient approaches for large-scale problems," *Computational Intelligence and Neuroscience*, vol. 2008, 2008.

[100] N. Mohammadiha and A. Leijon, "Nonnegative matrix factorization using projected gradient algorithms with sparseness constraints," in *IEEE Int. Symp. on Signal Process. and Information Technology (ISSPIT)*, dec. 2009, pp. 418–423.

[101] A. Cichocki, R. Zdunek, and S. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 5, may 2006.

[102] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[103] H. Hu, N. Mohammadiha, J. Taghia, A. Leijon, M. E. Lutman, and S. Wang, "Sparsity level in a non-negative matrix factorization based speech strategy in cochlear implants," in *Proc. European Signal Process. Conf. (EUSIPCO)*, aug. 2012, pp. 2432–2436.

[104] M. V. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as non-negative factorizations," *in special issue on Advances in Non-negative Matrix and Tensor Factorization, Computational Intelligence and Neuroscience Journal*, may 2008.

[105] C. Févotte and A. T. Cemgil, "Nonnegative matrix factorisations as probabilistic inference in composite models," in *Proc. European Signal Process. Conf. (EUSIPCO)*, vol. 47, 2009, pp. 1913–1917.

[106] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, 2009, article ID 785152, 17 pages.

[107] 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, pp. 793–830, 2009.

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

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

[110] M. D. Hoffman, D. M. Blei, and P. R. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *Proc. Int. Conf. Machine Learning (ICML)*, 2010, pp. 439–446.

[111] N. Mohammadiha, W. B. Kleijn, and A. Leijon, "Gamma hidden Markov model as a probabilistic nonnegative matrix factorization," in *Proc. European Signal Process. Conf. (EUSIPCO)*, sep. 2013.

[112] P. Smaragdis, B. Raj, and M. V. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *Advances in Models for Acoustic Process. Workshop, NIPS*. MIT Press, 2006.

[113] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. of the 22nd annual Int. ACM SIGIR Conf. on Research and development in information retrieval*, 1999.

[114] ——, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, pp. 177–196, 2001.

[115] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[116] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora." in *Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing*, vol. 1. Association for Computational Linguistics, 2009.

[117] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. Int. Conf. Machine Learning (ICML)*. ACM, 2006.

[118] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[119] M. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete latent variable decomposition of counts data," in *Advances in Neural Information Process. Systems (NIPS)*. MIT Press, 2007, pp. 1313–1320.

[120] P. Smaragdis, B. Raj, and M. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, apr. 2008, pp. 2069–2072.

[121] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *Advances in Neural Information Process. Systems (NIPS)*. MIT Press, 2009, pp. 1705–1713.

[122] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 1066–1074, 2007.

[123] 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.

[124] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Applications of Signal Process. Audio Acoust. (WASPAA)*, oct. 2003, pp. 177–180.

[125] P. Smaragdis, B. Raj, and M. V. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. of the Int. Conf. on Independent Component Analysis and Signal Separation*, sep. 2007.

[126] T. Virtanen and A. T. Cemgil, "Mixtures of gamma priors for non-negative matrix factorization based speech separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation*, 2009, pp. 646–653.

[127] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, 2008, pp. 1825–1828.

[128] M. N. Schmidt, "Single-channel source separation using non-negative matrix factorization," Ph.D. dissertation, Technical University of Denmark, 2008.

[129] 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.

[130] M. N. Schmidt and J. Larsen, "Reduction of non-stationary noise using a non-negative latent variable decomposition," in *IEEE Workshop on Machine Learning for Signal Process. (MLSP)*, oct. 2008, pp. 486–491.

[131] 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)*, may 2011, pp. 17–20.

[132] N. Mohammadiha and A. Leijon, "Model order selection for non-negative matrix factorization with application to speech enhancement," KTH Royal Institute of Technology, Tech. Rep., 2011. [Online]. Available: http://kth.diva-portal.org/smash/record.jsf?pid=diva2:447310

[133] 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 Applications of Signal Process. Audio Acoust. (WASPAA)*, oct. 2011, pp. 45–48.

[134] ——, "A new approach for speech enhancement based on a constrained nonnegative matrix factorization," in *IEEE Int. Symp. on Intelligent Signal Process. and Communication Systems (ISPACS)*, dec. 2011, pp. 1–5.

[135] N. Mohammadiha, J. Taghia, and A. Leijon, "Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, mar. 2012, pp. 4561–4564.

[136] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 1, pp. 1–12, jan. 2007.

[137] C. Févotte, "Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, may 2011, pp. 1980–983.

[138] R. Badeau, "Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (HR-NMF)," in *Proc. IEEE Workshop Applications of Signal Process. Audio Acoust. (WASPAA)*, 2011, pp. 253–256.

[139] C. Févotte, J. L. Roux, and J. R. Hershey, "Non-negative dynamical system with application to speech and audio," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, 2013.

[140] M. Kim, P. Smaragdis, G. G. Ko, and R. A. Rutenbar, "Stereophonic spectrogram segmentation using Markov random fields," in *IEEE Int. Workshop on Machine Learning for Signal Process. (MLSP)*, 2012, pp. 1–6.

[141] M. Kim and P. Smaragdis, "Single channel source separation using smooth nonnegative matrix factorization with Markov random fields," in *IEEE Int. Workshop on Machine Learning for Signal Process. (MLSP)*, sep. 2013.

[142] "Speech processing, transmission and quality aspects (STQ), distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," Tech. Rep. ETSI ES 202 050 V1.1.5, 2007.

[143] I.-T. P.862, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assesment of narrowband telephone networks and speech codecs," Tech. Rep., 2000.

[144] B. Raj, R. Singh, and T. Virtanen, "Phoneme-dependent NMF for speech enhancement in monaural mixtures," in *Proc. Int. Conf. Spoken Language Process. (Interspeech)*, 2011, pp. 1217–1220.

[145] 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, 2006.

[146] C. Févotte, R. Gribonval, and E. Vincent, "Bss_Eval toolbox user guide," IRISA, Rennes, France, Tech. Rep. 1706, apr. 2005.

[147] P. Smaragdis, C. Févotte, N. Mohammadiha, G. J. Mysore, and M. Hoffman, "A unified view of static and dynamic source separation using non-negative factorizations," 2013, to be submitted.

[148] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Simultaneous noise classification and reduction using a priori learned models," in *IEEE Int. Workshop on Machine Learning for Signal Process. (MLSP)*, sep. 2013.

[149] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "TIMIT acoustic-phonetic continuous speech corpus." Philadelphia: Linguistic Data Consortium, 1993.

[150] A. Berchtold and A. E. Raftery, "The mixture transition distribution model for high-order Markov chains and non-Gaussian time series," *Statistical Science*, vol. 17, no. 3, pp. 328–356, 2002.

[151] S. T. Roweis and Z. Ghahramani, "A unifying review of linear Gaussian models," *Neural computation*, vol. 11, no. 2, pp. 305–345, oct. 1999.

[152] A. Cichocki and R. Zdunek, "Multilayer nonnegative matrix factorization using projected gradient approaches," *Int. Journal of Neural Systems*, vol. 17, no. 6, pp. 431–446, 2007.

[153] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, pp. 915–936, 2003.

[154] B. K. Sriperumbudur and G. R. G. Lanckriet, "On the convergence of the concave-convex procedure," in *Advances in Neural Information Process. Systems (NIPS)*. MIT Press, 2009, pp. 1759–1767.

[155] *Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*, ITU-R Recommendation BS.1534-1 Std., 2001-2003. [Online]. Available: http://www.itu.int

[156] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Low-artifact source separation using probabilistic latent component analysis," in *Proc. IEEE Workshop Applications of Signal Process. Audio Acoust. (WASPAA)*, oct. 2013.

[157] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. IEEE Workshop Applications of Signal Process. Audio Acoust. (WASPAA)*, oct. 2009, pp. 121–124.

Part II

Included papers