

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

Copyright ©2013 Nasser Mohammadiha except where otherwise stated. All rights reserved.

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. Taghia, 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 Non-negative 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 Smaragdīs 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 Dic- tionaries	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 Sep-	C21
	arately	
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 Coef-	D2
	ficients	
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 Tem-	
	poral 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

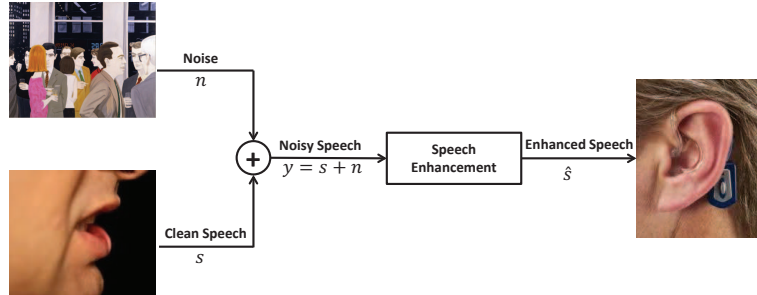


Figure 1: A simplified diagram of a speech enhancement system: the corrupting noise is assumed to be additive.

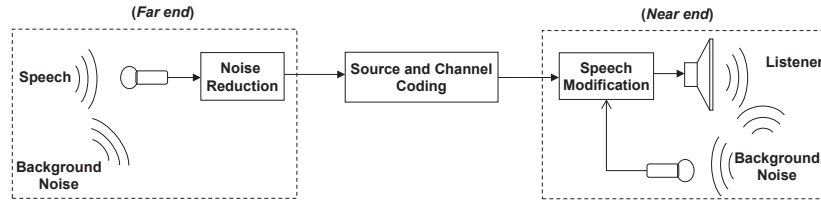


Figure 2: A schematic diagram of a speech communication system. Interfering background noise is present in both the speaker side (*far end*) and the listener side (*near end*).

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:

$$y_m = s_m + n_m. \quad (1)$$

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:

$$\hat{\mathbf{s}}_t = \mathbf{h}_t \odot \mathbf{y}_t, \quad (2)$$

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 independent², we can minimize the MSE for each individual frequency bin k separately:

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

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:

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

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

¹For 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.

²In 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:

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

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

The optimal weight vector can now be written as:

$$h_{kt} = \frac{\xi_{kt}}{\xi_{kt} + 1}. \quad (7)$$

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:

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

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), $E(|n_{kt}|^2)$. 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].

³For 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:

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

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:

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

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

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

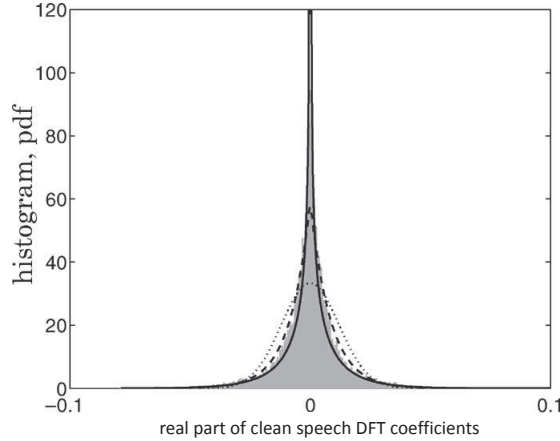


Figure 3: (Dotted) Gaussian, (dashed) Laplace, and (solid) gamma densities fitted to a histogram (shaded) of the real part of clean speech DFT coefficients. Frame length is 32 ms, the sampling rate is 8000 Hz (Source: [14]).

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]

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

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]:

$$\hat{x}_{kt} = E(x_{kt} | v_{kt}) = \frac{\int_0^\infty x_{kt} f(x_{kt}) f(v_{kt} | x_{kt}) dx_{kt}}{\int_0^\infty f(x_{kt}) f(v_{kt} | x_{kt}) dx_{kt}}, \quad (13)$$

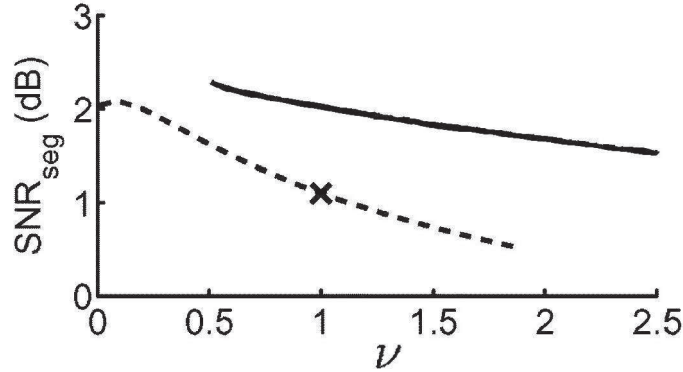


Figure 4: Comparing the performance of Gaussian and super-Gaussian models in terms of segmental SNR as a function of the parameter ν defined in (12). The “cross” corresponds to the Gaussian assumption, the solid and dashed lines correspond to setting $\gamma = 1$ and $\gamma = 2$ in (12), respectively (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 experiment⁴. 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].

⁴In 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).

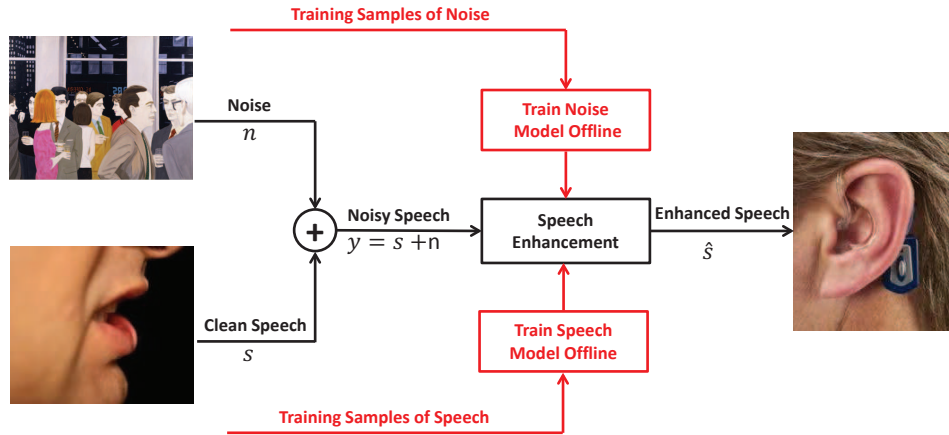


Figure 5: Schematic diagram of a typical supervised speech enhancement system (compare to Figure 1).

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 $P(s_t | z_t = j), j \in \{1 \dots J\}$ 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 $q_j = P(z_1 = j)$ and a transition matrix between the states, denoted by \mathbf{A} with elements $a_{ij} = P(z_t = j | z_{t-1} = i)$.

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 $f(\mathbf{s}_1^T, z_1^T; \lambda)$ is much easier than directly maximizing $f(\mathbf{s}_1^T; \lambda)$. In the E step of the EM algorithm, a lower bound is obtained on $\log(f(\mathbf{s}_1^T; \lambda))$, and in the M step, this lower bound is maximized [72, Chapter 9]. The EM lower bound takes the form

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

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 $Q(\hat{\lambda}, \lambda)$) 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):

$$y_m = s_m + n_m. \quad (15)$$

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

$$f(\mathbf{s}_1^T; \lambda) = \sum_{z_1} \dots \sum_{z_T} \prod_{t=1}^T a_{z_{t-1}z_t} f(\mathbf{s}_t | z_t; \boldsymbol{\theta}), \quad (16)$$

where $a_{z_0z_1} \triangleq P(z_1)$ is the probability of the initial state z_1 and $f(\mathbf{s}_t | z_t; \boldsymbol{\theta})$ is given by a state-dependent GMM:

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

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 adaptation 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:

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

where $D(\mathbf{V} \|\hat{\mathbf{V}})$ is a cost function [94, 95]. The NMF problem is 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]:

$$w_{ki} \leftarrow w_{ki} \frac{[\mathbf{V}\mathbf{H}^\top]_{ki}}{[\mathbf{W}\mathbf{H}\mathbf{H}^\top]_{ki}}, \quad \forall k, i, \quad (19)$$

$$h_{it} \leftarrow h_{it} \frac{[\mathbf{W}^\top \mathbf{V}]_{it}}{[\mathbf{W}^\top \mathbf{W}\mathbf{H}]_{it}}, \quad \forall i, t. \quad (20)$$

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

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

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:

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

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., $D(v_{kt} \parallel \hat{v}_{kt}) = D(\gamma v_{kt} \parallel \gamma \hat{v}_{kt})$). 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

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

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 $\mathbf{v}_t \approx \gamma_t \boldsymbol{\theta}_t = \gamma_t (\mathbf{W}\mathbf{h}_t)$. 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:

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

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

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 $D_{\text{PLCA}} = \sum_t \gamma_t D_{\text{KL}}(\boldsymbol{\lambda}_t \parallel \hat{\boldsymbol{\lambda}}_t)$ where $\boldsymbol{\lambda}_t = \mathbf{v}_t/\gamma_t$, $\hat{\boldsymbol{\lambda}}_t = \mathbf{W}\mathbf{h}_t$, and

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

$D_{\text{KL}}(\boldsymbol{\lambda}_t \|\hat{\boldsymbol{\lambda}}_t) = \sum_k \lambda_{kt} \log \frac{\lambda_{kt}}{\hat{\lambda}_{kt}}$ 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}_{tr} and \mathbf{U}_{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 $\mathbf{W} = [\mathbf{W}^{(s)} \mathbf{W}^{(n)}]$ (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., $\mathbf{v}_t \approx \mathbf{W}\mathbf{h}_t = [\mathbf{W}^{(s)} \mathbf{W}^{(n)}] \left[\mathbf{h}_t^{(s)\top} \mathbf{h}_t^{(n)\top} \right]^\top$. Finally, an estimate of the clean speech spectrum is obtained by a Wiener-type filtering as:

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

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-

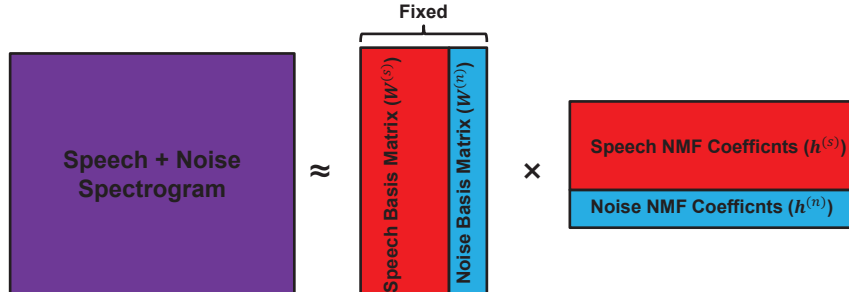


Figure 6: Applying NMF on noisy speech.

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 is 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:

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

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 approach⁶:

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

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,

⁶In 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:

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

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):

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

in which $\text{Gamma}(h; \phi, \theta) = \exp((\phi - 1) \log h - h/\theta - \log \Gamma(\phi) - \phi \log \theta)$ denotes the gamma density function with ϕ as the shape parameter and θ as the scale parameter, and $\Gamma(\phi)$ 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

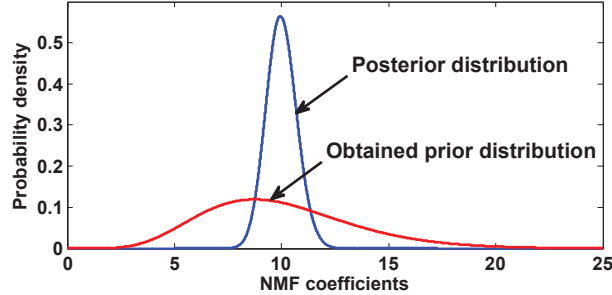


Figure 7: Using the posterior distribution of \mathbf{h}_{t-1} as a prior distribution for \mathbf{h}_t .

considered⁷:

$$v_{kt} = \sum_i q_{kit}, \quad (31)$$

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

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.

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

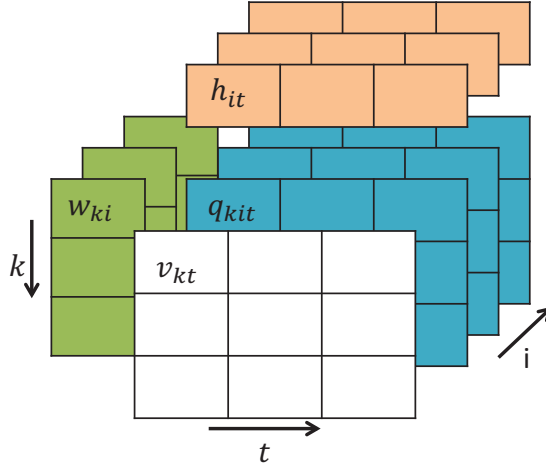


Figure 8: A schematic representation of (31), redrawn from [106].

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 fixed⁸. Assuming that speech and noise spectrograms are additive, the MMSE estimate of the clean speech signal is shown to be:

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

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.

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

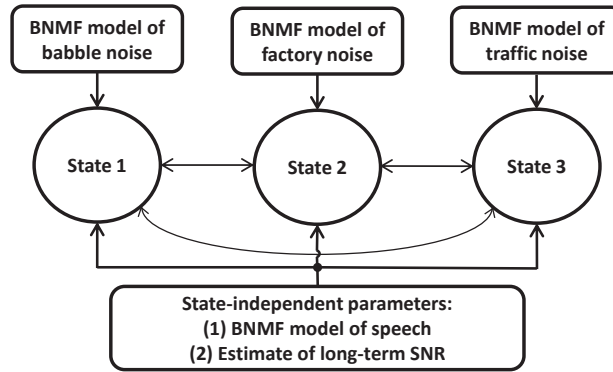


Figure 9: A schematic diagram of BNMF-HMM approach. Source: Paper B [27].

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-

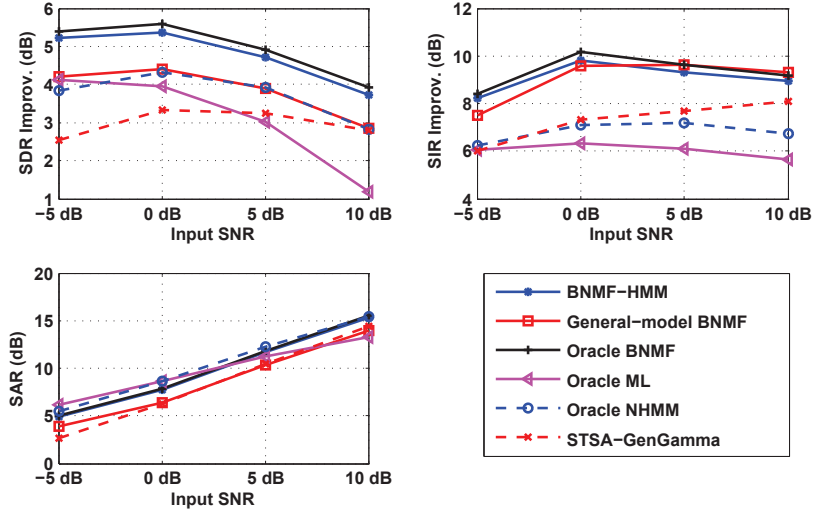


Figure 10: A comparison of NMF-based approaches with a state-of-the-art speech spectral amplitude estimator using super-Gaussian priors (Paper B [27]).

interesting 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

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

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

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

⁹These 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 .

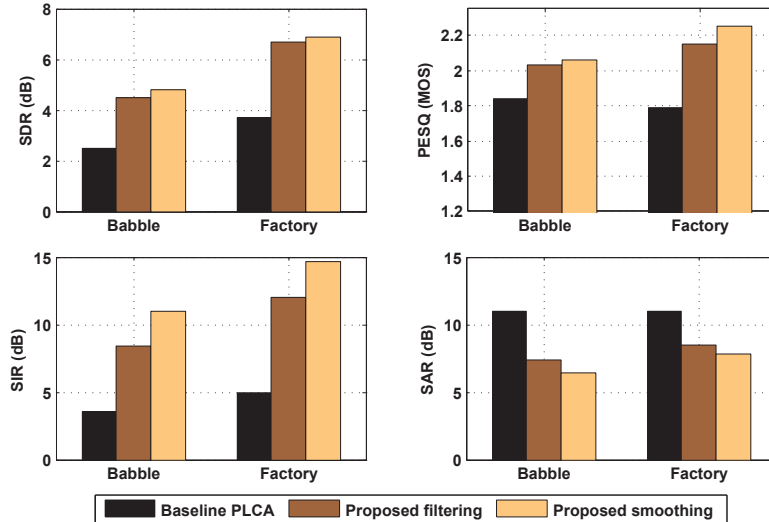


Figure 11: Performance of denoising algorithms for a noisy signal at a 0 dB input SNR (Paper E [63])

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

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

from $f(\mathbf{v}_t; \mathbf{W}, \overline{\mathbf{H}}\mathbf{h}_t, \gamma_t)$. 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:

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

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 zero¹². 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 $\hat{\mathbf{s}}_i^b$ 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].

¹²In 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

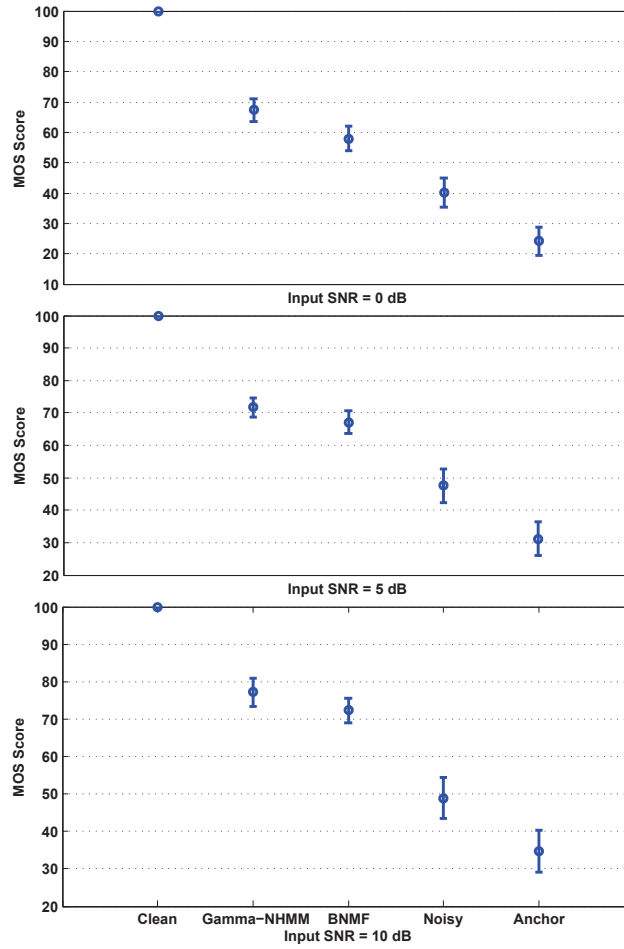


Figure 12: Results of MUSHRA test at 3 input SNRs: 0 dB, 5 dB, 10 dB (top to down) with 95% confidence interval. Source: Paper C [64].

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:

$$f(x_{kt} | z_t = i) = \text{Gamma}(x_{kt}; \alpha_{ki}, b_{ki}), \quad (37)$$

where x_{kt} represents speech magnitude-squared DFT coefficients, $z_t = i \in \{1, \dots, I\}$ is the hidden variable, $\text{Gamma}(x_{kt}; \alpha_{ki}, b_{ki})$ denotes a gamma density function (as defined in (30)) with α_{ki} and b_{ki} as the state-dependent shape and scale parameters¹³. 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

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

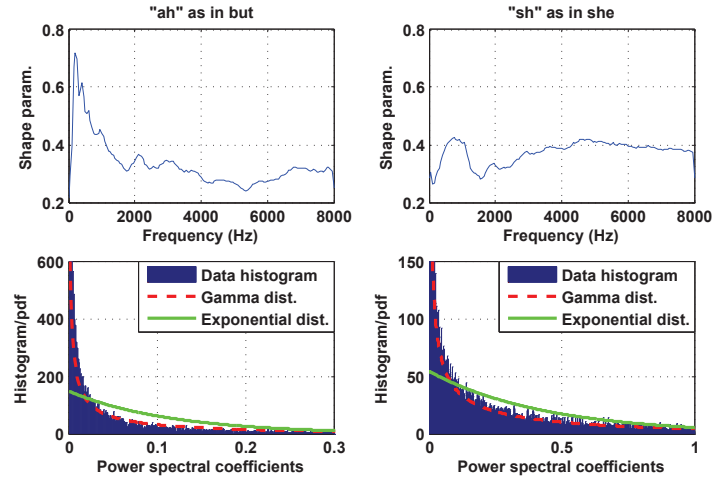


Figure 13: Experimental distribution of the speech power spectral coefficients. The bottom panel shows the fitted distributions and the histogram of the power spectral coefficients of “ah” at frequency 2500 Hz (left) and of “sh” at frequency 6000 Hz (right). Taken from Paper D [26].

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 non-stationary 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. Schaffer, *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 Non-linear 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. 1, 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*. New York: 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. Titterton, “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 iroquois 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 (IS-SPIT)*, 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, “Convolutional 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. 1980–1983.

- [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

Paper A

A New Linear MMSE Filter for Single Channel Speech Enhancement Based on Nonnegative Matrix Factorization

Nasser Mohammadiha, Timo Gerkmann, and Arne Leijon

Refereed paper published in
*Proceedings of IEEE Workshop on Applications of Signal Processing to
Audio and Acoustics (WASPAA)*, oct. 2011, pp. 45–48.

©2013 IEEE
Layout has been revised for thesis consistency

A New Linear MMSE Filter for Single Channel Speech Enhancement Based on Nonnegative Matrix Factorization

Nasser Mohammadiha, Timo Gerkmann, and Arne Leijon

Abstract

In this paper, a linear MMSE filter is derived for single-channel speech enhancement which is based on Nonnegative Matrix Factorization (NMF). Assuming an additive model for the noisy observation, an estimator is obtained by minimizing the mean square error between the clean speech and the estimated speech components in the frequency domain. In addition, the noise power spectral density (PSD) is estimated using NMF and the obtained noise PSD is used in a Wiener filtering framework to enhance the noisy speech. The results of the both algorithms are compared to the result of the same Wiener filtering framework in which the noise PSD is estimated using a recently developed MMSE-based method. NMF based approaches outperform the Wiener filter with the MMSE-based noise PSD tracker for different measures. Compared to the NMF-based Wiener filtering approach, Source to Distortion Ratio (SDR) is improved for the evaluated noise types for different input SNRs using the proposed linear MMSE filter.

1 Introduction

In this paper, we consider a supervised approach based on Nonnegative Matrix Factorization (NMF) to enhance the noisy speech signal. NMF finds a locally optimal solution to solve the matrix equation $X \approx TV$ under the nonnegativeness constraint on T and V . For NMF based speech enhancement or audio source separation, X is the magnitude or power spectrogram of the observed signal, where spectra are stored column-wise in X . NMF is applied to factorize the spectrogram into a matrix consisting of NMF basis vectors, T , and the NMF coefficients matrix, V , which represents the activity of each basis vector over time. NMF has been widely used for blind

source separation (BSS) and speech enhancement recently; after performing NMF on the mixed observed speech signal, in general, the separated or enhanced signal is obtained using one of the following approaches: 1) as a weighted sum of the basis vectors, where weighting factors are given by the NMF coefficients matrix, V [1–4]. 2) by a product of a Wiener-type soft mask and the observation matrix X [5]. In this paper we aim to obtain an optimal soft mask for enhancing the noisy signal.

We use a supervised algorithm in which the noise type is known a priori. One separate basis matrix is obtained for each noise type, and one basis matrix is derived for the speech signal using the training data. A standard NMF is used in the training part, and contribution is made to find a better enhancement procedure using the trained basis matrices. First, we consider a Wiener filtering framework in which the noise PSD is obtained using NMF. Second, we derive a linear minimum mean square error (LMMSE) estimator for the speech signal. Assuming an additive model for the noisy observation, mean square error between the clean speech and the estimated speech components is minimized in the frequency domain to find the estimated speech. The performance of the enhancement algorithms is measured with different instrumental measures including PESQ, SDR, segmental speech SNR, and segmental noise reduction.

2 Notation and Basic Concepts

To refer to the (k, τ) th entry of a matrix X , we use either of notations $X_{k,\tau}$ or $[X]_{k,\tau}$; \mathbf{x}_τ denotes the τ th column of a matrix X , and $[\mathbf{x}]_k$ denotes the k th element of the vector \mathbf{x} . Let $Y_{k,\tau}$ denote the DFT coefficient for frequency bin k and time-frame τ of the noisy signal. The observation for NMF is obtained by taking the (element-wise) p th power of the magnitude of the DFT coefficients ($X = |Y|^p$). Given the observation matrix X , there are different algorithms to perform the factorization [6, 7]. Here, we use generalized Kullback-Leibler divergence as the cost function:

$$D_{KL}(X||TV) = \sum_{k,\tau} (X_{k,\tau} \log \frac{X_{k,\tau}}{[TV]_{k,\tau}} + [TV]_{k,\tau} - X_{k,\tau}). \quad (1)$$

Factors T, V are found by iterating the following multiplicative rules [8] to minimize (1):

$$\begin{aligned} T_{k,i} &\leftarrow T_{k,i} \frac{\sum_\tau V_{i,\tau} (X_{k,\tau} / [TV]_{k,\tau})}{\sum_p V_{i,p}}, \\ V_{i,\tau} &\leftarrow V_{i,\tau} \frac{\sum_k T_{k,i} (X_{k,\tau} / [TV]_{k,\tau})}{\sum_q T_{q,i}}. \end{aligned} \quad (2)$$

After updating T , the columns of T are normalized such that each column sums to 1.

3 Noise PSD estimation Using NMF

In this section, we show how NMF can be used to estimate the noise PSD. The obtained noise PSD will be used in a Wiener filtering framework to enhance the noisy speech. NMF based algorithms consist of training and enhancement phases. For both steps, the given time-domain signal is segmented, windowed, and transformed into the frequency domain to obtain the spectrogram. During the training phase, NMF is applied to the observations from the clean speech and noise signals ($|S_{train}|^p$ and $|N_{train}|^p$) to obtain the speech basis matrix, T_S , and noise basis matrix, T_N :

$$(T_S, V) = \underset{T, Z}{\operatorname{argmin}} D_{KL}(|S_{train}|^p \|TZ), \quad (3)$$

$$(T_N, W) = \underset{T, Z}{\operatorname{argmin}} D_{KL}(|N_{train}|^p \|TZ), \quad (4)$$

where S_{train} and N_{train} are the DFT coefficients of the clean speech and noise signals, respectively. Now, the basis matrix for the observed noisy speech, T , is obtained by concatenating T_S and T_N as: $T = (T_S \ T_N)$. In the enhancement phase, an overlap-add framework is utilized to process the noisy speech. Given the vector of the observation at time frame τ (p th power of the magnitude of the DFT coefficients of the τ th frame of the noisy speech), $|\mathbf{y}_\tau|^p$, NMF is applied to find a linear approximation of $|\mathbf{y}_\tau|^p$ as: $|\mathbf{y}_\tau|^p \approx T\mathbf{u}_\tau$. In other words, keeping the basis matrix T fixed, NMF is performed to obtain the NMF coefficients vector \mathbf{u}_τ :

$$\mathbf{u}_\tau = \underset{\mathbf{z}}{\operatorname{argmin}} D_{KL}(|\mathbf{y}_\tau|^p \|T\mathbf{z}). \quad (5)$$

Partitioning \mathbf{u}_τ as: $\mathbf{u}_\tau = (\mathbf{v}_\tau^\top \ \mathbf{w}_\tau^\top)^\top$ (\top denotes the transpose), the clean speech component is approximated using $|\mathbf{s}_\tau|^p \approx T_S \mathbf{v}_\tau$, and the noise component is approximated as $|\mathbf{n}_\tau|^p \approx T_N \mathbf{w}_\tau$. An instantaneous estimate of the noise PSD is now obtained as:

$$|\widehat{N}_{k,\tau}|^2 = \left(\frac{[T_N \mathbf{w}_\tau]_k}{[T_S \mathbf{v}_\tau + T_N \mathbf{w}_\tau]_k} \times |Y_{k,\tau}|^p \right)^{2/p}. \quad (6)$$

Assuming some extent of stationarity of the noise, we can smooth this instantaneous estimate across time to get a better noise PSD estimate:

$$\left[\widehat{\sigma_N^2} \right]_{k,\tau} = \alpha \left[\widehat{\sigma_N^2} \right]_{k,\tau-1} + (1 - \alpha) |\widehat{N}_{k,\tau}|^2, \quad (7)$$

where $\left[\widehat{\sigma_N^2} \right]_{k,\tau}$ denotes the estimated noise PSD for frequency bin k and time-frame τ .

4 Linear MMSE Filter Based on NMF

In this section, we derive a new filter for single channel speech enhancement by minimizing the mean square error between the clean speech and the estimated speech components. We use $|\mathbf{y}_\tau|^p \approx T\mathbf{u}_\tau = T_S\mathbf{v}_\tau + T_N\mathbf{w}_\tau \approx |\mathbf{s}_\tau|^p + |\mathbf{n}_\tau|^p$, in which $T_S\mathbf{v}_\tau$ and $T_N\mathbf{w}_\tau$ are some random variables whose specific realizations are to be estimated; Given the observation $|\mathbf{y}_\tau|^p$, we can find the magnitude of the DFT coefficients of the enhanced speech as $|\widehat{S}_{k,\tau}| = \left(|\widehat{S}_{k,\tau}|^p\right)^{1/p}$ where $|\widehat{S}_{k,\tau}|^p = H_{k,\tau} |Y_{k,\tau}|^p$ is the linear MMSE estimate of the speech component. Assuming that p -th powers of the magnitude of the DFT coefficients at different frequencies are independent, we can minimize the mean square error

$$E\left(\left(|S_{k,\tau}|^p - H_{k,\tau} |Y_{k,\tau}|^p\right)^2\right) \approx E\left(\left([T_S\mathbf{v}_\tau]_k - H_{k,\tau} [T\mathbf{u}_\tau]_k\right)^2\right) \quad (8)$$

independently for each frequency bin k . H can be obtained by taking the derivative of (8) and making it equal to zero [9, Sec 11.3.1]:

$$0 = \frac{\partial E\left(\left([T_S\mathbf{v}_\tau]_k - H_{k,\tau} [T\mathbf{u}_\tau]_k\right)^2\right)}{\partial H_{k,\tau}} = E\left(-2 [T_S\mathbf{v}_\tau]_k [T\mathbf{u}_\tau]_k + 2H_{k,\tau} [T\mathbf{u}_\tau]_k^2\right) \quad (9)$$

and hence:

$$H_{k,\tau} = \frac{E([T_S\mathbf{v}_\tau]_k [T\mathbf{u}_\tau]_k)}{E([T\mathbf{u}_\tau]_k^2)}. \quad (10)$$

Assuming independency between the speech and noise components we get:

$$H_{k,\tau} = \frac{E([T_S\mathbf{v}_\tau]_k^2) + E([T_S\mathbf{v}_\tau]_k)E([T_N\mathbf{w}_\tau]_k)}{E([T_S\mathbf{v}_\tau]_k^2) + E([T_N\mathbf{w}_\tau]_k^2) + 2E([T_S\mathbf{v}_\tau]_k)E([T_N\mathbf{w}_\tau]_k)}. \quad (11)$$

Eq. (11) can be converted into a simpler form by assuming that the real and imaginary parts of the DFT coefficients of the speech (S) and noise (N) signals are zero mean normally distributed random variables; recalling that $|\mathbf{s}_\tau|^p \approx T_S\mathbf{v}_\tau$ (and $|\mathbf{n}_\tau|^p \approx T_N\mathbf{w}_\tau$), the relation between $E\left([T_S\mathbf{v}_\tau]_k^2\right)$ and $E([T_S\mathbf{v}_\tau]_k)$ (also $E\left([T_N\mathbf{w}_\tau]_k^2\right)$ and $E([T_N\mathbf{w}_\tau]_k)$) can be found simply for $p = 1, 2$, i.e.:

$$E([T_S\mathbf{v}_\tau]_k) \approx c\sqrt{E\left([T_S\mathbf{v}_\tau]_k^2\right)}, \quad (12)$$

where $c = \sqrt{\pi}/2$ for $p = 1$, and $c = \sqrt{2}/2$ for $p = 2$.

We can now continue to simplify equation (11). Dividing the denominator and numerator of (11) by $E\left([T_N\mathbf{w}_\tau]_k^2\right)$ and defining $\xi_{k,\tau} = \frac{E([T_S\mathbf{v}_\tau]_k^2)}{E([T_N\mathbf{w}_\tau]_k^2)}$,

and using (12) we get:

$$H_{k,\tau} \approx \frac{\xi_{k,\tau} + c^2 \sqrt{\xi_{k,\tau}}}{\xi_{k,\tau} + 1 + 2c^2 \sqrt{\xi_{k,\tau}}}, \quad (13)$$

in which we used:

$$\begin{aligned} \frac{E([T_S \mathbf{v}_\tau]_k) E([T_N \mathbf{w}_\tau]_k)}{E([T_N \mathbf{w}_\tau]_k^2)} &\approx c^2 \frac{\sqrt{E([T_S \mathbf{v}_\tau]_k^2)} \sqrt{E([T_N \mathbf{w}_\tau]_k^2)}}{E([T_N \mathbf{w}_\tau]_k^2)} \\ &= c^2 \sqrt{\xi_{k,\tau}}. \end{aligned} \quad (14)$$

$\xi_{k,\tau}$ represents the smoothed speech to noise ratio (*smoothed* SpNR). $E([T_N \mathbf{w}_\tau]_k^2)$ can be estimated by a low pass filter as:

$$E([T_N \mathbf{w}_\tau]_k^2) \approx \beta E([T_N \mathbf{w}_{\tau-1}]_k^2) + (1 - \beta) [T_N \mathbf{w}_\tau]_k^2. \quad (15)$$

$\xi_{k,\tau}$ can be found by following approximation: Define an *approximate* SpNR as $\eta_{k,\tau} = \frac{[T_S \mathbf{v}_\tau]_k^2}{E([T_N \mathbf{w}_\tau]_k^2)}$, and hence $\xi_{k,\tau} = E(\eta_{k,\tau})$; we propose a decision-directed estimator, similar to [10], for $\xi_{k,\tau}$ as¹:

$$\xi_{k,\tau} = \max \left(\xi_{min}, \gamma \frac{|\widehat{S}_{k,\tau-1}|^{2p}}{E([T_N \mathbf{w}_{\tau-1}]_k^2)} + (1 - \gamma) \eta_{k,\tau} \right). \quad (16)$$

In our simulations, fairly similar results were obtained using the following approximation of (13):

$$H_{k,\tau} \approx \frac{\xi_{k,\tau}}{\xi_{k,\tau} + 1}. \quad (17)$$

It is interesting to highlight the differences between (17) and the Wiener filter: Assuming the magnitude of the DFT coefficients of the noisy speech as the observation ($p = 1$) for NMF, and perfect nonnegative factorization for the clean speech and noise signals as $|\mathbf{s}_\tau| = T_S \mathbf{v}_\tau$, $|\mathbf{n}_\tau| = T_N \mathbf{w}_\tau$, (17) will be identical to the Wiener filter; however, by using the magnitude-squared DFT coefficients ($p = 2$) this is not true any more. Moreover, there is another implementation difference between the two filters: the Wiener filter is often implemented by estimating *a priori* SNR based on a *posteriori* SNR which is obtained from the noisy observation [11]; though, for implementing (11), as it is mentioned above, an *approximate* SpNR can be estimated using the separated speech and noise components from NMF; next, the *smoothed* SpNR is estimated using (16) and is used to implement (13) or (17). Since the *approximate* SpNR is based on an initial estimate of the speech component and not the noisy speech, the smoothing factor in (16)

¹In Eq. (12) of the IEEE published version of this paper, a “2” is missing in the power of $|\widehat{S}_{k,\tau-1}|$, which is corrected here in (16).

should be low enough to capture the speech variations quickly. Good results were obtained for $\gamma = 0.5 - 0.75$ while the results were not sensitive to the exact value of γ . Finally, note that the defined SpNR is not the same as the SNR which is usually defined as the ratios of the powers of the speech and noise signals.

The algorithm is summarized as:

1. Obtain the NMF coefficients vector \mathbf{u}_τ by applying NMF to the given observation at time frame τ as $|\mathbf{y}_\tau|^p \approx T\mathbf{u}_\tau$.
2. Find \mathbf{w}_τ from $\mathbf{u}_\tau = (\mathbf{v}_\tau^\top \mathbf{w}_\tau^\top)^\top$, then obtain an estimate of $E\left([T_N \mathbf{w}_\tau]_k^2\right)$ by smoothing $[T_N \mathbf{w}_\tau]_k^2$ over time (Eq. (15)).
3. Obtain the *approximate* SpNR, $\eta_{k,\tau}$, and *smoothed* SpNR, $\xi_{k,\tau}$ for all frequency bins (Eq. (16)).
4. Obtain the filter gain as (13) or (17).
5. The magnitude of the DFT coefficients of the enhanced speech are obtained as $|\widehat{S}_{k,\tau}| = (H_{k,\tau} |Y_{k,\tau}|^p)^{1/p}$.
6. Reconstruct the time domain signal using the noisy phase.

5 Evaluation

Both magnitude ($p = 1$) and squared magnitude ($p = 2$) of the DFT coefficients of the observed noisy speech signal are used as observation in NMF model for the enhancement task. In the following, the derived algorithm in section 4 (using Eq. (13)) is referred as 'LMMSE-Mag' and 'LMMSE-Pow' for $p = 1$ and $p = 2$, respectively. The estimated noise PSDs from Section 3 were used in combination with a Wiener filter to perform the enhancement and are referred as 'Wiener-Mag' and 'Wiener-Pow' for $p = 1$ and $p = 2$, respectively; in addition, noise PSD was estimated using a MMSE-based approach [12] which is one of the best algorithms for this purpose [13], and the same Wiener filter was used for the enhancement; in the following, this approach is called 'Wiener-UnS' to reflect the fact that this approach is an unsupervised filtering and does not have any training. The Wiener filter was implemented using the decision-directed approach [10] with the same parameters $10 \log_{10}(\xi_{min}) = -25\text{dB}$ and $\alpha = 0.98$ for all the approaches. The same lower bound ξ_{min} also was used in (16).

We used speech from the Grid Corpus and noise from the NOISEX-92 databases. All the signals are down-sampled to 16 KHz. The speech is degraded by adding babble noise or factory noise at 3 different SNRs: 0 dB, 5 dB, and 10 dB. A separate model is trained for each noise type, and one

speaker independent model is trained for the speech signal; this model was trained on a mixed group of 24 male and female speakers, and 8 sentences from each speaker were used. For all the approaches 10 sentences from each of the 8 speakers (4 male and 4 female, and none of them were used for the training), and a part of the noise signal which was not used for the training, were used for the performance evaluation. The results are averaged over the entire test set. To apply NMF we use a noise specific basis matrix; if noise type is not known a priori, some adapting procedures have to be used which we have not included in our simulations. For the speech and noise signals, 60 and 100 basis vectors are trained, respectively. The following parameters are obtained by performing a cross-validation test and are used in the simulations: $\alpha = 0.95$ in (7), $\beta = 0.95$ in (15), and $\gamma = 0.6$ in (16). The time frames have a length of 512 samples with 50% overlap, and are windowed using a Hann window.

The performance of the speech enhancement algorithms are evaluated using PESQ [14], and the Source to Distortion Ratio (SDR) which is defined as:

$$\text{SDR} = 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{artifact}}\|^2}, \quad (18)$$

where s_{target} , e_{interf} and e_{artifact} are target time-domain speech signal, interference, and artifact error terms defined in [15], and $\|\cdot\|^2$ denotes the energy. In order to analyze the results more specifically, segmental speech SNR ($\text{SNR}_{\text{seg-sp}}$), and segmental noise reduction (SegNR) are also measured as [16]:

$$\text{SNR}_{\text{seg-sp}} = \frac{1}{T} \sum_{\tau=1}^T 10 \log_{10} \left(\frac{\sum_{i=1}^I s_{i+\tau I}^2}{\sum_{i=1}^I (s_{i+\tau I} - \tilde{s}_{i+\tau I})^2} \right), \quad (19)$$

$$\text{SegNR} = \frac{1}{T} \sum_{\tau=1}^T 10 \log_{10} \left(\frac{\sum_{i=1}^I n_{i+\tau I}^2}{\sum_{i=1}^I \tilde{n}_{i+\tau I}^2} \right), \quad (20)$$

where I denotes the length of the frame, and T the number of frames; These measures are obtained in a shadow filtering framework: the filter is computed from the noisy speech signal ($s + n$) and is used to obtain \tilde{s}, \tilde{n} . \tilde{s} is the output of the enhancement system when the clean speech, s , is the input to the filter; similarly, \tilde{n} is the output of the enhancement system when only the noise, n , is the input to the filter.

5.1 Results and Discussion

Figure 1 shows the improvement in PESQ and SDR for different algorithms. The results show that a NMF-based filter which is derived using the magnitude of the DFT coefficients ($p = 1$) of the noisy speech gives a better result compared to the same type of the NMF-based filter which is derived using the magnitude-squared DFT coefficients ($p = 2$). This is true for both the Wiener and the proposed LMMSE filters, and agrees with the previous

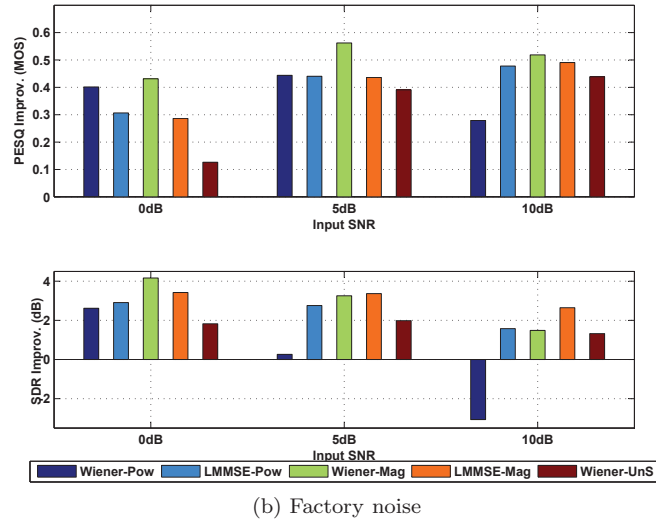
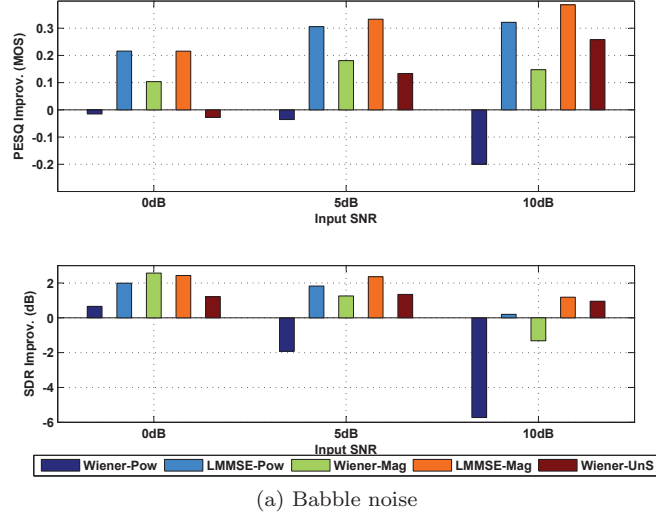


Figure 1: PESQ and SDR improvements for babble and factory noises.

applications of NMF in source separation (e.g. [1,2]). For the Wiener based algorithms, the difference between these two cases is much higher than that between the LMMSE algorithms. The NMF-based algorithms with $p = 1$ mostly result to a better performance than the Wiener-UnS algorithm, especially at low input SNRs. In the most cases the SDR improvement for

LMMSE-Mag is the highest among all the algorithms for both noise types. For the babble noise, the PESQ improvement is higher for the proposed LMMSE algorithms compared to the Wiener based algorithms. For the factory noise, the Wiener based algorithms often provide better PESQ for the enhanced speech signal than the LMMSE algorithms, especially for low input SNRs.

For input SNRs for which PESQ and SDR improvements are not pointing in the same direction (for instance factory noise at 5 dB input SNR) it becomes more difficult to compare different algorithms; hence, we performed an informal listening test and found that if the difference in the PESQ improvements is not high, and at the same time the difference in the SDR improvements is high, the algorithm with the higher SDR is preferred; for example, the LMMSE-Pow was preferred over the Wiener-Pow algorithm for factory noise at 5 dB input SNR. This is because LMMSE-Pow provides much higher SDR even though both methods provide similar PESQ scores for the enhanced speech. This can be expected since none of these measures completely model the speech quality. Even fairly similar scores were obtained for the LMMSE-Mag and Wiener-Mag for the factory noise at 5 dB input SNR. These results might be explained by looking at Figure 2.

Figure 2 shows the stacked results for Segmental Noise Reduction, SegNR, and Segmental Speech SNR, $\text{SNR}_{\text{seg-sp}}$, for factory noise which are shown in the bottom and top of the figure respectively. For both measures a high value is desired. $\text{SNR}_{\text{seg-sp}}$ is inversely proportional to the speech distortion. Wiener based approaches provide a higher SegNR and lower $\text{SNR}_{\text{seg-sp}}$ compared to the proposed LMMSE algorithms. The PESQ improvements for the Wiener based approaches are obtained mainly because of the high SegNR while for the proposed LMMSE filters the PESQ improvements are obtained mainly because of the high $\text{SNR}_{\text{seg-sp}}$ (and hence less speech distortion).

6 Conclusions

Two types of NMF-based algorithms were obtained in this paper: first, a Wiener filter was considered in which noise PSD was estimated using NMF. Second, a LMMSE filter was derived by minimizing the mean square error between the clean speech and the estimated speech components in the frequency domain. The proposed LMMSE filters were shown to be promising and gave better SDR improvements compared to the Wiener-based algorithms in most of the test cases; LMMSE filters gave a higher PESQ improvements for the babble noise for all the simulated input SNRs although for the factory noise it was not the case. Most of the NMF-based approaches gave better SDR and PESQ improvements compared to the Wiener filtering method in which a recently developed unsupervised

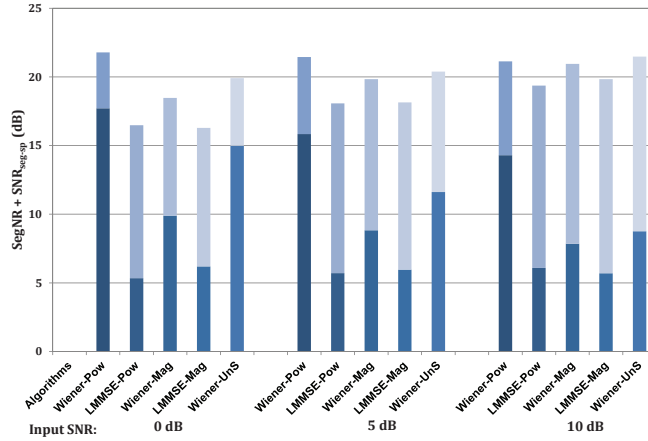


Figure 2: Stacked presentation of Segmental Noise Reduction (SegNR, bottom), and Segmental Speech SNR ($\text{SNR}_{\text{seg-sp}}$, top) for factory noise.

approach was used to estimate the noise PSD.

References

- [1] 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.
- [2] 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.
- [3] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, “Speech denoising using nonnegative matrix factorization with priors,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, 2008, pp. 4029–4032.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Process. Systems (NIPS)*. MIT Press, 2000, pp. 556–562.
- [9] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. West Sussex, England: John Wiley & Sons, 2006.
- [10] Y. Ephraim and I. Cohen, *Recent Advancements in Speech Enhancement*. in *The Electrical Engineering Handbook*, CRC Press, 2005.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.

Paper B

Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization

Nasser Mohammadiha, Paris Smaragdis, and Arne Leijon

Refereed article published in
IEEE Transactions on Audio, Speech and Language Processing, vol. 21,
no. 10, pp. 2140–2151, oct. 2013.

©2013 IEEE
Layout has been revised for thesis consistency

Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization

Nasser Mohammadiha, Paris Smaragdis, and Arne Leijon

Abstract

Reducing the interference noise in a monaural noisy speech signal has been a challenging task for many years. Compared to traditional unsupervised speech enhancement methods, e.g., Wiener filtering, supervised approaches, such as algorithms based on hidden Markov models (HMM), lead to higher-quality enhanced speech signals. However, the main practical difficulty of these approaches is that for each noise type a model is required to be trained a priori. In this paper, we investigate a new class of supervised speech denoising algorithms using nonnegative matrix factorization (NMF). We propose a novel speech enhancement method that is based on a Bayesian formulation of NMF (BNMF). To circumvent the mismatch problem between the training and testing stages, we propose two solutions. First, we use an HMM in combination with BNMF (BNMF-HMM) to derive a minimum mean square error (MMSE) estimator for the speech signal with no information about the underlying noise type. Second, we suggest a scheme to learn the required noise BNMF model online, which is then used to develop an unsupervised speech enhancement system. Extensive experiments are carried out to investigate the performance of the proposed methods under different conditions. Moreover, we compare the performance of the developed algorithms with state-of-the-art speech enhancement schemes using various objective measures. Our simulations show that the proposed BNMF-based methods outperform the competing algorithms substantially.

1 Introduction

Estimating the clean speech signal in a single-channel recording of a noisy speech signal has been a research topic for a long time and is of interest

for various applications including hearing aids, speech/speaker recognition, and speech communication over telephone and internet. A major outcome of these techniques is the improved quality and reduced listening effort in the presence of an interfering noise signal.

In general, speech enhancement methods can be categorized into two broad classes: unsupervised and supervised. Unsupervised methods include a wide range of approaches such as spectral subtraction [1], Wiener and Kalman filtering, e.g., [2, 3], short-time spectral amplitude (STSA) estimators [4], estimators based on super-Gaussian prior distributions for speech DFT coefficients [5–8], and schemes based on periodic models of the speech signal [9]. In these methods, a statistical model is assumed for the speech and noise signals, and the clean speech is estimated from the noisy observation without any prior information on the noise type or speaker identity. However, the main difficulty of most of these methods is estimation of the noise power spectral density (PSD) [10–12], which is a challenging task if the background noise is non-stationary.

For the supervised methods, a model is considered for both the speech and noise signals and the model parameters are estimated 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. Some examples of this class of algorithms include the codebook-based approaches, e.g., [13, 14] and hidden Markov model (HMM) based methods [15–19]. One advantage of these methods is that there is no need to estimate the noise PSD using a separate algorithm.

The supervised approaches have been shown to produce better quality enhanced speech signals compared to the unsupervised methods [14, 16], which can be expected as more prior 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 [14, 16], or can be provided by a separate acoustic environment classification algorithm [20]. The primary goal of this work is to propose supervised and unsupervised speech enhancement algorithms based on nonnegative matrix factorization (NMF) [21, 22].

NMF is a technique to project a nonnegative matrix \mathbf{y} onto a space spanned by a linear combination of a set of basis vectors, i.e., $\mathbf{y} \approx \mathbf{b}\mathbf{v}$, where both \mathbf{b} and \mathbf{v} are nonnegative matrices. In speech processing, \mathbf{y} is usually the spectrogram of the speech signal with spectral vectors stored by column, \mathbf{b} is the basis matrix or dictionary, and \mathbf{v} is referred to as the NMF coefficient or activation matrix. NMF has been widely used as a source separation technique applied to monaural mixtures, e.g., [23–25]. More recently, NMF has also been used to estimate the clean speech from a noisy observation [26–31].

When applied to speech source separation, a good separation can be ex-

pected only when speaker-dependent basis 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 [29, 31]. Nevertheless, there might be some scenarios (such as speech degraded with multitalker babble noise) for which the basis matrices of speech and noise are quite similar. In these cases, although the traditional NMF-based approaches can be used to get state-of-the-art performance, other constraints can be imposed into NMF to obtain a better noise reduction. For instance, assuming that the babble waveform is obtained as a sum of different speech signals, a nonnegative hidden Markov model is proposed in [26] to model the babble noise in which the babble basis is identical to the speech basis. Another fundamental issue in basic NMF is that it ignores the important temporal dependencies of the audio signals. Different approaches have been proposed in the literature to employ temporal dynamics in NMF, e.g., [23–25, 27, 30, 31].

In this paper, we first propose a new supervised NMF-based speech enhancement system. In the proposed method, the temporal dependencies of speech and noise signals are used to construct informative prior distributions that are applied in a Bayesian framework to perform NMF (BNMF). We then develop an HMM structure with output density functions given by BNMF to simultaneously classify the environmental noise and enhance the noisy signal. Therefore, the noise type does not need to be specified a priori. Here, the classification is done using the noisy input and is not restricted to be applied at only the speech pauses as it is in [16], and it does not require any additional noise PSD tracking algorithm, as it is required in [14].

Moreover, we propose an unsupervised NMF-based approach in which the noise basis matrix is learned online from the noisy mixture. Although online dictionary learning from clean data has been addressed in some prior works, e.g., [32, 33], our causal method learns the noise basis matrix from the noisy mixture. The main contributions of this work can be summarized as:

1. We present a review of state-of-the-art NMF-based noise reduction approaches.
2. We propose a speech enhancement method based on BNMF that inherently captures the temporal dependencies in the form of hierarchical prior distributions. Some preliminary results of this approach has been presented in [31]. Here, we further develop the method and evaluate its performance comprehensively. In particular, we present an approach to construct SNR-dependent prior distributions.
3. An environmental noise classification technique is suggested and is combined with the above BNMF approach (BNMF-HMM) to develop an unsupervised speech enhancement system.

Table 1: The table summarizes some of the notations that are consistently used in the paper.

k	frequency index
t	time index
X	a scalar random variable
$\mathbf{Y} = [Y_{kt}]$	a matrix of random variables
\mathbf{Y}_t	t -th column of \mathbf{Y}
$\mathbf{y} = [y_{kt}]$	a matrix of observed magnitude spectrogram
\mathbf{y}_t	t -th column of \mathbf{y}
$\mathbf{b}^{(s)}$	speech parameters ($\mathbf{b}^{(s)}$ is the speech basis matrix)
$\mathbf{b}^{(n)}$	noise parameters ($\mathbf{b}^{(n)}$ is the noise basis matrix)
$\mathbf{b} = [\mathbf{b}^{(s)} \mathbf{b}^{(n)}]$	mixture parameters (\mathbf{b} is the mixture basis matrix)

4. A causal online dictionary learning scheme is proposed that learns the noise basis matrix from the noisy observation. Our simulations show that the final unsupervised noise reduction system outperforms state-of-the-art approaches significantly.

The rest of the paper is organized as follows: The review of the NMF-based speech enhancement algorithms is presented in Section 2. In Section 3, we describe our main contributions, namely the BNMF-based noise reduction, BNMF-HMM structure, and online noise dictionary learning. Section 4 presents our experiments and results with supervised and unsupervised noise reduction systems. Finally, Section 5 concludes the study.

2 Review of State-of-the-art NMF-Based Speech Enhancement

In this section, we first explain a basic NMF approach, and then we review NMF-based speech enhancement. Let us represent the random variables associated with the magnitude of the discrete Fourier transform (DFT) coefficients of the speech, noise, and noisy signals as $\mathbf{S} = [S_{kt}]$, $\mathbf{N} = [N_{kt}]$ and $\mathbf{Y} = [Y_{kt}]$, respectively, where k and t denote the frequency and time indices, respectively. The actual realizations are shown in small letters, e.g., $\mathbf{y} = [y_{kt}]$. Table 1 summarizes some of the notations that are frequently used in the paper. To obtain a nonnegative decomposition of a given matrix \mathbf{x} , a cost function is usually defined and is minimized. Let us denote the basis matrix and NMF coefficient matrix by \mathbf{b} and \mathbf{v} , respectively. Nonnegative

factorization is achieved by solving the following optimization problem:

$$(\mathbf{b}, \mathbf{v}) = \underset{\mathbf{b}, \mathbf{v}}{\operatorname{argmin}} D(\mathbf{y} \|\mathbf{b}\mathbf{v}) + \mu h(\mathbf{b}, \mathbf{v}), \quad (1)$$

where $D(\mathbf{y} \|\hat{\mathbf{y}})$ is a cost function, $h(\cdot)$ is an optional regularization term, and μ is the regularization weight. The minimization in (1) is performed under the nonnegativity constraint of \mathbf{b} and \mathbf{v} . The common choices for the cost function include Euclidean distance [21], generalized Kullback-Leibler divergence [21,34], Itakura-Saito divergence [25], and the negative likelihood of data in the probabilistic NMFs [35]. Depending on the application, the sparsity of the activations \mathbf{v} and the temporal dependencies of input data \mathbf{x} are two popular motivations to design the regularization function, e.g., [24,27,36,37]. Since (1) is not a convex problem, iterative gradient descent or expectation-maximization (EM) algorithms are usually followed to obtain a locally optimal solution for the problem [21,25,35].

Let us consider a supervised denoising approach where the basis matrix of speech $\mathbf{b}^{(s)}$ and the basis matrix of noise $\mathbf{b}^{(n)}$ are learned using the appropriate training data in advance. The common assumption used to model the noisy speech signal is the additivity of speech and noise spectrograms, i.e., $\mathbf{y} = \mathbf{s} + \mathbf{n}$. Although in the real world problems this assumption is not justified completely, the developed algorithms have been shown to produce satisfactory results, e.g., [24]. The basis matrix of the noisy signal is obtained by concatenating the speech and noise basis matrices as $\mathbf{b} = [\mathbf{b}^{(s)} \mathbf{b}^{(n)}]$. Given the magnitude of DFT coefficients of the noisy speech at time t , \mathbf{y}_t , the problem in (1) is now solved—with \mathbf{b} held fixed—to obtain the noisy NMF coefficients \mathbf{v}_t . The NMF decomposition takes the form $\mathbf{y}_t \approx \mathbf{b}\mathbf{v}_t = [\mathbf{b}^{(s)} \mathbf{b}^{(n)}][(\mathbf{v}_t^{(s)})^\top (\mathbf{v}_t^{(n)})^\top]^\top$, where \top denotes transposition. Finally, an estimate of the clean speech DFT magnitudes is obtained by a Wiener-type filtering as:

$$\hat{\mathbf{s}}_t = \frac{\mathbf{b}^{(s)}\mathbf{v}_t^{(s)}}{\mathbf{b}^{(s)}\mathbf{v}_t^{(s)} + \mathbf{b}^{(n)}\mathbf{v}_t^{(n)}} \odot \mathbf{y}_t, \quad (2)$$

where the division is performed element-wise, and \odot denotes an element-wise multiplication. The clean speech waveform is estimated using the noisy phase and inverse DFT. One advantage of the NMF-based approaches over the HMM-based [16,17] or codebook-driven [14] approaches is that NMF automatically captures the long-term levels of the signals, and no additional gain modeling is necessary.

Schmidt *et al.* [28] presented an NMF-based unsupervised batch algorithm for noise reduction. In this approach, it is assumed that the entire noisy signal is observed, and 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 (2). 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 [28] is applicable only in the batch mode, which is usually not practical in the real world.

In [27], 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 $\hat{\mathbf{s}}_t = \mathbf{b}^{(s)} \mathbf{v}_t^{(s)}$. Although it is not explicitly mentioned in [27], 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.

In [29], authors propose a linear minimum mean square error (MMSE) estimator for NMF-based speech enhancement. In this work, NMF is applied to \mathbf{y}_t^p (i.e., $\mathbf{y}_t^p = \mathbf{b} \mathbf{v}_t$, where $p = 1$ corresponds to using magnitude of DFT coefficients and $p = 2$ corresponds to using magnitude-squared DFT coefficients) in a frame by frame routine. Then, a gain variable \mathbf{g}_t is estimated to filter the noisy signal as: $\hat{\mathbf{s}}_t = (\mathbf{g}_t \odot \mathbf{y}_t^p)^{1/p}$. Assuming that the basis matrices of speech and noise are obtained during the training stage, and that the NMF coefficients \mathbf{V}_t are random variables, \mathbf{g}_t is derived such that the mean square error between \mathbf{S}_t^p and $\widehat{\mathbf{S}}_t^p$ is minimized. The optimal gain is shown to be:

$$\mathbf{g}_t = \frac{\boldsymbol{\xi}_t + c^2 \sqrt{\boldsymbol{\xi}_t}}{\boldsymbol{\xi}_t + 1 + 2c^2 \sqrt{\boldsymbol{\xi}_t}}, \quad (3)$$

where c is a constant that depends on p [29] and $\boldsymbol{\xi}_t$ is called the smoothed speech to noise ratio that is estimated using a decision-directed approach. For a theoretical comparison of (3) to a usual Wiener filter see [29]. The conducted simulations show that the results using $p = 1$ are superior to those using $p = 2$ (which is in line with previously reported observations, e.g., [24]) and that both of them are better than the results of a state-of-the-art Wiener filter.

A semi-supervised approach is proposed in [30] to denoise a noisy signal using NMF. In this method, a nonnegative hidden Markov model (NHMM) is used to model the speech magnitude spectrogram. Here, the HMM state-dependent output density functions are assumed to be a mixture of multi-

nomial distributions, and thus, the model is closely related to probabilistic latent component analysis (PLCA) [35]. 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 then used to construct a factorial HMM. When applied for noise reduction, first 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 also 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 [28]. 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 [30] suffers from high computational complexity.

A linear nonnegative dynamical system is presented in [38] to model temporal dependencies in NMF. The proposed causal filtering and fixed-lag smoothing algorithms use Kalman-like prediction in NMF and PLCA. Compared to the ad-hoc methods that use temporal correlations to design regularity functions, e.g., [27, 37], this approach suggests a solid framework to incorporate temporal dynamics into the system. Also, the computational complexity of this method is significantly less than [30].

Raj *et al.* [39] proposed a phoneme-dependent approach to use NMF for speech enhancement in which a set of basis vectors are 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 using a set of speaker-dependent basis matrices the estimation of the clean speech is carried out.

NMF-based noise PSD estimation is addressed in [37]. In this work, the speech and noise basis matrices are trained offline, after which a constrained NMF is applied to the noisy spectrogram in a frame by frame basis. To utilize the time dependencies of the speech and noise signals, an l_2 -norm regularization term is added to the cost function. This penalty term encourages consecutive speech and noise NMF coefficients to take similar values, and hence, to model the signals' time dependencies. The instantaneous noise periodogram is obtained similarly to (2) by switching the role of speech and noise approximates. This estimate is then smoothed over time using an exponential smoothing to get a less-fluctuating estimate of the noise PSD, which can be combined with any algorithm that needs a noise PSD, e.g., Wiener filter.

3 Speech Enhancement Using Bayesian NMF

In this section, we present our Bayesian NMF (BNMF) based speech enhancement methods. In the following, an overview of the employed BNMF is provided first, which was originally proposed in [34]. Our proposed extensions of this BNMF to modeling a noisy signal, namely BNMF-HMM and Online-BNMF are given in Subsections 3.1 and 3.2, respectively. Subsection 3.3 presents a method to construct informative priors to use temporal dynamics in NMF.

The probabilistic NMF in [34] assumes that an input matrix is stochastic, and to perform NMF as $\mathbf{y} \approx \mathbf{b}\mathbf{v}$ the following model is considered:

$$Y_{kt} = \sum_i Z_{kit}, \quad (4)$$

$$\begin{aligned} f_{Z_{kit}}(z_{kit}) &= \mathcal{PO}(z_{kit}; b_{ki}v_{it}) \\ &= (b_{ki}v_{it})^{z_{kit}} e^{-b_{ki}v_{it}} / (z_{kit}!), \end{aligned} \quad (5)$$

where Z_{kit} are latent variables, $\mathcal{PO}(z; \lambda)$ denotes the Poisson distribution, and $z!$ is the factorial of z . A schematic representation of this model is shown in Figure 1.

As a result of (4) and (5), Y_{kt} is assumed Poisson-distributed and integer-valued. In practice, the observed spectrogram is first scaled up and then rounded to the closest integer numbers to avoid large quantization errors. The maximum likelihood (ML) estimate of the parameters \mathbf{b} and \mathbf{v} can be obtained using an EM algorithm [34], and the result would be identical to the well-known multiplicative update rules for NMF using Kullback-Leibler (KL-NMF) divergence [21].

In the Bayesian formulation, the nonnegative factors are further assumed to be random variables. In this hierarchical model, gamma prior distributions are considered to govern the basis (\mathbf{B}) and NMF coefficient (\mathbf{V}) matrices:

$$\begin{aligned} f_{V_{it}}(v_{it}) &= \mathcal{G}(v_{it}; \phi_{it}, \theta_{it}/\phi_{it}), \\ f_{B_{ki}}(b_{ki}) &= \mathcal{G}(b_{ki}; \psi_{ki}, \gamma_{ki}/\psi_{ki}), \end{aligned} \quad (6)$$

in which $\mathcal{G}(v; \phi, \theta) = \exp((\phi - 1) \log v - v/\theta - \log \Gamma(\phi) - \phi \log \theta)$ denotes the gamma density function with ϕ as the shape parameter and θ as the scale parameter, and $\Gamma(\phi)$ is the gamma function. ϕ, θ, ψ , and γ are referred to as the hyperparameters.

As the exact Bayesian inference for (4), (5), and (6) is difficult, a variational Bayes approach has been proposed in [34] to obtain the approximate posterior distributions of \mathbf{B} and \mathbf{V} . 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.

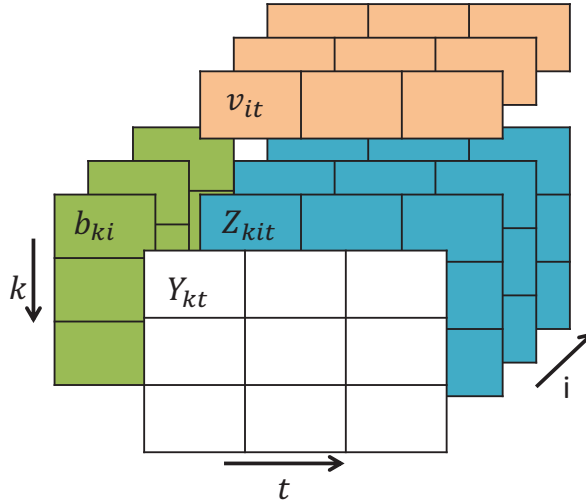


Figure 1: A schematic representation of (4) and (5) [34]. Each time-frequency bin of a magnitude spectrogram (Y_{kt}) is assumed to be a sum of some Poisson-distributed hidden random variables (Z_{kit}).

More specifically for this Bayesian NMF, in an iterative scheme, the current estimates of the posterior distributions of \mathbf{Z} are used to update the posterior distributions of \mathbf{B} and \mathbf{V} , and these new posteriors are used to update the posteriors of \mathbf{Z} in the next iteration. The iterations are carried on until convergence. The posterior distributions for $Z_{k,:,t}$ are shown to be multinomial density functions ($:$ denotes 'all the indices'), while for B_{ki} and V_{it} they are gamma density functions. Full details of the update rules can be found in [34]. This variational approach is much faster than an alternative Gibbs sampler, and its computational complexity can be comparable to that of the ML estimate of the parameters (KL-NMF).

3.1 BNMF-HMM for Simultaneous Noise Classification and Reduction

In the following, we describe the proposed BNMF-HMM noise reduction scheme in which the state-dependent output density functions are instances of the BNMF explained in the introductory part of this section. Each state of the HMM corresponds to one specific noise type. Let us consider a set of noise types for which we are able to gather some training data, and let us denote the cardinality of the set by M . We can train a BNMF model for each of these noise types given its training data. Moreover, we consider

a universal BNMF model for speech that can be trained a priori. Note that the considered speech model does not introduce any limitation in the method since we train a model for the speech signal in general, and we do not use any assumption on the identity or gender of the speakers.

The structure of the BNMF-HMM is shown in Figure 2. Each state of the HMM has some state-dependent parameters, which are the noise BNMF model parameters. Also, all the states share some state-independent parameters, which consist of the speech BNMF model and an estimate of the long-term signal to noise ratio (SNR) that will be used for the enhancement. To complete the Markovian model, we need to predefine an empirical state transition matrix (whose dimension is $M \times M$) and an initial state probability vector. For this purpose, we assign some high values to the diagonal elements of the transition matrix, and we set the rest of its elements to some small values such that each row of the transition matrix sums to one. Each element of the initial state probability vector is also set to $1/M$.

We model the magnitude spectrogram of the clean speech and noise signals by (4). To obtain a BNMF model, we need to find the posterior distribution of the basis matrix, and optimize for the hyperparameters if desired. During training, we assign some sparse and broad prior distributions to \mathbf{B} and \mathbf{V} according to (6). For this purpose, ψ and γ are chosen such that the mean of the prior distribution for \mathbf{B} is small, and its variance is very high. On the other hand, ϕ and θ are chosen such that the prior distribution of \mathbf{V} has a mean corresponding to the scale of the data and has a high variance to represent uncertainty. To have good initializations for the posterior means, the multiplicative update rules for KL-NMF are applied first for a few iterations, and the result is used as the initial values for the posterior means. After the initialization, variational Bayes (as explained before) is run until convergence. We also optimize the hyperparameters using Newton's method, as proposed in [34].

In the following, the speech and noise random basis matrices are denoted by $\mathbf{B}^{(s)}$ and $\mathbf{B}^{(n)}$, respectively. A similar notation is used to distinguish all the speech and noise parameters.

Let us denote the hidden state variable at each time frame t by X_t , which can take one of the M possible outcomes $x_t = 1, 2, \dots, M$. The noisy magnitude spectrogram, given the state X_t , is modeled using (4). Here, we use the additivity assumption to approximate the state-dependent distribution of the noisy signal, i.e., $\mathbf{y}_t = \mathbf{s}_t + \mathbf{n}_t$. To obtain the distribution of the noisy signal, given the state X_t , the parameters of the speech and noise basis matrices ($\mathbf{B}^{(s)}$ and $\mathbf{B}^{(n)}$) are concatenated to obtain the parameters of the noisy basis matrix \mathbf{B} . Since the sum of independent Poisson random variables is Poisson, (4) leads to:

$$f_{Y_{kt}}(y_{kt} | x_t, \mathbf{b}, \mathbf{v}_t) = \frac{\lambda_{kt}^{y_{kt}} e^{-\lambda_{kt}}}{y_{kt}!}, \quad (7)$$

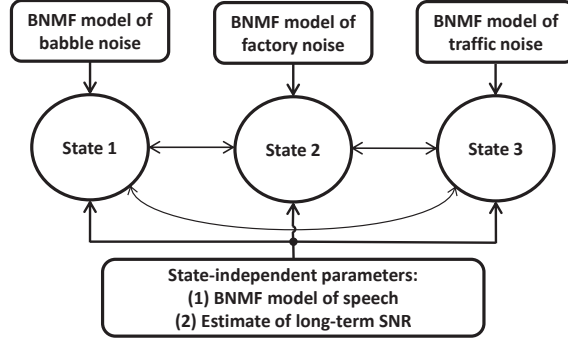


Figure 2: A block diagram representation of BNMF-HMM with three states.

where $\lambda_{kt} = \sum_i b_{ki} v_{it}$. Note that although the basis matrix \mathbf{b} is state-dependent, to keep the notations uncluttered, we skip writing this dependency explicitly.

The state-conditional likelihood of the noisy signal can now be computed by integrating over \mathbf{B} and \mathbf{V}_t as:

$$\begin{aligned} f_{Y_{kt}}(y_{kt} | x_t) &= \int \int f_{Y_{kt}, \mathbf{B}, \mathbf{V}_t}(y_{kt}, \mathbf{b}, \mathbf{v}_t | x_t) d\mathbf{b} d\mathbf{v}_t \\ &= \int \int f_{Y_{kt}}(y_{kt} | \mathbf{b}, \mathbf{v}_t, x_t) \\ &\quad f_{\mathbf{B}, \mathbf{V}_t}(\mathbf{b}, \mathbf{v}_t | x_t) d\mathbf{b} d\mathbf{v}_t. \end{aligned} \quad (8)$$

The distribution of \mathbf{y}_t is obtained by assuming that different frequency bins are independent [5, 7]:

$$f_{\mathbf{Y}_t}(\mathbf{y}_t | x_t) = \prod_k f_{Y_{kt}}(y_{kt} | x_t). \quad (9)$$

As the first step of the enhancement, variational Bayes approach is applied to approximate the posterior distributions of the NMF coefficient vector \mathbf{V}_t by maximizing the variational lower bound on (9). Here, we assume that the state-dependent posterior distributions of \mathbf{B} are time-invariant and are identical to those obtained during the training. Moreover, we use the temporal dynamics of noise and speech to construct informative prior distributions for \mathbf{V}_t , which is explained in Subsection 3.3. After convergence of the variational learning, we will have the parameters (including expected values) of the posterior distributions of \mathbf{V}_t as well as the latent variables \mathbf{Z}_t .

The MMSE estimate [40] of the speech DFT magnitudes can be shown to be [15, 26]:

$$\hat{s}_{kt} = E(S_{kt} | \mathbf{y}_t) = \frac{\sum_{x_t=1}^M \xi_t(\mathbf{y}_t, x_t) E(S_{kt} | x_t, \mathbf{y}_t)}{\sum_{x_t=1}^M \xi_t(\mathbf{y}_t, x_t)}, \quad (10)$$

where

$$\begin{aligned} \xi_t(\mathbf{y}_t, x_t) &= f_{\mathbf{Y}_t, X_t}(\mathbf{y}_t, x_t | \mathbf{y}_1^{t-1}) \\ &= f_{\mathbf{Y}_t}(\mathbf{y}_t | x_t) f_{X_t}(x_t | \mathbf{y}_1^{t-1}), \end{aligned} \quad (11)$$

in which $\mathbf{y}_1^{t-1} = \{\mathbf{y}_1, \dots, \mathbf{y}_{t-1}\}$. Here, $f_{X_t}(x_t | \mathbf{y}_1^{t-1})$ is computed using the forward algorithm [41]. Since (8) can not be evaluated analytically, one can either use numerical methods or use approximations to calculate $f_{Y_{kt}}(y_{kt} | x_t)$. Instead of expensive stochastic integrations, we approximate (8) by evaluating the integral at the mean value of the posterior distributions of \mathbf{B} and \mathbf{V}_t :

$$f_{Y_{kt}}(y_{kt} | x_t) \approx f_{Y_{kt}}(y_{kt} | \mathbf{b}', \mathbf{v}'_t, x_t), \quad (12)$$

where $\mathbf{b}' = E(\mathbf{B} | \mathbf{y}_t, x_t)$, and $\mathbf{v}'_t = E(\mathbf{V}_t | \mathbf{y}_t, x_t)$ are the posterior means of the basis matrix and NMF coefficient vector that are obtained using variational Bayes. Other types of point approximations have also been used for gain modeling in the context of HMM-based speech enhancement [17, 18].

To finish our derivation, we need to calculate the state-dependent MMSE estimate of the speech DFT magnitudes $E(S_{kt} | x_t, \mathbf{y}_t)$. First, let us rewrite (4) for the noisy signal as:

$$Y_{kt} = S_{kt} + N_{kt} = \sum_{i=1}^{I^{(s)}} Z_{kit}^{(s)} + \sum_{i=1}^{I^{(n)}} Z_{kit}^{(n)} = \sum_{i=1}^{I^{(s)}+I^{(n)}} Z_{kit}, \quad (13)$$

where $I^{(s)}$ and $I^{(n)}$ are the number of speech and noise basis vectors, respectively, given X_t . Then,

$$\begin{aligned} E(S_{kt} | x_t, \mathbf{y}_t) &= E\left(\sum_{i=1}^{I^{(s)}} Z_{kit}^{(s)} | x_t, \mathbf{y}_t\right) \\ &= \sum_{i=1}^{I^{(s)}} E\left(Z_{kit}^{(s)} | x_t, \mathbf{y}_t\right). \end{aligned} \quad (14)$$

The posterior expected values of the latent variables in (14) are obtained during variational Bayes and are given by [34]:

$$E(Z_{kit} | x_t, \mathbf{y}_t) = \frac{e^{E(\log B_{ki} + \log V_{it} | x_t, \mathbf{y}_t)}}{\sum_{i=1}^{I^{(s)}+I^{(n)}} e^{E(\log B_{ki} + \log V_{it} | x_t, \mathbf{y}_t)}} y_{kt}. \quad (15)$$

Finally, using (15) in (14), we get

$$E(S_{kt} | x_t, \mathbf{y}_t) = \frac{\sum_{i=1}^{I^{(s)}} e^{E(\log B_{ki} + \log V_{it} | x_t, \mathbf{y}_t)}}{\sum_{i=1}^{I^{(s)} + I^{(n)}} e^{E(\log B_{ki} + \log V_{it} | x_t, \mathbf{y}_t)}} y_{kt}. \quad (16)$$

As mentioned before, the posterior distributions of \mathbf{B} and \mathbf{V} are gamma density functions and the required expected values to evaluate (16) are available in closed form. The time-domain enhanced speech signal is reconstructed using (10) and the noisy phase information.

Eq. (16) includes Wiener filtering (2) as a special case. When the posterior distributions of the basis and NMF coefficients are very sharp (which happens for large shape parameters in the gamma distribution), $E(\log V_{it} | x_t, \mathbf{y}_t)$ approaches the logarithm of the mean value of the posterior distribution, $\log(E(V_{it} | x_t, \mathbf{y}_t))$. This can be easily verified by considering that for very large arguments the logarithm provides an accurate approximation to the digamma function. Therefore, for large posterior shape parameters (16) converges asymptotically to (2). In this case, the mean values of the posterior distributions are used to design the Wiener filter.

We can use $\xi_t(\mathbf{y}_t, x_t)$ to classify the acoustic noise more explicitly. For this purpose, we compute the posterior state probability as:

$$f(x_t | \mathbf{y}_1^t) = \frac{f(\mathbf{y}_t, x_t | \mathbf{y}_1^{t-1})}{\sum_{x_t} f(\mathbf{y}_t, x_t | \mathbf{y}_1^{t-1})}. \quad (17)$$

To reduce fluctuations, it is helpful to smooth (17) over time. Other likelihood-based classification techniques have been used in [14, 16] for HMM-based and codebook-driven denoising approaches. In [14], a long-term noise PSD is computed using a separate noise PSD tracking algorithm and is used to select one of the available noise models to enhance the noisy signal. Alternatively, in [16], a single noise HMM is selected during periods of speech pauses and is used to enhance the noisy signal until the next speech pause when a new selection is made. Our proposed classification in (17) neither needs an additional noise PSD tracking algorithm, nor requires a voice activity detector.

3.2 Online Noise Basis Learning for BNMF

We present our scheme to learn the noise basis matrix from the noisy data in this subsection. The online-adapted noise basis is then employed to enhance the noisy signal using the BNMF approach, similarly to Subsection 3.1 with only one state in the HMM. We continue to use a universal speech model that is learned offline.

To update the noise basis, we store N_1 past noisy magnitude DFT frames into a buffer $\mathbf{u} \in \mathbb{R}_+^{K \times N_1}$, where K is the length of \mathbf{y}_t . The buffer will

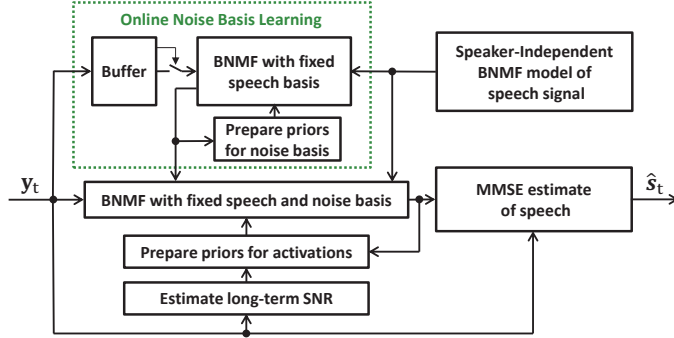


Figure 3: Block diagram representation of BNMF with online noise basis learning. \mathbf{y}_t and $\hat{\mathbf{s}}_t$ are the short-time spectral amplitudes of the noisy and enhanced speech signals, respectively, at time frame t . The goal of the "Prepare priors" boxes is to recursively update the prior distributions, which will be also discussed in Subsection 3.3.

be updated when a new noisy frame arrives. Then, keeping the speech basis unchanged, variational Bayes is applied to \mathbf{u} to find the posterior distributions of both the speech and noise NMF coefficients and noise basis matrix.

Let us denote the noise dictionary at time index $t - 1$ by $f_{\mathbf{B}_{t-1}^{(n)}}(\mathbf{b}_{t-1}^{(n)} | \mathbf{y}_1^{t-1})$. To maintain a slowly varying basis matrix, we flatten $f_{\mathbf{B}_{t-1}^{(n)}}(\mathbf{b}_{t-1}^{(n)} | \mathbf{y}_1^{t-1})$ and use it as the prior distribution for the noise basis matrix at time t . Accordingly, using the notation from (6), we set $\gamma^{(n)} = E(\mathbf{B}_t^{(n)}) = E(\mathbf{B}_{t-1}^{(n)} | \mathbf{y}_1^{t-1})$, and $\psi_{ki}^{(n)}$ is set to a high value ($\psi_{ki}^{(n)} = \psi^{(n)} \gg 1, k = 1, \dots, K, i = 1, \dots, I^{(n)}$) to avoid overfitting. With a high value for the shape parameter, the posterior distributions are flattened only slightly to obtain a quite sharp prior distribution. Therefore, the posteriors of the noise basis matrix are encouraged to follow the prior patterns unless the noise spectrogram changes heavily. Figure 3 shows a simplified diagram of the online BNMF approach. The top part of the figure (dashed-line box) illustrates the online noise basis learning.

Two points have to be considered to complete the online learning. As we do not expect the noise type to change rapidly, we can reduce the computational complexity by updating the noise dictionary less frequently. Also, as an inherent property of NMF, good initializations can improve the dictionary learning. To address these two issues, we use a simple approach based on a sliding window concept. Let us define a local buffer $\mathbf{m} \in \mathbb{R}_+^{K \times N_2}$ that

stores the last N_2 observed noisy DFT magnitudes. Every time we observe a new frame, the columns in $\underline{\mathbf{m}}$ are shifted to the left and the most recent frame is stored at the rightmost column. When the local buffer is full, i.e., N_2 new frames have been observed, a number of frames (let's say q frames) that have the lowest energies are chosen to update the main buffer $\underline{\mathbf{n}}$. Note that to do this we do not use any voice activity detector. Hence, the columns in $\underline{\mathbf{n}}$ are shifted to the left and new data is stored on the rightmost columns of the buffer. We now apply the KL-NMF on $\underline{\mathbf{n}}$ for few iterations, and use the obtained basis matrix to initialize the posterior means of the noise basis matrix. Then, the iterations of variational Bayes (using both speech and noise basis matrices) are continued until convergence.

One of the important parameters in our online learning is N_1 , size of the main buffer. Although a large buffer reduces the overfitting risk, it slows down the adaption speed of the basis matrix. The latter causes the effect of the previous noise to fade out slowly, which will be illustrated in the following example. In our experiments, we set $N_1 = 50$, $N_2 = 15$, $q = 5$. Our approach of the basis adaption is independent of the underlying SNR.

Figure 4 provides a demonstration of the online noise basis learning using a toy example. For this example, a noisy signal (at 0 dB SNR) is obtained by adding two different sinusoidal noise signals to the speech waveform at a sampling rate of 16 kHz. A frame length of 32 ms with 50% overlap and a Hann window was utilized to implement the DFT. We learned a single noise basis vector ($I^{(n)} = 1$) from the noisy mixture. As depicted in the lower panel of Figure 4, the noise basis is adapted correctly to capture the changes in the noise spectrum. BNMF-based speech enhancement resulted to a 13 dB improvement in source to distortion ratio (SDR) [42] and a 0.9 MOS improvement in PESQ [43] for this example.

As Figure 4 demonstrates, the proposed online learning has introduced a latency of around 15 frames in the adaption of the noise basis. In general, this delay depends on both N_2 and the time alignment of the signals, but it is always upper bounded by $2N_2 - q$ short-time frames. Moreover, Figure 4 shows a side effect of the sliding window where the effect of the previous noise is fed out slowly (depending on the parameters N_1 , N_2 and q). However, in a practical scenario, the effect of this latency and slow decay are not as clear as this toy example because the noise characteristics change gradually and not abruptly.

An additional approach to adapt the noise basis is to update the basis matrix in each short-time frame. In this view, variational Bayes is applied to each noisy frame to obtain the posterior distribution of both the NMF coefficients and the noise basis matrix. However, our simulations showed that this approach is not robust enough to changes in the noise type. In fact, to capture the noise spectrogram changes and at the same time not overfit to a single frame, a tradeoff has to be considered in constructing the priors for the noise dictionary, which was difficult to achieve in our simulations.

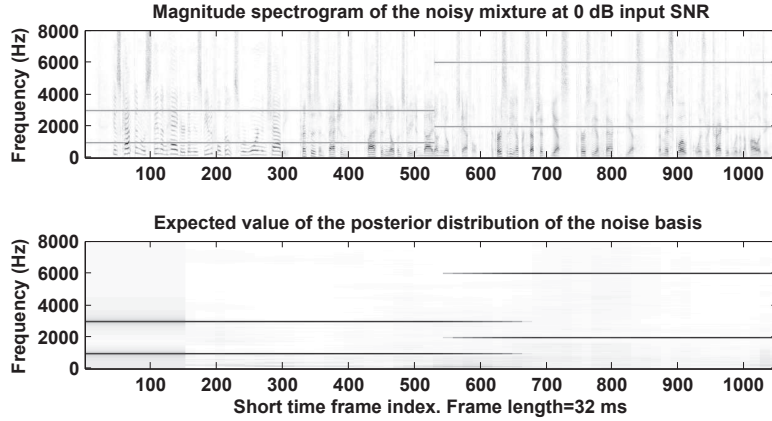


Figure 4: Demonstration of the noise basis adaptation. The top panel shows a mixture magnitude spectrogram in which a sinusoidal noise signal (having two harmonics corresponding to the horizontal lines) is added to a speech signal at 0 dB input SNR. The bottom panel depicts a single noise basis vector over time that is adapted using the noisy mixture. See the text for more explanation.

3.3 Informative Priors for NMF Coefficients

To apply variational Bayes to the noisy signal, we use the temporal dependencies of data to assign prior distributions for the NMF coefficients \mathbf{V} . Both BNMF-based methods from Subsections 3.1 and 3.2 use this approach to recursively update the prior distributions. To model temporal dependencies and also to account for the non-stationarity of the signals, we obtain a prior for \mathbf{V}_t by widening the posterior distributions of \mathbf{V}_{t-1} . Recalling (6), let the state-conditional prior distributions be: $f_{V_{it}}(v_{it} | x_t) = \mathcal{G}(v_{it}; \phi_{it}[x_t], \theta_{it}[x_t]/\phi_{it}[x_t])$ where state dependency is made explicit through the notation $[x_t]$. For this gamma distribution we have:

$$E(V_{it} | x_t) = \theta_{it}[x_t], \quad \frac{\sqrt{\text{var}(V_{it} | x_t)}}{E(V_{it} | x_t)} = \frac{1}{\sqrt{\phi_{it}[x_t]}}, \quad (18)$$

where $\text{var}(\cdot)$ represents the variance. We assign the following recursively updated mean value to the prior distribution:

$$\theta_{it}[x_t] = \alpha \theta_{i,t-1}[x_t] + (1 - \alpha) E(V_{i,t-1} | \mathbf{y}_{t-1}, x_t), \quad (19)$$

where the value of α controls the smoothing level to obtain the prior. Note that due to the recursive updating, θ_{it} is dependent on all the observed noisy data \mathbf{y}_1^{t-1} .

In (18), different shape parameters are used for the speech and noise NMF coefficients, but they are constant over time. Thus, $\phi_{it} = \phi_{i,t-1} = \dots \phi_{i1}$, also $\phi_{it} = \phi^{(s)}$ for $i = 1, \dots, I^{(s)}$, and $\phi_{it} = \phi^{(n)}$ for $i = I^{(s)} + 1, \dots, I^{(s)} + I^{(n)}$. Moreover, different noise types are allowed to have different shape parameters. In this form of prior, the ratio between the standard deviation and the expected value is the same for all the NMF coefficients for a source signal. The shape parameter ϕ represents the uncertainty of the prior which in turn corresponds to the non-stationarity of the signal being processed. We can learn this parameter in the training stage using the clean speech or noise signals. Hence, at the end of the training stage, the shape parameters of the posterior distributions of all the NMF coefficients are calculated and their mean value is taken for this purpose. Using this approach for the speech signal results in $\phi^{(s)} = 3 \sim 5$. However, the noise reduction simulations suggest that having an uninformative prior for speech (a small value for $\phi^{(s)}$) leads to a better performance unless the noise signal is more non-stationary than the speech signal, e.g., keyboard or machine gun noises. Therefore, in our experiments we used a relatively flat prior for the speech NMF coefficients ($\phi^{(s)} \ll 1$) that gives the speech BNMF model greater flexibility.

Our experiments show that the optimal amount of smoothing in (19) depends on the long-term SNR (or global SNR). For low SNRs (high level of noise) a strong smoothing ($\alpha \rightarrow 1$) improves the performance by reducing unwanted fluctuations while for high SNRs a milder smoothing ($\alpha \rightarrow 0$) is preferred. The latter case corresponds to obtaining the mean value θ directly using the information from the previous time frame. Here, in contrast to [31], we use an SNR-dependent value for the smoothing factor. Figure 5 shows an $\alpha - \text{SNR}$ curve that we obtained using computer simulations and was used in our experiments.

To calculate the long-term SNR from the noisy data, we implemented the approach proposed in [44] that works well enough for our purpose. This approach assumes that the amplitude of the speech waveform is gamma-distributed with a shape parameters fixed at 0.4, and that the background noise is Gaussian-distributed, and that speech and noise are independent. Under these assumptions, authors have modeled the amplitude of the noisy waveform with a gamma distribution and have shown that the maximum likelihood estimate of the shape parameter is uniquely determined from the long-term SNR [44].

4 Experiments and Results

We evaluate and compare the proposed NMF-based speech enhancement systems in this section. The experiments are categorized as supervised and unsupervised speech enhancement methods. In Subsection 4.1, we evaluate

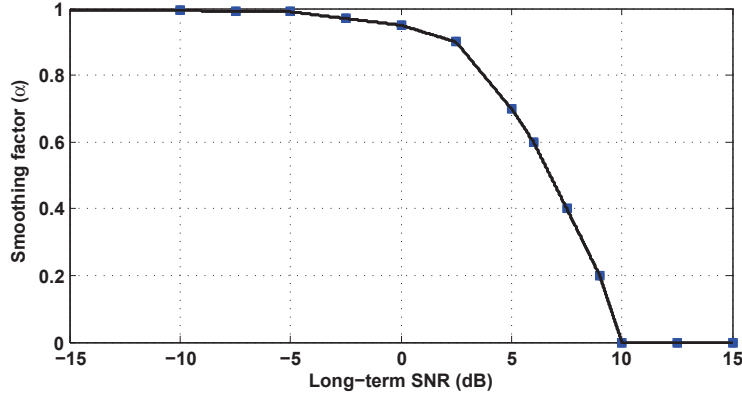


Figure 5: An empirical α -SNR curve, which is used in our experiments. The figure shows that for low input SNRs (high noise levels) a high degree of smoothing should be applied to update the mean values of the prior distributions for NMF coefficients (19), and vice versa.

the noise reduction systems where for each noise type we have access to some training data. Evaluation of the unsupervised denoising schemes is presented in Subsection 4.2, where we assume that we do not have training data for some of the noise types.

In our simulations, all the signals were down-sampled to 16 kHz and the DFT was implemented using a frame length of 512 samples and 0.5-overlapped Hann windows. The core test set of the TIMIT database (192 sentences) [45] was exploited for the noise reduction evaluation. The signal synthesis was performed using the overlap-and-add procedure.

For all the BNMF-based methods, a universal speaker-independent speech model with 60 basis vectors is learned using the training data from the TIMIT database. The choice of dictionary size is motivated by our previous study [46]. Moreover, for the BNMF-based approaches the long-term SNR was estimated using [44] and we used Figure 5 to apply an SNR-dependent smoothing to obtain the priors.

As reviewed in Section 2, the method introduced in [30] factorizes the whole spectrogram of the noisy signal, and therefore, is not causal. In order to make it more practical, we considered two causal extensions of this work and evaluated their performance in this section. The first extension is a supervised approach that works frame by frame. Here, we trained one universal NHMM (100 states and 10 basis vectors per state) for speech and one single-state NHMM for each noise type. To achieve causality, we simply replaced the forward-backward algorithm with the forward algorithm

in which the NMF coefficients from the previous timestamp were used to initialize the current ones. As the other extension, we adapted an online noise dictionary learning, similarly to Subsection 3.2.

4.1 Noise Reduction Using a-Priori Learned NMF Models

We evaluated five variants of NMF-based enhancement methods for three noise types. The considered noise types include factory and babble noises from the NOISEX-92 database [47] and city traffic noise from Sound Ideas [48]. Although all of these three noises are considered non-stationary, the city traffic noise is very non-stationary since it includes mainly horn sounds. We implemented five NMF-based algorithms including:

1. BNMF-HMM: we used (10) in which the noise-type is not known in advance.
2. General-model BNMF: we trained a single noise dictionary by applying BNMF on a long signal obtained by concatenating the training data of all three noises. For the enhancement, (16) was used regardless of the underlying noise type.
3. Oracle BNMF: this is similar to BNMF-HMM but the only difference is that instead of the proposed classifier an oracle classifier is used to choose a noise model for enhancement, i.e., the noise type is assumed to be known a priori and its offline-learned basis matrix is used to enhance the noisy signal. Therefore, this approach is an ideal case of BNMF-HMM.
4. Oracle ML: this supervised method is the maximum likelihood implementation of the Oracle BNMF in which KL-NMF in combination with (2) is used to enhance the noisy signal. Similar to the previous case, an oracle classifier is used to choose a noise model for enhancement. The term ML reflects the fact that KL-NMF arises as the maximum likelihood solution of (4) and (5).
5. Oracle NHMM: this is basically the supervised causal NHMM, as explained earlier in Section 4. Similar to cases (3) and (4), the noise type is assumed to be known in advance.

The number of basis vectors in the noise models were set using simulations performed on a small development set. For BNMF and KL-NMF methods, we trained 100 basis vectors for each noise type. Also, 200 basis vectors were learned for the general noise model. For NHMM, a single state with 100 basis vectors were learned for factory and city traffic noises while 30

basis vectors were pre-trained for babble noise since it provided a better performance.

The performance of the NMF-based methods is compared to a speech short-time spectral amplitude estimator using super-Gaussian prior distributions [7], which is referred to as STSA-GenGamma. Here, we used [12] to track the noise PSD, and we set $\gamma = \nu = 1$ since it is shown to be one of the best alternatives [7]. This algorithm is considered in our simulations as a state-of-the-art benchmark to compare NMF-based systems.

Figure 6 shows the source to distortion ratio (SDR), source to interference ratio (SIR), and source to artifact ratio (SAR) from the BSS-Eval toolbox [42]. SDR measures the overall quality of the enhanced speech while SIR and SAR are proportional to the amount of noise reduction and inverse of the speech distortion, accordingly. For SDR and SIR, the improvements gained by the noise reduction systems are shown. Several interesting conclusions can be drawn from this figure.

The simulations show that the Oracle BNMF has led to the best performance, which is closely followed by BNMF-HMM. The performance of these two systems is quite close with respect to all three measures. This shows the superiority of the BNMF approach, and also, it indicates that the HMM-based classification scheme is working successfully. Another interesting 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. These results were confirmed by our informal listening tests.

Moreover, the figure shows that the Oracle NHMM and General-model BNMF methods lead to similar SDR values. However, these two methods process the noisy signal differently. The NHMM method does not suppress a lot of noise but it does not distort the speech signal either (i.e., SAR is high). This is reversed for the General-model BNMF. Furthermore, comparing BNMF-HMM and General-model BNMF confirms an already reported observation [14, 16] that using many small noise-dependent models is superior to a large noise-independent model.

Figure 7 provides the experimental results using segmental SNR (SegSNR) [49, ch. 10], which is limited to the range $[-10\text{dB}, 30\text{dB}]$, and perceptual evaluation of speech quality (PESQ) [43]. As it can be seen in the figure, the BNMF-based methods have led to the highest SegSNR and PESQ improvements. These results verify again the excellence of the BNMF strategies. Moreover, it is interesting to note that the NHMM method has not been very successful in improving the quality of the noisy speech with respect to the PESQ measure.

To study specifically the classification part of the BNMF-HMM algorithm, we analyzed the output of the classifier. Figure 8 provides the result of this experiment. To have a clearer representation, the probability of each noise type in (17) is smoothed over time and is depicted in Figure 8. Here,

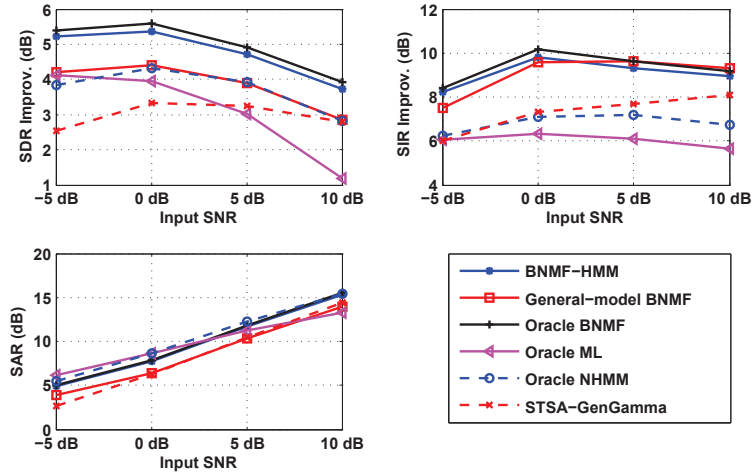


Figure 6: BSS-Eval measures [42] to evaluate and compare the supervised NMF-based denoising algorithms. The BNMf-based schemes are described in Subsection 3.1. Here, the prefix "Oracle" is used for the variants where the noise type is known a priori. The results are averaged over different noise types. For the SDR and SIR, improvements gained by the enhancement systems are shown.

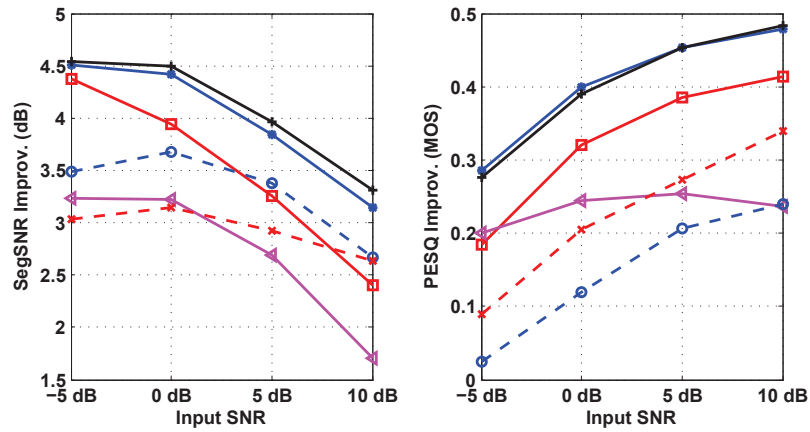


Figure 7: PESQ and Segmental SNR (SegSNR) improvements gained by the supervised enhancement systems. Legend of this figure is similar to that of Figure 6.

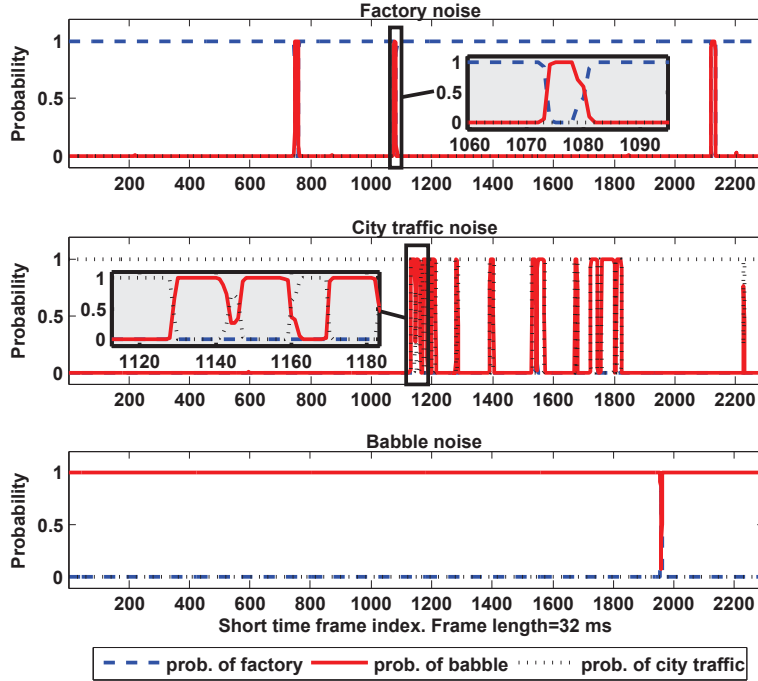


Figure 8: Result of the noise classifier where (17) is smoothed over time and is plotted for a mixture at 0 dB input SNR. The underlying noise type is given in the titles of the subplots (which corresponds to factory, city traffic, and babble noises, respectively, from top to bottom). In each subplot, the probability of three noise classes (factory, city traffic, and babble noises) are shown. For visibility, two small segments are magnified and shown in the figure.

the classifier is applied to a noisy signal at 0 dB input SNR. The underlying noise type is given as the titles of the subplots. As it can be seen in the figure, the classifier works reasonably well in general. Most of the wrong classifications correspond to the case where the true noise type is confused with the babble noise. One reason for this confusion is due to the nature of babble noise. If the short-time spectral properties of the noise are not very different from those of babble, the union of speech and babble basis vectors can explain any noisy signal by providing a very good fit to the speech part. However, as shown in Figure 6 and Figure 7, this confusion has reduced the performance only very marginally.

4.2 Experiments with Unsupervised Noise Reduction

This subsection is devoted to investigating the performance of the unsupervised NMF-based speech enhancement systems. For this purpose, we considered 6 different noise types including factory and babble noises from the NOISEX-92 database [47], and city traffic, highway traffic, ocean, and hammer noises from Sound Ideas [48]. Among these, ocean noise can be seen as a stationary signal in which the noise level changes up to ± 20 dB. All the signals were concatenated before processing.

We evaluated three NMF-based enhancement systems using a general speech model, which is learned similarly to Subsection 4.1. We considered Online BNMF (Subsection 3.2) and Online NHMM (as explained earlier in Section 4). Additionally, we included the BNMF-HMM in the comparison. The considered BNMF-HMM model was identical to that of Subsection 4.1, i.e., we learned only three models for factory, babble and city traffic noises. For the other noise types, the method is allowed to use any of these models to enhance the noisy signal according to (10). Furthermore, we included two state-of-the-art approaches in our experiments: The STSA-GenGamma approach, identical to that of Subsection 4.1, and a Wiener filter in which the noise PSD was estimated using [12] and a decision-directed approach [50] was used to implement the filter. Here, the final gain applied to the noisy signal was limited to be larger than 0.1, for perceptual reasons [51].

For the online BNMF and online NHMM algorithms, we learned $I^{(n)} = 30$ basis vectors for noise. Learning a large basis matrix in this case can lead to overfitting since the dictionary is adapted given a small number of observations ($N_1 = 50$ in our experiments). This was also verified in our computer simulations. Hence, in contrast to the supervised methods for which we learned 100 basis vectors for each noise, we learned a smaller dictionary for online algorithms.

Figure 9 shows the objective measures from BSS-Eval [42] for different algorithms. As it can be seen in the figure, Online BNMF has outperformed all the other systems. This method introduces the least distortion in the enhanced speech signal while performing moderate noise reduction. On the other hand, Wiener filter and STSA-GenGamma reduce the interfering noise greatly with the cost of introducing artifacts in the output signal.

Online NHMM outperforms the Wiener and STSA-GenGamma algorithms at low input SNRs with respect to SDR but for high input SNRs the performance of the algorithm is the worst among all the competing methods. Also, the amount of noise suppression using Online NHMM is the least among different methods.

Moreover, Figure 9 shows that STSA-GenGamma provides a higher-quality enhanced speech signal than the Wiener filter. This is reported frequently in the literature, e.g. [7].

Another interesting result that can be seen in Figure 9 is that Online

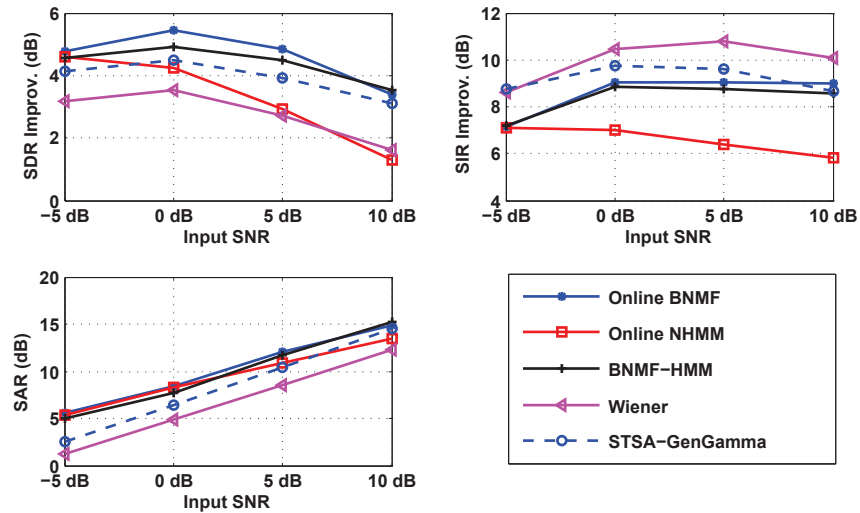


Figure 9: SDR and SIR improvements and SAR measure [42] to evaluate and compare the unsupervised NMF-based denoising algorithms. For the Online BNMF and Online NHMM variants, the noise basis matrix is learned online from the noisy data, explained in Subsection 3.2. The results are averaged over different noise types. For the BNMF-HMM approach, similar to Figure 6, only three noise models are learned.

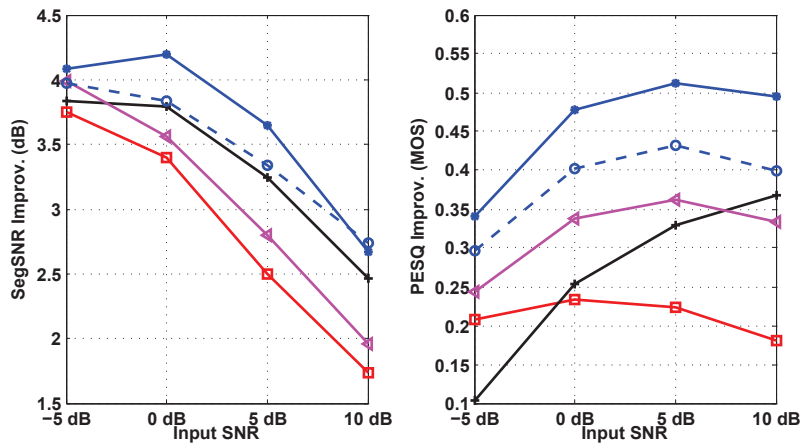


Figure 10: PESQ and Segmental SNR (SegSNR) improvements gained by the unsupervised enhancement systems. Legend of this figure is similar to that of Figure 9.

BNMF outperforms the BNMF-HMM. The difference in the performance is even larger with respect to SegSNR and PESQ, shown in Figure 10. As it is shown in this figure, Online BNMF outperforms the BNMF-HMM (and the other methods) with a large margin.

To have a better understanding on how Online BNMF and BNMF-HMM schemes behave for different noise types, we evaluated SDR and PESQ over short intervals of time. To do so, the noisy and enhanced speech signals were windowed into segments of 5 seconds and then for each segment a SDR and PESQ value was calculated. Figure 11 shows such results as a function of window index. The boundary of the underlying noise types is shown in green in six different levels in which segments belong to factory, babble, city traffic, highway traffic, ocean, and hammer noises, respectively from left to right. As can be seen in the figure, for the first three noise types for which a noise-dependent BNMF model is learned offline the BNMF-HMM approach works marginally better than the Online BNMF. But, for the last three noise types Online BNMF outperforms BNMF-HMM significantly. The difference is highest for the hammer noise; this is due to our observation that the hammer noise differs more from either factory, babble or city traffic noises than highway traffic or ocean noises do. Therefore, neither of the pre-trained models can explain the hammer noise well, and as a result, the overall performance of the BNMF-HMM degrades whenever there is a large mismatch between the training and the testing signals.

A final remark about the Online BNMF and BNMF-HMM can be made considering the computational complexity. In our simulations (where we didn't use parallel processing techniques), Online BNMF runs twice as fast as BNMF-HMM with three states. Moreover, our Matlab implementation of the Online BNMF runs in approximately 5-times real time in a PC with 3.8 GHz Intel CPU and 2 GB RAM.

5 Conclusions

This paper investigated the application of NMF in speech enhancement systems. We developed speech enhancement methods using a Bayesian formulation of NMF (BNMF). We proposed two BNMF-based systems to enhance the noisy signal in which the noise type is not known a priori. We developed an HMM in which the output distributions are assumed to be BNMF (BNMF-HMM). The developed method performs a simultaneous noise classification and speech enhancement and therefore does not require the noise type in advance. Another unsupervised system was constructed by learning the noise BNMF model online, and is referred to as Online BNMF.

Our experiments showed that a noise reduction system using a maximum likelihood (ML) version of NMF—with a universal speaker-independent speech model—does not outperform state-of-the-art approaches. However,

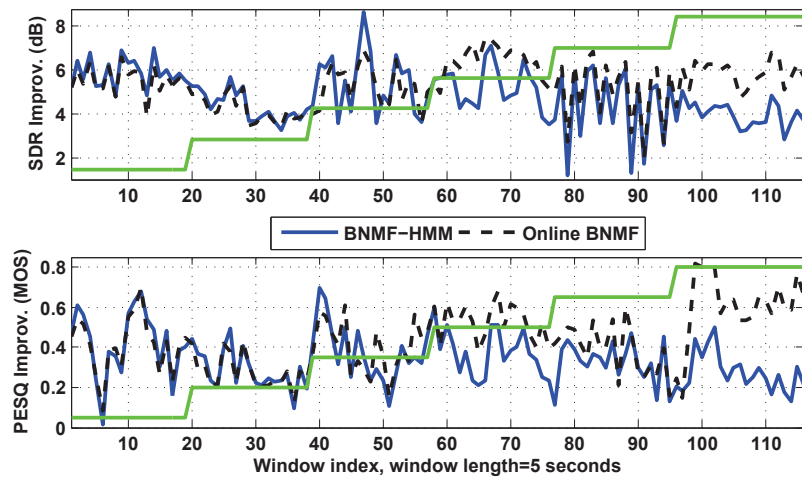


Figure 11: SDR and PESQ measured over short intervals of 5-second long. Six different levels shown in green correspond to factory, babble, city traffic, highway traffic, ocean, and hammer noises, respectively from left to right. For the BNMF-HMM approach, only three noise models corresponding to the first three noises are learned; for the other noise types, the estimator chooses a model that can describe the noisy observation better than the other models.

by incorporating the temporal dependencies in form of prior distributions and using optimal MMSE filters, the performance of the NMF-based methods increased considerably. The Online BNMF method is faster than the BNMF-HMM and was shown to be superior when the underlying noise type was not included in the training data. Our simulations showed that the suggested systems outperform the Wiener filter and an MMSE estimator of speech short-time spectral amplitude (STSA) using super-Gaussian priors with a high margin while they are not restricted to know any priori information that is difficult to obtain in practice.

Acknowledgment

The authors are grateful to Gautham J. Mysore for providing a Matlab implementation of the NHMM approach in [30].

References

- [1] 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.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] I. Cohen, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *Signal Process.*, vol. 86, no. 4, pp. 698–709, apr. 2006.
- [7] 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.

-
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.

-
- [19] 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.
- [20] 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.
- [21] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Process. Systems (NIPS)*. MIT Press, 2000, pp. 556–562.
- [22] 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.
- [23] P. Smaragdis, "Convolutional 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.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] 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.

- [30] 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.
- [31] 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.
- [32] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.
- [33] A. Lefevre, F. Bach, and C. Févotte, "Online algorithms for nonnegative matrix factorization with the Itakura-Saito divergence," in *Proc. IEEE Workshop Applications of Signal Process. Audio Acoust. (WASPAA)*, 2011, pp. 313–316.
- [34] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, 2009, article ID 785152, 17 pages.
- [35] 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.
- [36] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.
- [37] N. Mohammadiha, T. Gerkmann, and A. Leijon, "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.
- [38] 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.
- [39] 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.
- [40] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall, 1993.

-
- [41] 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.
- [42] 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.
- [43] 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.
- [44] C. Kim and R. M. Stern, “Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis,” in *Proc. Int. Conf. Spoken Language Process. (Interspeech)*, 2008, pp. 2598–2601.
- [45] 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.
- [46] 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>
- [47] A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, no. 3, pp. 247–251, jul. 1993.
- [48] B. Nimens *et al.*, “Sound ideas: sound effects collection,” ser. 6000, <http://www.sound-ideas.com/6000.html>.
- [49] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 1st ed. Boca Raton, FL: CRC Press, 2007.
- [50] Y. Ephraim and I. Cohen, *Recent Advancements in Speech Enhancement*. in *The Electrical Engineering Handbook*, CRC Press, 2005.
- [51] D. Malah, R. V. Cox, and A. J. Accardi, “Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 2, mar. 1999, pp. 789–792.

Paper C

Nonnegative HMM for Babble Noise Derived From Speech HMM: Application to Speech Enhancement

Nasser Mohammadiha and Arne Leijon

Refereed article published in
IEEE Transactions on Audio, Speech and Language Processing, vol. 21,
no. 5, pp. 998–1011, may 2013.

©2013 IEEE
Layout has been revised for thesis consistency

Nonnegative HMM for Babble Noise Derived From Speech HMM: Application to Speech Enhancement

Nasser Mohammadiha and Arne Leijon

Abstract

Deriving a good model for multitalker babble noise can facilitate different speech processing algorithms, e.g. noise reduction, to reduce the so-called cocktail party difficulty. In the available systems, the fact that the babble waveform is generated as a sum of N different speech waveforms is not exploited explicitly. In this paper, first we develop a gamma hidden Markov model for power spectra of the speech signal, and then formulate it as a sparse nonnegative matrix factorization (NMF). Second, the sparse NMF is extended by relaxing the sparsity constraint, and a novel model for babble noise (gamma nonnegative HMM) is proposed in which 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. Finally, a noise reduction algorithm is proposed using the derived speech and babble models. All of the stationary model parameters are estimated using the expectation-maximization (EM) algorithm, whereas the time-varying parameters, i.e. the gain parameters of speech and babble signals, are estimated using a recursive EM algorithm. The objective and subjective listening evaluations show that the proposed babble model and the final noise reduction algorithm significantly outperform the conventional methods.

1 Introduction

Multispeaker babble noise is one of the frequently encountered interferences in daily life that greatly degrades the quality and intelligibility of a target speech signal. The problem of understanding the desired speech in the presence of other interfering speech signals and background noise (also known as

the “cocktail party problem”) has received great attention since it was popularized by Cherry in 1953 [1]. Different auditory aspects of this problem are investigated (e.g. [2,3]), and the intelligibility of speech in the presence of multitalker babble noise is examined (e.g. [4]). In addition, there have been few studies that have addressed some babble-specific signal processing techniques to improve speech perception in the presence of background babble noise. In [5], considering a single-channel observation of the babble noise, a framework was proposed to characterize the underlying babble signal. Also, the effect of the number of conversations and speakers was investigated, and a system was proposed to identify the number of speakers in a presented babble noise; moreover, it was shown that this information is beneficial for speaker recognition.

On the contrary, little attention has been paid to develop mathematical babble-specific models that can also be used in signal processing algorithms, e.g. speech enhancement. The goal of speech enhancement algorithms is to improve the quality and intelligibility of the noisy speech, e.g., [6–11], and among different applications, it is very beneficial for hearing aid users [12]. Various classes of single channel model-based speech enhancement approaches have been proposed in the literature. In these methods, for each type of signal (speech or noise) a model is considered and the model parameters are obtained using the training samples of that signal. Then, the task of the speech enhancement is done by defining an interactive model between the speech and noise signals. Some examples of this class of algorithms include the codebook-based approaches [13] and HMM-based methods [14–16]. However, none of these methods exploit the fact that the babble is generated by adding different speech signals, and hence the structure of the considered model for babble noise is similar to that of other noise types. In this paper, we derive a statistical model for babble noise, which takes into account the fact that the babble is generated by adding speech signals of M independent speakers. Then, we propose a single-channel speech enhancement framework that utilizes the derived babble model to enhance the noisy speech signal.

The proposed babble model is based on the nonnegative matrix factorization (NMF). NMF is a technique to approximate a nonnegative matrix \mathbf{X} by a nonnegative linear combination of some basis vectors [17], i.e. $\mathbf{X} \approx \mathbf{TV}$. In speech processing: \mathbf{X} is the spectrogram of the signal with short-time spectral vectors stored as columns in \mathbf{X} , \mathbf{T} is the basis matrix or basis spectral vectors, and \mathbf{V} is called the NMF coefficient matrix. NMF has been used successfully in different fields including blind source separation [18–20], and speech enhancement [21–25]. The “pure addition” property of NMF makes it a powerful technique to be used whenever some nonnegative quanta are added to each other. In the case of babble noise, spectral vectors of different speech signals can be added to generate a spectral vector of babble.

The basic idea of NMF-based speech enhancement algorithms is that, for

each signal, an NMF model is considered and its parameters are obtained using the training data. Then, a mixing model is defined, which usually involves the assumption that the spectrograms of the noise and speech signals are additive, and speech enhancement is carried out by a Wiener-type filtering approach. Two important shortcomings of NMF have to be considered when designing NMF-based speech enhancement systems:

1) The correlation between consecutive time-frames is not handled directly in a standard NMF. To overcome this problem, several approaches have been proposed [21–25]. For instance, a semi-supervised approach (where the noise type is not known a priori) was proposed in [22], which was based on a nonnegative hidden Markov model (NHMM) where the correlation of the signals were taken into account by the transition probability matrix of the underlying HMM. In [25], a Bayesian NMF based speech enhancement algorithm was proposed in which the temporal correlation of the underlying speech and noise signals was exploited through the informative prior distributions.

2) For some noise signals, the noise basis matrix is quite similar to the basis matrix of the speech signal, e.g. the basis matrix of the babble noise should be quite similar to the basis matrix of the speech signal. As a result, the performance of the noise reduction algorithms is usually worse in the case of babble noise [21, 24]. This issue has not been addressed in the available systems and is one of the main focuses of this study.

In this paper, first we derive an ergodic gamma-HMM model for the power spectral coefficients of the speech signal. Next, we formulate the speech model as a sparse NMF. Then, by relaxing the sparsity constraint, we derive a gamma nonnegative hidden Markov model (gamma-NHMM) for babble noise in which the basis matrix is identical to the speech basis matrix, and only the activity of the basis vectors segregates the speech from the babble signal. Moreover, an expectation-maximization (EM) algorithm is proposed to estimate the model parameters. In addition, to employ the derived babble model for speech enhancement, an HMM-based speech enhancement framework in the time-frequency domain is proposed where each power spectral vector of the power spectrogram of speech and babble signals are modeled by the gamma-HMM and gamma-NHMM models, respectively. The proposed framework differs from the state-of-the-art HMM-based approaches [14–16] as we directly model the spectral vectors with HMM. In the available HMM-based methods, the waveform signal is modeled as an autoregressive (AR) process, and hence the waveforms of speech and noise signals are modeled by HMM. Thus, this new framework facilitates a new class of HMM-based speech enhancement algorithms. Similar to [16], the interaction model for the noisy speech signal is constructed by considering a prior distribution over the long-term energy levels of the speech and noise signals. A recursive EM algorithm [26, 27] is developed to estimate the time-varying parameters of these distributions online. The excellence

of the proposed babble model and noise reduction scheme is demonstrated through objective evaluations and a subjective listening test.

The rest of the paper is organized as follows: The gamma-HMM speech signal model is developed in Section 2. In Section 3, the gamma-NHMM model of babble noise is derived. In Section 4, the mixed signal model and noise reduction algorithm is constructed. The estimation of the stationary model parameters and time-varying parameters is described in Section 5. The objective and subjective examination of the noise reduction algorithms are presented in Section 6. Finally, Section 7 concludes the study.

2 Speech Signal Model

2.1 Single-voice Gamma HMM

We model the magnitude-squared DFT coefficients (periodogram coefficients) of the speech signal using an \bar{N} -state HMM with gamma distributions as output probability density functions. Throughout this paper, random variables are represented with capital letters, e.g. $\mathbf{X} = [X_{kt}]$ denotes the matrix of random variables associated with the DFT coefficients of the clean speech, where k is the frequency bin and t denotes the time-frame index. The corresponding realizations are shown with small letters, e.g. $\mathbf{x}=[x_{kt}]$. Also, let $|\cdot|^2$ represents the element-wise magnitude-square operator. The conditional distribution of $|X_{kt}|^2$ is given as:

$$f\left(|x_{kt}|^2 \mid \bar{S}_t = i, G_t = g_t\right) = \frac{\left(|x_{kt}|^2\right)^{\alpha_k - 1}}{\left(g_t b_{ki}\right)^{\alpha_k} \Gamma\left(\alpha_k\right)} e^{-|x_{kt}|^2 / \left(g_t b_{ki}\right)}, \quad (1)$$

where the conditional density $f_{X|Y}(x \mid Y = y)$ is simply shown as $f(x \mid Y = y)$ to keep notations uncluttered, and $\Gamma(\cdot)$ is the Gamma function. Here, \bar{S}_t is the random hidden state of the speech signal, α_k is the shape parameter, b_{ki} is the scale parameter, and G_t is the stochastic gain parameter, which is discussed later. The expected value and variance of X_{kt} are defined as: $E(|X_{kt}|^2 \mid \bar{S}_t = i, G_t = g_t) = \alpha_k g_t b_{ki}$, and $\text{var}(|X_{kt}|^2 \mid \bar{S}_t = i, G_t = g_t) = \alpha_k (g_t b_{ki})^2$.

The gamma assumption for a magnitude-squared DFT coefficient in (1) is motivated by the super-Gaussianity of the speech DFT coefficients [9, 11]. Denote the real and imaginary parts of the DFT coefficient by $\text{Re}\{X_{kt}\}$ and $\text{Im}\{X_{kt}\}$, respectively. Assuming that $\text{Re}\{X_{kt}\}$ and $\text{Im}\{X_{kt}\}$ have a two-sided generalized gamma distribution is equivalent to assuming that $|\text{Re}\{X_{kt}\}|$ and $|\text{Im}\{X_{kt}\}|$ have a generalized gamma distribution. Then, it can be easily shown that $|\text{Re}\{X_{kt}\}|^2$ and $|\text{Im}\{X_{kt}\}|^2$ have a gamma distribution if the γ parameter of the generalized gamma distributions equals 2 (see [11] for a general discussion and definition of γ).

We use the standard assumption that $\text{Re}\{X_{kt}\}$ and $\text{Im}\{X_{kt}\}$ are independent and identically distributed. Since the sum of two independent gamma random variables (RV) with equal scale parameters is a gamma RV, $|X_{kt}|^2 = |\text{Re}\{X_{kt}\}|^2 + |\text{Im}\{X_{kt}\}|^2$ will have a gamma distribution.

In general, state-conditional densities can describe different parts of the speech signal depending on the total number of states. For example, when 50~60 states are available, each state roughly corresponds to one phoneme.

The short-term stochastic gain parameter G_t in (1) is considered to model the long-term changes in the speech energy level over time. Since G_t is nonnegative, we choose to have a gamma distribution to govern G_t in order to simplify the resulting algorithm:

$$f(g_t) = \frac{g_t^{\phi-1}}{\theta_t^\phi \Gamma(\phi)} e^{-g_t/\theta_t}, \quad (2)$$

where ϕ and θ_t are the shape and scale parameters, respectively. In this model, the long-term speech level is modeled by the time-varying scale parameter θ_t , while relative signal-energy levels for different states are modeled by b_{ki} (see Figure 1). Also, we have: $E(G_t) = \phi\theta_t$, and $\text{var}(G_t) = \phi\theta_t^2$. Since $\sqrt{\text{var}(G_t)}/E(G_t) = \sqrt{\phi}$, by using (2) we assume that in the log-domain the standard deviation of outcomes of G_t from its mean value is approximately constant, independent of the long-term level of speech. Considering that $E(G_t)$ is updated for different speech levels, the above assumption of gamma distribution for G_t is reasonable.

The complete HMM output density functions can now be expressed as:

$$f(|\mathbf{x}_t|^2 | \bar{S}_t = i, G_t = g_t) = \prod_{k=1}^K f(|x_{kt}|^2 | \bar{S}_t = i, G_t = g_t), \quad (3)$$

where we have assumed that DFT coefficients at different frequency bins are conditionally independent [6, 9, 11]. The state-conditional probability of the observed power spectral coefficients of the speech signal (which will be used for parameter estimation) can now be computed by integrating out the gain variable. Using the properties of the generalized inverse Gaussian distribution (see Appendix 8.2), this can be obtained in a closed form as:

$$\begin{aligned} f(|\mathbf{x}_t|^2 | \bar{S}_t = i) &= \int_0^\infty f(|\mathbf{x}_t|^2 | \bar{S}_t = i, G_t = g_t) f(g_t) dg_t \\ &= \frac{2\tau^{\nu/2} \mathcal{K}_\nu(2\sqrt{\rho\tau})}{\rho^{\nu/2} \theta_t^\phi \Gamma(\phi)} \prod_{k=1}^K \frac{|x_{kt}|^{2(\alpha_k-1)}}{b_{ki}^{\alpha_k} \Gamma(\alpha_k)}, \end{aligned} \quad (4)$$

where we have defined $\rho = 1/\theta_t$, $\nu = \phi - \sum_{k=1}^K \alpha_k$, $\tau = \sum_{k=1}^K |x_{kt}|^2 b_{ki}^{-1}$, and $\mathcal{K}_\nu(\cdot)$ denotes a modified Bessel function of the second kind.

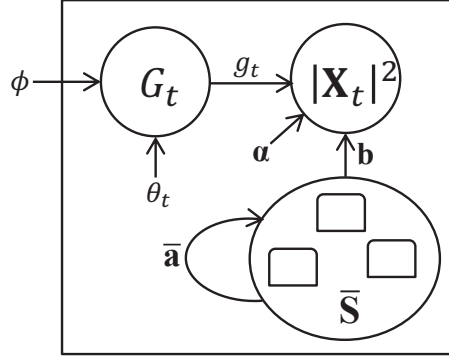


Figure 1: A schematic representation of the HMM with gain modeling. $\bar{\mathbf{a}}$ denotes the transition probability matrix.

The sequence of hidden states is characterized by a first-order Markov model, with the transition probability matrix $\bar{\mathbf{a}}$, and with the elements $\bar{a}_{ij} = f(\bar{S}_t = j \mid \bar{S}_{t-1} = i)$. As we are modeling speech in general, and not a specific utterance, the state Markov chain is considered to be fully connected, and hence *ergodic*, with the time-invariant state probability mass vector $\bar{\mathbf{p}}$, and with the elements $\bar{p}_i = f(\bar{S}_t = i)$.

2.2 Gamma-HMM as a Probabilistic NMF

Instead of denoting the hidden state by its index number, as $\bar{S}_t = i$, we can denote the random discrete state by a one-of- \bar{N} indicator column vector $\bar{\mathbf{S}}_t$, where $\bar{S}_{it} = 1$ and $\bar{S}_{jt} = 0, j \neq i$. Using this notation, the selected state-conditional set of scale parameters $\mathbf{b}_i = (b_{1i}, \dots, b_{Ki})^\top$, given a particular state $\bar{\mathbf{S}}_t$ with $\bar{S}_{it} = 1$, can be simply expressed by $\mathbf{b}\bar{\mathbf{S}}_t$, where all of the \bar{N} state-conditional scale parameter vectors have been collected as columns in the “basis” matrix $\mathbf{b} = (\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_{\bar{N}})$.

The complete sequence of scale parameter vectors for the complete spectrum sequence $|\mathbf{X}|^2 = (|\mathbf{X}_1|^2, |\mathbf{X}_2|^2, \dots, |\mathbf{X}_t|^2, \dots)$ can then be expressed as $\mathbf{b}\bar{\mathbf{S}}$ where $\bar{\mathbf{S}} = (\bar{\mathbf{S}}_1, \bar{\mathbf{S}}_2, \dots, \bar{\mathbf{S}}_t, \dots)$ is the random sequence of the state indicator vectors. The probability of the complete sequence $|\mathbf{x}|^2$ of the observed short-time spectra, given any state sequence $\bar{\mathbf{s}}$ and gain factors $\mathbf{g} = (g_1, g_2, \dots, g_t, \dots)$, can now be obtained as:

$$f\left(|\mathbf{x}|^2 \mid \bar{\mathbf{s}}, \mathbf{g}\right) = \prod_t f\left(|\mathbf{x}_t|^2 \mid \bar{\mathbf{s}}_t, g_t\right), \quad (5)$$

$$f\left(|\mathbf{x}_t|^2 \mid \bar{\mathbf{s}}_t, g_t\right) = \prod_k \frac{\left(|x_{kt}|^2\right)^{\alpha_k - 1}}{\left(g_t [\mathbf{b}\bar{\mathbf{s}}_t]_k\right)^{\alpha_k} \Gamma(\alpha_k)} e^{-|x_{kt}|^2 / \left(g_t [\mathbf{b}\bar{\mathbf{s}}_t]_k\right)}, \quad (6)$$

where $[\cdot]_k$ denotes the k^{th} element of the vector, and we have: $E(|X_{kt}|^2 \mid \bar{\mathbf{s}}_t, g_t) = \alpha_k g_t [\mathbf{b}\bar{\mathbf{s}}_t]_k$. Eq. (6) can be used to derive an NMF representation of any observed nonnegative matrix $|\mathbf{x}|^2$. In order to show this, we approximate an observed sequence by its expected value, under the model assumptions, and show that the expected value is the product of two nonnegative matrices. To compute the expected value, the posterior distribution of the state and gain variables are employed. That is, an NMF approximation $\widehat{|\mathbf{x}_t|^2}$ of an input vector $|\mathbf{x}_t|^2$ is given as:

$$\widehat{|\mathbf{x}_t|^2} = \sum_{\bar{\mathbf{s}}_t} \int E\left(|\mathbf{X}_t|^2 \mid \bar{\mathbf{s}}_t, g_t\right) f\left(\bar{\mathbf{s}}_t, g_t \mid |\mathbf{x}|^2\right) dg_t. \quad (7)$$

Let us define $\hat{\mathbf{b}}$ with elements $\hat{b}_{ki} = \alpha_k b_{ki}$. Noting from (6) that $E(|\mathbf{X}_t|^2 \mid \bar{\mathbf{s}}_t, g_t) = g_t \hat{\mathbf{b}}\bar{\mathbf{s}}_t$, (7) can be written as:

$$\widehat{|\mathbf{x}_t|^2} = \hat{\mathbf{b}} \sum_{\bar{\mathbf{s}}_t} \int g_t \bar{\mathbf{s}}_t f\left(\bar{\mathbf{s}}_t \mid |\mathbf{x}|^2\right) f\left(g_t \mid \bar{\mathbf{s}}_t, |\mathbf{x}|^2\right) dg_t. \quad (8)$$

Here, the conditional state probabilities $f(\bar{\mathbf{s}}_t \mid |\mathbf{x}|^2)$ can be calculated using the forward-backward algorithm [28]. Since g_t depends only on the current observation, $f(g_t \mid \bar{\mathbf{s}}_t, |\mathbf{x}|^2) = f(g_t \mid \bar{\mathbf{s}}_t, |\mathbf{x}_t|^2)$. The posterior distribution of the gain variable is a generalized inverse Gaussian distribution (this is derived in Appendix 8.2 and will also be used in Subsection 5.1). Thus, the required integration in (8) is available in a closed form (Eq. (61)). Denoting $E(g_t \mid \bar{\mathbf{s}}_t, |\mathbf{x}_t|^2) = \int g_t f(g_t \mid \bar{\mathbf{s}}_t, |\mathbf{x}_t|^2) dg_t$, and $\mathbf{u}_t = \sum_{\bar{\mathbf{s}}_t} \bar{\mathbf{s}}_t f(\bar{\mathbf{s}}_t \mid |\mathbf{x}|^2) E(g_t \mid \bar{\mathbf{s}}_t, |\mathbf{x}_t|^2)$, we can write: $\widehat{|\mathbf{x}_t|^2} = \hat{\mathbf{b}}\mathbf{u}_t$. Hence, the proposed gamma-HMM model can be used to factorize a nonnegative matrix $|\mathbf{x}|^2$ into a nonnegative basis matrix $\hat{\mathbf{b}}$ and an NMF coefficients matrix \mathbf{u} as: $|\mathbf{x}|^2 \approx \hat{\mathbf{b}}\mathbf{u}$. In an extremely sparse case where $f(\bar{\mathbf{s}}'_t \mid |\mathbf{x}|^2) = 1$ only for one state $\bar{\mathbf{s}}'_t$, depending on time t , and all of the other states have a zero probability, this model reduces to: $|\mathbf{x}|^2 \approx \hat{\mathbf{b}}\mathbf{u}'$ with $\mathbf{u}'_t = \bar{\mathbf{s}}'_t E(g_t \mid \bar{\mathbf{s}}'_t, |\mathbf{x}_t|^2)$.

3 Probabilistic Model of Babble Noise

We model 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 , $|\mathbf{V}_t|^2$, is given by:

$$E\left(|\mathbf{V}_t|^2\right) = \sum_{m=1}^M E\left(|\mathbf{X}_{mt}|^2\right), \quad (9)$$

where each random vector $|\mathbf{X}_{mt}|^2$ is independently generated by an instance of the gamma-HMM described in Section 2. Note that in (9) different weights are used for different speakers as a consequence of the gain modeling in Section 2. That is, there is a hidden speaker-dependent gain (g_{mt}) in (9) (see also (1)). Eq (9) provides a simplified model of real-life babble noise because we are not modeling reverberations here. There might also be additional noise with a recorded babble, which is not considered in (9). However, it must be mentioned that all of the babble model parameters will be estimated given a babble training data set, with no information about M . Therefore, the estimated parameters will be such that the model explains the considered babble as well as possible.

In this view, the babble noise is still described by an HMM with discrete states defined by the combination of the states for each of the M sources. Since the speech signal has \bar{N} states, there are \bar{N}^M possible discrete states for the babble. As the discrete HMM for the clean speech is already an approximation, and speech should probably rather be modeled with a continuous-state HMM, it would be preferable to describe the babble sequence with a continuous-state HMM. On the other hand, an exact implementation of the EM algorithm for HMMs with a continuous-state 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 [29]. Hence, it would be preferable to avoid a continuous-state structure whenever it is possible. Furthermore, in a real babble noise only a finite number of states (say representative states) would be sufficient for practical purposes to model the normalized spectral shape of the signal; this is indicated, for example, by the success of the vector quantization techniques to quantize continuous signals with a limited number of centroid vectors effectively. Based on these reasons, in the following we model the babble noise with a discrete-state HMM.

Using the text following (6), (9) can be written as:

$$\begin{aligned} E\left(|\mathbf{V}_t|^2 \mid \bar{\mathbf{s}}_{1t}, g_{1t}, \dots, \bar{\mathbf{s}}_{Mt}, g_{Mt}\right) &= \\ \sum_{m=1}^M E\left(|\mathbf{X}_{mt}|^2 \mid \bar{\mathbf{s}}_{mt}, g_{mt}\right) &= \\ \sum_{m=1}^M g_{mt} \hat{\mathbf{b}} \bar{\mathbf{s}}_{mt} = \hat{\mathbf{b}} \sum_{m=1}^M g_{mt} \bar{\mathbf{s}}_{mt}, & \quad (10) \end{aligned}$$

where $\hat{b}_{ki} = \alpha_k b_{ki}$ as before. In the babble HMM, we will now approximate the sum over m in (10) by the babble hidden state vectors and the gain variables. Let us denote the babble hidden state vector at time t by $\check{\mathbf{S}}_t$ (as opposed to the speech state indicator shown by $\bar{\mathbf{S}}_t$) and its realizations by $\check{\mathbf{s}}_t$ that can take one of the \check{N} possible state value vectors $\{\check{\mathbf{s}}'_1, \check{\mathbf{s}}'_2, \dots, \check{\mathbf{s}}'_{\check{N}}\}$. Note that \check{N} is different from the number of speakers in the babble, which is shown by M in (10). Also, denote the stochastic babble gain by random variable H_t and its realizations by h_t .

The power spectrum values of the babble, as defined by (9), (10) are not exactly gamma-distributed¹, given the hidden state. However, our informal simulations showed that the babble DFT coefficients also have super-Gaussian distributions. This is understandable, considering the similarity of speech and babble. Hence, the same argument that was used for the speech model in Subsection 2.1 can be used here to motivate that gamma distribution is a good approximation for the distribution of the babble spectra. We now extend the clean-speech model in (6), just slightly, to model the density of the babble short-time power spectrum as:

$$f(|\mathbf{v}_t|^2 | \check{\mathbf{s}}_t, h_t) = \prod_k \frac{\left(|v_{kt}|^2\right)^{\beta_k-1}}{\left(h_t [\mathbf{b}\check{\mathbf{s}}_t]_k\right)^{\beta_k} \Gamma(\beta_k)} e^{-|v_{kt}|^2 / (h_t [\mathbf{b}\check{\mathbf{s}}_t]_k)}, \quad (11)$$

here, the main new feature is that the hidden state vectors $\check{\mathbf{s}}$ are not indicator vectors (columns of $\bar{\mathbf{s}}$ were one-of- \bar{N} indicator vectors in (6)). This is a result of (9). More specifically, if we set $\beta_k = \alpha_k$ and $\check{\mathbf{s}}_t = \sum_m g_{mt} \bar{\mathbf{s}}_{mt}$, then (11) leads to the same expected value as in (10) with $h_t = 1$. In this context, $\check{\mathbf{s}}_t$ is the weighted sum of the M indicator vectors. The shape parameters β_k are still assumed to be independent of the hidden states, but may be different from the shape parameters α_k of the clean speech model. In this approach, the babble signal is generated as a weighted sum of different clean speech waveforms, thus, the ‘‘basis’’ matrix \mathbf{b} is assumed to be the same and only the weights of the basis vectors are different for the speech and the babble signals. The short-term stochastic gain H_t in (11) is assumed to have a gamma distribution as:

$$f(h_t) = \frac{h_t^{\psi-1}}{\gamma_t^\psi \Gamma(\psi)} e^{-h_t/\gamma_t}. \quad (12)$$

In (12), the scale parameter γ_t represents the long-term energy level of the babble signal, and ψ is the shape parameter. An EM-based algorithm is proposed in Subsection 5.2 to estimate \check{N} babble state value vectors, $\check{\mathbf{s}}'_i$ for $i = 1, \dots, \check{N}$, the state transition probabilities $\check{a}_{ij} = f(\check{\mathbf{S}}_t = \check{\mathbf{s}}'_j | \check{\mathbf{S}}_{t-1} = \check{\mathbf{s}}'_i), \beta_k, \psi$, and γ_t given the recorded babble noise. Eq. (11) is referred as gamma-NHMM since the described model performs an NMF on the scale parameters of the HMM output distributions, which are gamma distributions.

¹Eq. (10) is defined for the expected values; to obtain the exact distribution of the babble power spectral vectors, given the hidden states for all of the speakers, both the summation of individual gamma distributions and the distribution of the cross terms have to be considered.

4 Speech Enhancement Method

A noise reduction scheme is proposed in this section to enhance the speech signal that is degraded by the babble noise. The mixed signal model is described in Subsection 4.1, which is used later in Subsection 4.2 to derive an MMSE estimator for the speech signal.

In the proposed models, the power spectra of the clean speech and the babble noise are conditionally gamma-distributed. Even though a gamma distribution might be a good approximation for the power spectra of the mixed signal, obtaining the MMSE estimator for the clean speech signal is difficult for this case ([30] proposes a solution using this approximation). Therefore, in this part of the paper (which provides an application of the developed babble model) we limit the models to use exponential distributions for the speech and the babble power spectral coefficients ($\alpha_k = 1$ in (1), $\beta_k = 1$ in (11)). This corresponds to the assumption that speech and babble DFT coefficients have complex Gaussian distributions, which have been used successfully in the literature (e.g. [6, 31]). In the experimental section, we show that even with this additional simplification the proposed noise reduction method outperforms the competing algorithms. To keep the generality of the speech and babble models for potential future applications, the proposed parameter estimation algorithm in Section 5 will be given for the general gamma case.

4.1 Clean Speech Mixed with Babble

Assuming that the DFT coefficients of the clean speech and babble noise are complex Gaussian, DFT coefficients of the mixed signal \mathbf{Y} ,

$$\mathbf{Y}_t = \mathbf{X}_t + \mathbf{V}_t, \quad (13)$$

will also have complex Gaussian distribution. Let us represent the composite state of the mixed signal by \mathbf{S}_t that can take one of the $N = \bar{N}\bar{N}$ possible outcomes. Defining $\sigma_{Y_{kt}}^2 = E(|Y_{kt}|^2 | \mathbf{s}_t, g_t, h_t) = E(|X_{kt}|^2 | \bar{\mathbf{s}}_t, g_t) + E(|V_{kt}|^2 | \bar{\mathbf{s}}_t, h_t)$ we have:

$$f(y_{kt} | g_t, h_t, \mathbf{s}_t) = \frac{1}{\pi \sigma_{Y_{kt}}^2} e^{-\frac{|y_{kt}|^2}{\sigma_{Y_{kt}}^2}}, \quad (14)$$

and also

$$f(\mathbf{y}_t | g_t, h_t, \mathbf{s}_t) = \prod_{k=1}^K f(y_{kt} | g_t, h_t, \mathbf{s}_t). \quad (15)$$

The state-conditional distribution of the mixed signal can be obtained by integrating out the gain variables as:

$$f(\mathbf{y}_t | \mathbf{s}_t) = \int \int f(\mathbf{y}_t | g_t, h_t, \mathbf{s}_t) f(g_t, h_t) dg_t dh_t. \quad (16)$$

The required expectations to calculate $\sigma_{\hat{y}_{kt}}^2$ in (14) are obtained considering the models given in (6) and (11). The analytical evaluation of (16) turns out to be difficult; although numerical methods can be used to calculate the required integrations, we approximate the integrand by its behavior near to its maximum by applying Laplace approximation [32, Sec. 4.4]. Hence, we first derive an EM algorithm to obtain the state-dependent Maximum a-Posteriori estimates g'_t and h'_t in Appendix 8.1 based on the following optimization problem:

$$\{g'_t, h'_t\} = \arg \max_{g_t, h_t} f(\mathbf{y}_t | g_t, h_t, \mathbf{s}_t) f(g_t, h_t), \quad (17)$$

then (16) is approximated by

$$f(\mathbf{y}_t | \mathbf{s}_t) \approx f(\mathbf{y}_t | g'_t, h'_t, \mathbf{s}_t) f(g'_t, h'_t) \frac{2\pi}{\sqrt{\det(A_{\mathbf{s}_t})}}, \quad (18)$$

where $\det(A_{\mathbf{s}_t})$ is the determinant of the negative Hessian of $\log f(\mathbf{y}_t, g_t, h_t | \mathbf{s}_t)$ with respect to g_t, h_t , evaluated at the maximum point. The expression for the Hessian matrix is also given in Appendix 8.1.

It should also be mentioned that in [16] an EM algorithm was developed to find the mode of the joint distribution and then $f(y | \mathbf{s}_t)$ was approximated by $f(y, g'_t, h'_t | \mathbf{s}_t)$.

4.2 Clean Speech Estimator

The posterior distribution of the clean speech DFT coefficients given the noisy observations can be written as [14, 16]:

$$f(\mathbf{x}_t | \mathbf{y}_1^t) = \frac{\sum_{\mathbf{s}_t} \eta_t(\mathbf{s}_t) f(\mathbf{x}_t, \mathbf{y}_t | \mathbf{s}_t)}{f(\mathbf{y}_t | \mathbf{y}_1^{t-1})}, \quad (19)$$

where $\mathbf{y}_1^{t2} = \{\mathbf{y}_{t1}, \mathbf{y}_{t1+1}, \dots, \mathbf{y}_{t2}\}$, and $\eta_t(\mathbf{s}_t) = f(\mathbf{s}_t | \mathbf{y}_1^{t-1})$ is the probability of being in the composite state \mathbf{s}_t at time t given all of the noisy observations until time $t - 1$, and is calculated as:

$$\eta_t(\mathbf{s}_t) = f(\mathbf{s}_t | \mathbf{y}_1^{t-1}) = \sum_{\mathbf{s}_{t-1}} a_{\mathbf{s}_{t-1}, \mathbf{s}_t} f(\mathbf{s}_{t-1} | \mathbf{y}_1^{t-1}), \quad (20)$$

with $a_{\mathbf{s}_{t-1}, \mathbf{s}_t} = f(\mathbf{S}_t = \mathbf{s}_t | \mathbf{S}_{t-1} = \mathbf{s}_{t-1}) = \bar{a}_{\bar{\mathbf{s}}_{t-1}, \bar{\mathbf{s}}_t} \bar{a}_{\bar{\mathbf{s}}_{t-1}, \bar{\mathbf{s}}_t}$ because of the independency of the speech and the noise Markov chains, and $f(\mathbf{s}_{t-1} | \mathbf{y}_1^{t-1})$ is the scaled forward variable obtained using the forward algorithm [28]. The

joint distribution of \mathbf{X}_t and \mathbf{Y}_t can also be written as:

$$\begin{aligned} f(\mathbf{x}_t, \mathbf{y}_t | \mathbf{s}_t) &= \int \int f(\mathbf{x}_t, \mathbf{y}_t, g_t, h_t | \mathbf{s}_t) dg_t dh_t \approx \\ f(\mathbf{x}_t | \mathbf{y}_t, g'_t, h'_t, \mathbf{s}_t) &\int \int f(\mathbf{y}_t, g_t, h_t | \mathbf{s}_t) dg_t dh_t \approx \\ f(\mathbf{x}_t | \mathbf{y}_t, g'_t, h'_t, \mathbf{s}_t) f(\mathbf{y}_t, g'_t, h'_t | \mathbf{s}_t) &\frac{2\pi}{\sqrt{\det(A_{\mathbf{s}_t})}}, \end{aligned} \quad (21)$$

where the second line is obtained using a point approximation for $f(\mathbf{x}_t | \mathbf{y}_t, g_t, h_t, \mathbf{s}_t)$ (similarly to [16]), and the last line is obtained by using approximation (18). We can also write:

$$\begin{aligned} f(\mathbf{y}_t | \mathbf{y}_1^{t-1}) &= \sum_{\mathbf{s}_t} \int \eta_t(\mathbf{s}_t) f(\mathbf{x}_t, \mathbf{y}_t | \mathbf{s}_t) d\mathbf{x}_t \\ &\approx \sum_{\mathbf{s}_t} \eta_t(\mathbf{s}_t) f(\mathbf{y}_t, g'_t, h'_t | \mathbf{s}_t) \frac{2\pi}{\sqrt{\det(A_{\mathbf{s}_t})}}. \end{aligned} \quad (22)$$

Denoting $\zeta_t(\mathbf{s}_t, \mathbf{y}_t) = \eta_t(\mathbf{s}_t) f(\mathbf{y}_t, g'_t, h'_t | \mathbf{s}_t) \frac{2\pi}{\sqrt{\det(A_{\mathbf{s}_t})}}$, and using (21) and (22), (19) can be written as:

$$f(\mathbf{x}_t | \mathbf{y}_1^t) = \frac{\sum_{\mathbf{s}_t} \zeta_t(\mathbf{s}_t, \mathbf{y}_t) f(\mathbf{x}_t | \mathbf{y}_t, g'_t, h'_t, \mathbf{s}_t)}{\sum_{\mathbf{s}_t} \zeta_t(\mathbf{s}_t, \mathbf{y}_t)}. \quad (23)$$

Because of the Gaussian assumption, calculating the state-conditional posterior distribution of the clean speech DFT coefficients is straightforward and is given by a complex Gaussian distribution with the mean value obtained via the Wiener filtering:

$$E(X_{kt} | \mathbf{y}_t, g'_t, h'_t, \mathbf{s}_t) = C_{X_{kt}} (C_{X_{kt}} + C_{V_{kt}})^{-1} y_{kt}, \quad (24)$$

and the covariance matrix given as:

$$\begin{aligned} E\left(|X_{kt} - E(X_{kt} | \mathbf{y}_t, g'_t, h'_t, \mathbf{s}_t)|^2 | \mathbf{y}_t, g'_t, h'_t, \mathbf{s}_t\right) \\ = C_{X_{kt}} - C_{X_{kt}} (C_{X_{kt}} + C_{V_{kt}})^{-1} C_{X_{kt}}, \end{aligned} \quad (25)$$

where $C_{X_{kt}} = E(|X_{kt}|^2 | g'_t, \mathbf{s}_t) = \alpha_k g'_t [\mathbf{b}\bar{\mathbf{s}}_t]_k$ and $C_{V_{kt}} = E(|V_{kt}|^2 | h'_t, \mathbf{s}_t) = \beta_k h'_t [\mathbf{b}\bar{\mathbf{s}}_t]_k$. By using (24) and (23), the MMSE estimator of the clean speech DFT coefficients is derived as:

$$\begin{aligned} \hat{\mathbf{x}}_t &= E(\mathbf{X}_t | \mathbf{y}_1^t) \\ &= \frac{\sum_{\mathbf{s}_t} \zeta_t(\mathbf{s}_t, \mathbf{y}_t) E(\mathbf{X}_t | \mathbf{y}_t, g'_t, h'_t, \mathbf{s}_t)}{\sum_{\mathbf{s}_t} \zeta_t(\mathbf{s}_t, \mathbf{y}_t)}, \end{aligned} \quad (26)$$

or equivalently as $\hat{x}_{kt} = \kappa_{kt} y_{kt}$ where the gain parameter is given by:

$$\kappa_{kt} = \frac{\sum_{\mathbf{s}_t} \zeta_t(\mathbf{s}_t, \mathbf{y}_t) C_{X_{kt}} (C_{X_{kt}} + C_{V_{kt}})^{-1}}{\sum_{\mathbf{s}_t} \zeta_t(\mathbf{s}_t, \mathbf{y}_t)}. \quad (27)$$

5 Parameter Estimation

5.1 Speech Model Training

The EM-based Baum-Welch algorithm is followed to train speech and noise models [28, 33]. The parameters of the speech model are denoted by $\lambda = \{\bar{\mathbf{a}}, \mathbf{b}, \boldsymbol{\alpha}, \phi, \theta\}$. Letting the training data consist of R speech utterances, it is assumed that the time-dependent scale parameter of the stochastic gain, θ , remains constant during each utterance for simplicity, hence, denoted by θ_r in the following.

Denote the whole training set by $\bar{\mathbf{o}} = \{(\bar{\mathbf{o}}_{1,1} \dots \bar{\mathbf{o}}_{1,T_1}), \dots, (\bar{\mathbf{o}}_{R,1} \dots \bar{\mathbf{o}}_{R,T_R})\}$ where $\bar{\mathbf{o}}_{r,t} = [\bar{o}_{r,kt}]$ represents the speech power spectral vector of the r^{th} sentence at time t . Similarly, let $\bar{\mathbf{Z}} = \{\bar{\mathbf{S}}, \mathbf{G}\}$ represent the hidden variables in the model, which are not observed. Then, the maximization step in the EM algorithm consists of maximizing

$$Q(\hat{\lambda}, \lambda) = \sum_{\bar{\mathbf{s}}} \int f(\bar{\mathbf{z}} | \bar{\mathbf{o}}, \lambda) \log(f(\bar{\mathbf{z}}, \bar{\mathbf{o}} | \hat{\lambda})) d\mathbf{g}, \quad (28)$$

w.r.t $\hat{\lambda}$, where λ is the estimated parameters from the previous iteration of the EM algorithm. $Q(\hat{\lambda}, \lambda)$ can be written as:

$$Q(\hat{\lambda}, \lambda) = \hat{Q}(\hat{\lambda}, \lambda) + \sum_{r,t,i} \omega_{t,r}(i) \int f(g_{r,t} | \bar{\mathbf{o}}_{r,t}, \bar{S}_{r,t} = i, \lambda) \cdot \left(\log f(g_{r,t} | \hat{\lambda}) + \log f(\bar{\mathbf{o}}_{r,t} | g_{r,t}, \bar{S}_{r,t} = i, \hat{\lambda}) \right) dg_{r,t}, \quad (29)$$

for $r = 1 \dots R$, $t = 1 \dots T_r$, and $i = 1 \dots \bar{N}$. Here, $\hat{Q}(\hat{\lambda}, \lambda)$ includes the terms for optimizing the transition matrix $\hat{\mathbf{a}}$ and is maximized using the standard Baum-Welch algorithm. The posterior state probabilities

$$\omega_{t,r}(i) = f(\bar{S}_{r,t} = i | \bar{\mathbf{o}}, \lambda), \quad (30)$$

are obtained by the forward-backward algorithm [28]. To obtain the new parameters, (29) is differentiated w.r.t the parameters of interest, and the result is set to zero. The objective function in (29) is separable for on the one hand $\{\hat{\mathbf{b}}, \hat{\boldsymbol{\alpha}}\}$, and on the other hand $\{\hat{\phi}, \hat{\theta}\}$. First, consider $\{\hat{\mathbf{b}}, \hat{\boldsymbol{\alpha}}\}$;

obtaining the gradient w.r.t. \hat{b}_{ki} and setting it to zero yields the following estimate:

$$\hat{b}_{ki} = \frac{\sum_{r,t} \omega_{t,r}(i) \bar{o}_{r,kt} E_{G_{r,t}|\bar{S}_{r,t},\lambda}(G_{r,t}^{-1})}{\hat{\alpha}_k \sum_{r,t} \omega_{t,r}(i)} = \frac{\mu_{ki}^o}{\hat{\alpha}_k}, \quad (31)$$

where $E_{G_{r,t}|\bar{S}_{r,t},\lambda}(\cdot)$ is the expectation w.r.t. the posterior distribution of the gain variable $f(g_{r,t} | \bar{o}_{r,t}, \bar{S}_{r,t} = i, \lambda)$, and μ_{ki}^o is defined as:

$$\mu_{ki}^o = \frac{\sum_{r,t} \omega_{t,r}(i) \bar{o}_{r,kt} E_{G_{r,t}|\bar{S}_{r,t},\lambda}(G_{r,t}^{-1})}{\sum_{r,t} \omega_{t,r}(i)}. \quad (32)$$

Inserting (31) into (29), and setting the gradient of the objective function w.r.t. $\hat{\alpha}_k$ to zero yields:

$$\begin{aligned} \varphi(\hat{\alpha}_k) - \log(\hat{\alpha}_k) &= \frac{1}{\sum_{r,t,i} \omega_{t,r}(i)} \times \\ &\sum_{r,t,i} \omega_{t,r}(i) \left(\log \bar{o}_{r,kt} - E_{G_{r,t}|\bar{S}_{r,t},\lambda}(\log G_{r,t}) - \log \mu_{ki}^o \right), \end{aligned} \quad (33)$$

where $\varphi(u) = \frac{d}{du} \log \Gamma(u)$ is the digamma function, and μ_{ki}^o is defined in (32). Hence, (33) is solved first, e.g. by Newton's method, and the obtained $\hat{\alpha}_k$ is inserted into (31) to estimate \hat{b}_{ki} . Similarly, $\hat{\phi}$ and $\hat{\theta}$ can be obtained by first estimating the shape parameter ϕ as:

$$\begin{aligned} \varphi(\hat{\phi}) - \log(\hat{\phi}) &= \frac{1}{\sum_{t,r,i} \omega_{t,r}(i)} \times \\ &\sum_{t,r,i} \omega_{t,r}(i) \left(E_{G_{r,t}|\bar{S}_{r,t},\lambda}(\log G_{r,t}) - \log \mu_r^g \right), \end{aligned} \quad (34)$$

with μ_r^g defined as:

$$\mu_r^g = \frac{\sum_{t,i} \omega_{t,r}(i) E_{G_{r,t}|\bar{S}_{r,t},\lambda}(G_{r,t})}{\sum_{t,i} \omega_{t,r}(i)}, \quad (35)$$

and then using $\hat{\phi}$ to obtain $\hat{\theta}_r = \mu_r^g / \hat{\phi}$. Since the gamma probability density function given in (1) is log concave in α_k and b_{ki} around the stationary points $\hat{b}_{ki}, \hat{\alpha}_k$, these update rules are guaranteed to increase the overall log likelihood score of the parameters [34]. To perform the updates, it is required to calculate the posterior expected values of the functions of the gain variables. The posterior distribution of the gain variable, $f(g_{r,t} | \bar{o}_{r,t}, \bar{S}_{r,t} = i, \lambda)$, is obtained in Appendix 8.2 and is a generalized inverse Gaussian. The required expected values $E_{G_{r,t}|\bar{S}_{r,t},\lambda}(G_{r,t})$, $E_{G_{r,t}|\bar{S}_{r,t},\lambda}(G_{r,t}^{-1})$, and $E_{G_{r,t}|\bar{S}_{r,t},\lambda}(\log G_{r,t})$ are given with (61), (62), and (63), respectively.

5.2 Babble Model Training

The parameters of the babble model are denoted by $\lambda = \{\bar{\mathbf{a}}, \check{\mathbf{s}}', \boldsymbol{\beta}, \psi, \gamma\}$ where $\check{\mathbf{s}}' = \{\check{\mathbf{s}}'_1, \check{\mathbf{s}}'_2, \dots, \check{\mathbf{s}}'_{\check{N}}\}$ is the set of the \check{N} babble state value vectors. These vectors are in principle the weighting factors associated with the basis matrix \mathbf{b} . Letting the training data consist of R different recordings of babble noise, similar to the speech model, it is assumed that the time-dependent scale parameter of the stochastic gain H remains constant during each recording, hence, denoted by γ_r in the following.

Denote the whole training set by $\check{\mathbf{o}}$. The main difference between noise training and speech training is that for the babble training we must also update the babble state value vectors $\check{\mathbf{s}}'_i$ for $i = 1, \dots, \check{N}$ simultaneously with the other parameters. The update rules for $\bar{\mathbf{a}}, \gamma$, and ψ are similar to the update rules of $\bar{\mathbf{a}}, \theta$, and ϕ , respectively. The estimation of $\boldsymbol{\beta}$ and $\check{\mathbf{s}}'$ are coupled. Hence, it is easier to optimize the EM help function $Q(\hat{\lambda}, \lambda)$ w.r.t. these parameters separately, given the previous estimates of them. Obtaining the derivative of $Q(\hat{\lambda}, \lambda)$ w.r.t. β_k and setting it to zero yields the following estimate for β_k :

$$\varphi(\hat{\beta}_k) = \frac{\sum_{r,t,i} \omega_{t,r}(\check{\mathbf{s}}'_i)}{\sum_{r,t,i} \omega_{t,r}(\check{\mathbf{s}}'_i)} \times (\log \check{o}_{r,kt} - \log [\mathbf{b}\check{\mathbf{s}}'_i]_k - E_{H_{r,t}|\check{\mathbf{s}}_{r,t},\lambda}(\log H_{r,t})). \quad (36)$$

The update rule for $\check{\mathbf{s}}'_i$ cannot be obtained in a closed form. Here, we present an approach based on the concave-convex procedure (CCCP) [35, 36] to iteratively maximize the EM help function $Q(\hat{\lambda}, \lambda)$. CCCP is a procedure to find a local minimum of a nonconvex function and is often used to minimize a cost function that can be written as a difference of convex functions. The core idea of this procedure is that the nonconvex function is approximated by a convex function, which can be easily minimized, and then the procedure is iterated until a local minimum is found. The negative EM help function for the babble model is given as:

$$\begin{aligned} -Q(\hat{\lambda}, \lambda) &= Q' + \\ &\sum_{t,r,i} \omega_{t,r}(\check{\mathbf{s}}'_i) \sum_k \left(\frac{\check{o}_{r,kt} E_{H_{r,t}|\check{\mathbf{s}}_{r,t},\lambda}(H_{r,t}^{-1})}{[\mathbf{b}\hat{\check{\mathbf{s}}}'_i]_k} + \hat{\beta}_k \log [\mathbf{b}\hat{\check{\mathbf{s}}}'_i]_k \right) \\ &= Q' + \sum_i \left(P_1(\hat{\check{\mathbf{s}}}'_i) + P_2(\hat{\check{\mathbf{s}}}'_i) \right), \end{aligned} \quad (37)$$

where Q' is independent of the state variables $\check{\mathbf{s}}'$. Due to the summation, it is optimal to minimize (37) w.r.t. each $\hat{\check{\mathbf{s}}}'_i$ independently. It can be easily

shown that P_1 is a convex function where P_2 is a concave function. Using the CCCP procedure, a convex problem is generated as:

$$\begin{aligned} \hat{\mathbf{s}}'_i(l+1) = \underset{\mathbf{x}}{\operatorname{argmin}} P_1(\mathbf{x}) + \mathbf{x}^\top \nabla P_2(\hat{\mathbf{s}}'_i(l)), \quad (38) \\ \text{s.t.} \quad \mathbf{x} \geq 0 \end{aligned}$$

where $\nabla P_2(\mathbf{s})$ represents the gradient of P_2 evaluated at \mathbf{s} , and l is the iteration number. This constrained problem can be solved by usual convex optimization tools, e.g. the interior-point methods [37, ch.11]. This procedure is followed iteratively until a locally optimal solution $\hat{\mathbf{s}}'_i$ is obtained. Even though the CCCP procedure does not lead to a closed form solution, it makes the solution faster and more robust. The required derivatives to solve the above problem are given in Appendix 8.3. In summary, the following algorithm is pursued to find the optimal babble state value vectors:

1. Initialize $\hat{\mathbf{s}}'_i(1)$ using the solution obtained in the previous iteration of the EM algorithm for $i \in \{1 \dots \check{N}\}$, set $l = 1$.
2. For each i , iterate between the following steps until convergence (usually 2–3 iterations are enough):
 - (a) Calculate $\nabla P_2(\hat{\mathbf{s}}'_i(l))$ (64).
 - (b) Solve problem (38) using the interior-point methods.
 - (c) $l = l + 1$.

Since the parameter estimation framework is based on the EM, initialization of the algorithm is important. To assign the initial values for $\hat{\mathbf{s}}'_i$ (before the first iteration of the EM), we generated two minutes of 10-person babble noise using the TIMIT database (2f+2m + 1f_{-1.25dB}+1f_{-3dB}+1f_{-6dB}+1m_{-1.25dB}+1m_{-3dB}+1m_{-6dB}), and the gamma-HMM (Subsection 2.2) was applied independently to each speaker's spectrogram to find the NMF weighting vector \mathbf{u}_t . Then, all of the ten \mathbf{u}_t vectors were summed together to obtain $\mathbf{u}_t^{\text{babble}}$. At the end, a K-means clustering procedure was applied to cluster all of the columns of $\mathbf{u}^{\text{babble}}$ into \check{N} groups whose mean values were used to initialize the state value vectors ($\hat{\mathbf{s}}'_i$) for the babble training.

5.3 Updating Time-varying Parameters

The scale parameters of the stochastic gains are time-variant and, thus, for the purposes of noise reduction they have to be estimated online given only the noisy signal. In the following, the parameters $\lambda_t = \{\theta_t, \gamma_t\}$ are updated in a recursive manner after the estimation of the clean speech signal. Therefore, given the noisy signal, a correction term is calculated and is added to

the current estimates to obtain the new estimate of the parameters. In the remainder of this section, this correction term is obtained and is used to update the time-varying parameters (Eq. (46) and (48)). An algorithm was presented in [27] to estimate the HMM parameters online, that was based on the recursive EM algorithm and the stochastic approximation [26, 38]. Here, we follow a similar procedure as described in [16, 27] to update λ_t . 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. 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. Denote the hidden random variables of the EM algorithm as $\mathbf{Z}_t = \{\mathbf{S}_t, G_t, H_t\}$. Given a noisy observation at time t , \mathbf{y}_t , a new estimate of the parameters $\hat{\lambda}_t$ is obtained by solving:

$$\hat{\lambda}_t = \arg \max_{\lambda_t} Q_t \left(\lambda_t, \hat{\lambda}_1^{t-1} \right), \quad \text{where}$$

$$Q_t \left(\lambda_t, \hat{\lambda}_1^{t-1} \right) = \sum_{\mathbf{s}_1^t} \int \int f \left(\mathbf{z}_1^t \mid \mathbf{y}_1^t, \hat{\lambda}_1^{t-1} \right) \times \log f \left(\mathbf{y}_1^t, \mathbf{z}_1^t \mid \lambda_t \right) d\mathbf{g}_1^t d\mathbf{h}_1^t, \quad (39)$$

where $\hat{\lambda}_1^{t-1} = \{\hat{\lambda}_1, \dots, \hat{\lambda}_{t-1}\}$, $\mathbf{z}_1^t = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_t\}$. As it is shown in [27], the objective function in (39) can be simplified, up to an additive constant, as:

$$Q_t \left(\lambda_t, \hat{\lambda}_1^{t-1} \right) = \text{const.} + \sum_{\tau=1}^t \mathcal{L}_{\tau|t} \left(\lambda_t, \hat{\lambda}_1^{\tau-1} \right), \quad (40)$$

with

$$\mathcal{L}_{\tau|t} \left(\lambda_t, \hat{\lambda}_1^{\tau-1} \right) = \sum_{\mathbf{s}_\tau} \int \int f \left(\mathbf{z}_\tau \mid \mathbf{y}_1^t, \hat{\lambda}_1^{\tau-1} \right) \times \left(\log f \left(g_\tau \mid \mathbf{s}_\tau, \lambda_t \right) + \log f \left(h_\tau \mid \mathbf{s}_\tau, \lambda_t \right) \right) dg_\tau dh_\tau. \quad (41)$$

The parameters of interest can be updated as [27]:

$$\hat{\lambda}_t = \hat{\lambda}_{t-1} + \left(-\frac{\partial^2 Q_t \left(\lambda_t, \hat{\lambda}_1^{t-1} \right)}{\partial \lambda_t^2} \right)^{-1} \frac{\partial \mathcal{L}_{t|t} \left(\lambda_t, \hat{\lambda}_1^{t-1} \right)}{\partial \lambda_t} \Bigg|_{\hat{\lambda}_{t-1}}. \quad (42)$$

Using the Bayes rule, the posterior probability of the hidden states can be written as:

$$\begin{aligned} f(\mathbf{z}_\tau | \mathbf{y}_1^t, \hat{\lambda}_1^{\tau-1}) &= \frac{f(\mathbf{z}_\tau, \mathbf{y}_\tau | \mathbf{y}_1^{\tau-1}, \mathbf{y}_{\tau+1}^t, \hat{\lambda}_1^{\tau-1})}{f(\mathbf{y}_\tau | \mathbf{y}_1^{\tau-1}, \mathbf{y}_{\tau+1}^t, \hat{\lambda}_1^{\tau-1})} = \\ &= \frac{f(\mathbf{s}_\tau | \mathbf{y}_1^{\tau-1}, \mathbf{y}_{\tau+1}^t, \hat{\lambda}_1^{\tau-1}) f(\mathbf{y}_\tau, h_\tau, g_\tau | \mathbf{s}_\tau, \hat{\lambda}_1^{\tau-1})}{f(\mathbf{y}_\tau | \mathbf{y}_1^{\tau-1}, \mathbf{y}_{\tau+1}^t, \hat{\lambda}_1^{\tau-1})}, \end{aligned} \quad (43)$$

where the standard Markov chain property (independency of the observations given the hidden states) is used to get the second line. To reduce the computation effort, (43) is approximated as (this is also done implicitly in [16]):

$$\begin{aligned} f(\mathbf{z}_\tau | \mathbf{y}_1^t, \hat{\lambda}_1^{\tau-1}) &\approx \\ &= \frac{f(\mathbf{s}_\tau | \mathbf{y}_1^{\tau-1}, \hat{\lambda}_1^{\tau-1}) f(\mathbf{y}_\tau, h_\tau, g_\tau | \mathbf{s}_\tau, \hat{\lambda}_1^{\tau-1})}{f(\mathbf{y}_\tau | \mathbf{y}_1^{\tau-1}, \hat{\lambda}_1^{\tau-1})}. \end{aligned} \quad (44)$$

For $t = \tau$, the above approximation is exact. Using (18), (22), and (44) in (41) yields:

$$\begin{aligned} \mathcal{L}_{\tau|t}(\lambda_t, \hat{\lambda}_1^{\tau-1}) &\approx \sum_{\mathbf{s}_\tau} \omega_\tau(\mathbf{s}_\tau, \mathbf{y}_\tau) \times \\ &(\log f(g'_\tau | \mathbf{s}_\tau, \lambda_t) + \log f(h'_\tau | \mathbf{s}_\tau, \lambda_t)), \end{aligned} \quad (45)$$

where $\omega_\tau(\mathbf{s}_\tau, \mathbf{y}_\tau) = \frac{\zeta_\tau(\mathbf{s}_\tau, \mathbf{y}_\tau)}{\sum_{\mathbf{s}_\tau} \zeta_\tau(\mathbf{s}_\tau, \mathbf{y}_\tau)}$ is the scaled forward variable, and $\zeta_t(\mathbf{s}_t, \mathbf{y}_t) = \eta_t(\mathbf{s}_t) f(\mathbf{y}_t, g'_t, h'_t | \mathbf{s}_t) \frac{2\pi}{\sqrt{\det(A_{\mathbf{s}_t})}}$ as in Subsection 4.2. Evaluating (42) for θ_t and γ_t yields:

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \frac{1}{\mathcal{I}_t(\hat{\theta}_{t-1})} \sum_{\mathbf{s}_t} \omega_t(\mathbf{s}_t, \mathbf{y}_t) \left(\frac{-\phi}{\hat{\theta}_{t-1}} + \frac{g'_t}{\hat{\theta}_{t-1}^2} \right), \quad (46)$$

$$\mathcal{I}_t(\hat{\theta}_{t-1}) = \sum_{\tau=1}^t \sum_{\mathbf{s}_\tau} \omega_\tau(\mathbf{s}_\tau, \mathbf{y}_\tau) \left(\frac{-\phi}{\hat{\theta}_{t-1}^2} + \frac{2g'_t}{\hat{\theta}_{t-1}^3} \right). \quad (47)$$

and

$$\hat{\gamma}_t = \hat{\gamma}_{t-1} + \frac{1}{\mathcal{I}_t(\hat{\gamma}_{t-1})} \sum_{\mathbf{s}_t} \omega_t(\mathbf{s}_t, \mathbf{y}_t) \left(\frac{-\psi}{\hat{\gamma}_{t-1}} + \frac{h'_t}{\hat{\gamma}_{t-1}^2} \right), \quad (48)$$

$$\mathcal{I}_t(\hat{\gamma}_{t-1}) = \sum_{\tau=1}^t \sum_{\mathbf{s}_\tau} \omega_\tau(\mathbf{s}_\tau, \mathbf{y}_\tau) \left(\frac{-\psi}{\hat{\gamma}_{t-1}^2} + \frac{2h'_t}{\hat{\gamma}_{t-1}^3} \right). \quad (49)$$

To ensure the required positivity of the step sizes $1/\mathcal{I}_t(\hat{\theta}_{t-1})$ in (46) and $1/\mathcal{I}_t(\hat{\gamma}_{t-1})$ in (48) [27, 38], and to take care of the time-variant parameters, we can modify $\mathcal{I}_t(\hat{\lambda}_{t-1})$ slightly by adding a restriction and forgetting factors to reduce the effect of the previous observations as [16, 27]:

$$\begin{aligned} \mathcal{I}_t(\hat{\theta}_{t-1}) &= \xi_\theta \mathcal{I}_{t-1}(\hat{\theta}_{t-2}) + \\ &\quad \max \left(\beta_\theta, \sum_{\mathbf{s}_t} \omega_t(\mathbf{s}_t, \mathbf{y}_t) \left(\frac{-\phi}{\hat{\theta}_{t-1}^2} + \frac{2g'_t}{\hat{\theta}_{t-1}^3} \right) \right), \end{aligned} \quad (50)$$

$$\begin{aligned} \mathcal{I}_t(\hat{\gamma}_{t-1}) &= \xi_\gamma \mathcal{I}_{t-1}(\hat{\gamma}_{t-2}) + \\ &\quad \max \left(\beta_\gamma, \sum_{\mathbf{s}_t} \omega_t(\mathbf{s}_t, \mathbf{y}_t) \left(\frac{-\psi}{\hat{\gamma}_{t-1}^2} + \frac{2h'_t}{\hat{\gamma}_{t-1}^3} \right) \right), \end{aligned} \quad (51)$$

with $0 < \xi_\theta, \xi_\gamma < 1$, and $0 < \beta_\theta, \beta_\gamma$.

6 Experiments and Results

The capability of the proposed models and the performance of the developed noise reduction algorithm is investigated in various ways. In Subsection 6.1, the details of the implementation of the proposed system is explained. In Subsection 6.2, the developed noise reduction scheme is evaluated and compared to state-of-the-art methods using different objective measures and a subjective listening test. The performance of the developed system is compared to that of the Bayesian NMF (BNMF) based approach [25] and the ETSI (European Telecommunications Standards Institute) front end Wiener filtering [39].

In the BNMF approach, to utilize the temporal correlation of the underlying speech and noise signals, the posterior distributions of the NMF coefficients at the past time instances were widened and applied as the new prior distribution through the Bayesian framework to obtain an MMSE estimator for the speech signal [25]. A comparison to the BNMF method has been motivated by the analogy of the proposed babble model and nonnegative matrix factorization. Also, as it is reported in [25], the BNMF-based noise reduction approach outperforms different competing algorithms. On the other hand, the ETSI two-stage Wiener filter is carefully tuned for good performance in denoising speech [39], and it is considered here to compare the performance of the model-based approaches to a standard approach that does not benefit from trained noise-specific models.

6.1 System Implementation

The proposed models for speech and babble signals were trained using the TIMIT and NOISEX-92 databases, respectively. All of the signals were down-sampled to 16-kHz and the DFT was implemented using a frame length of 320 samples with 50% overlapped windows using a Hann window. For speech, 600 sentences from the training set of the TIMIT were used as training data, and for babble noise the first 75% of the signal was used for training while the rest of the signal was used for the test purposes. To investigate the performance of the algorithms as a function of the number of speakers in the babble noise, a different set of babble training and testing data was used, which is explained in Subsection 6.2.3. Also, the core test set of the TIMIT database (192 sentences) was exploited for the noise reduction evaluation. The signal synthesis was performed using the overlap-and-add procedure.

For the speech model, $\bar{N} = 55$ states were trained in order to roughly identify these states by different phonemes. For the babble model, a discrete HMM with \bar{N} states was considered. As a result, the final mixed signal model includes $N = \bar{N}\bar{N}$ states. To carry out the speech enhancement and calculate the final speech gain κ_{kt} (27), the weighted sum of the state-conditional Wiener filters has to be calculated while for each of the N states, the MAP estimate of the stochastic gain variables has to be performed, which is time consuming in general. Although a large value for \bar{N} may approximate the underlying continuous state-space of the babble model better, it will result in a computationally more expensive system and for \bar{N} larger than 50 a pruning algorithm [15,16] has to be implemented to keep the level of complexity practical. In our experiments, we set $\bar{N} = 10$ (except Subsection 6.2.3) since the performance was quite similar for \bar{N} in the range of 10 to 200. Moreover, we observed that a high shape parameter (5~30) for the stochastic gain variables makes the MAP estimation faster while the performance remains similar. Hence, in our simulations the shape parameters of both the stochastic gain variables were set to 15 although the data-driven estimate of the shape parameter of the speech stochastic gain variable was less than one (this is an indication of a high variation in the state-conditioned energy level of the signal). For this setup, our Matlab implementation runs in approximately 5-times real time² using a PC with 3.8 GHz Intel CPU and 2 GB RAM. The online parameter estimation (50,51) was done using $\xi_\theta = 0.99$, $\xi_\gamma = 0.98$, and $\beta_\theta = \beta_\gamma = 100$, which were set experimentally.

Additionally, motivated by our previous work [24], an exponential smoothing was performed as $\bar{\kappa}_{kt} = 0.4\bar{\kappa}_{k(t-1)} + 0.6\kappa_{kt}$ and the speech signal was estimated as $\hat{x}_{kt} = \bar{\kappa}_{kt}y_{kt}$. This smoothing slightly improves the qual-

²By real time, we mean that the processing of the current frame finishes before the next frame arrives.

ity of the estimated speech signal by smoothing out the gain fluctuations. For the BNMF approach [25], 60 basis vectors for speech and 100 basis vectors for babble were trained using the same training material as explained above. For this method, an informative prior was only used for babble NMF coefficients since applying informative prior for speech NMF coefficients did not result in better noise reduction performance, as also mentioned in [25].

6.2 Evaluations

In this section, we evaluate the proposed system and compare its performance with that of BNMF [25] and ETSI front end Wiener filtering [39]. First we present a general comparison of methods, and then some specific aspects are highlighted. Finally, the results of the subjective listening tests are given.

6.2.1 Objective Evaluation of the Noise Reduction

Five different objective measures were considered for the evaluation: (1) source to distortion ratio (SDR) [40] that represents the overall quality of speech; (2) long-term signal to noise ratio (SNR); (3) segmental SNR (SegSNR) [41, ch. 10], which was limited to the range $[-10 \text{ dB}, 30 \text{ dB}]$; (4) spectral distortion (SD) [42], for which the time-frames with powers 40 dB less than the long-term power level were excluded from the evaluations; (5) perceptual evaluation of speech quality (PESQ) [43]. The evaluation is performed at three input SNRs: 0, 5, and 10 dB.

The results are presented in Figure 2. For SDR, SNR, and SegSNR the improvements in dB (e.g. $\Delta\text{SDR} = \text{SDR}_{\text{enhanced}} - \text{SDR}_{\text{noisy}}$) are shown in this figure for readability. For PESQ and SD the actual values for the enhanced signals and for the noisy input signal are shown. A high degree of consistency can be seen between the different measures. The results show that the two model-based approaches lead to much better improvements than the Wiener filtering. The proposed method outperforms the BNMF in all of the input SNRs in the sense of SDR, SNR, SegSNR, and SD. For PESQ, gamma-NHMM results to slightly better PESQ improvement at 0 dB input SNR, while BNMF leads to slightly better improvements at 5 and 10 dB SNRs. However, the difference between PESQ values for two algorithms is marginal in all three SNRs.

6.2.2 Effect of Systems on Speech and Noise Separately

A desired feature of a noise reduction system is that the speech signal remains undistorted. In order to compare this aspect of the algorithms, segmental speech SNR ($\text{SNR}_{\text{seg-sp}}$), and segmental noise reduction (SegNR) [44] were measured in a shadow filtering framework. Hence, the enhancement

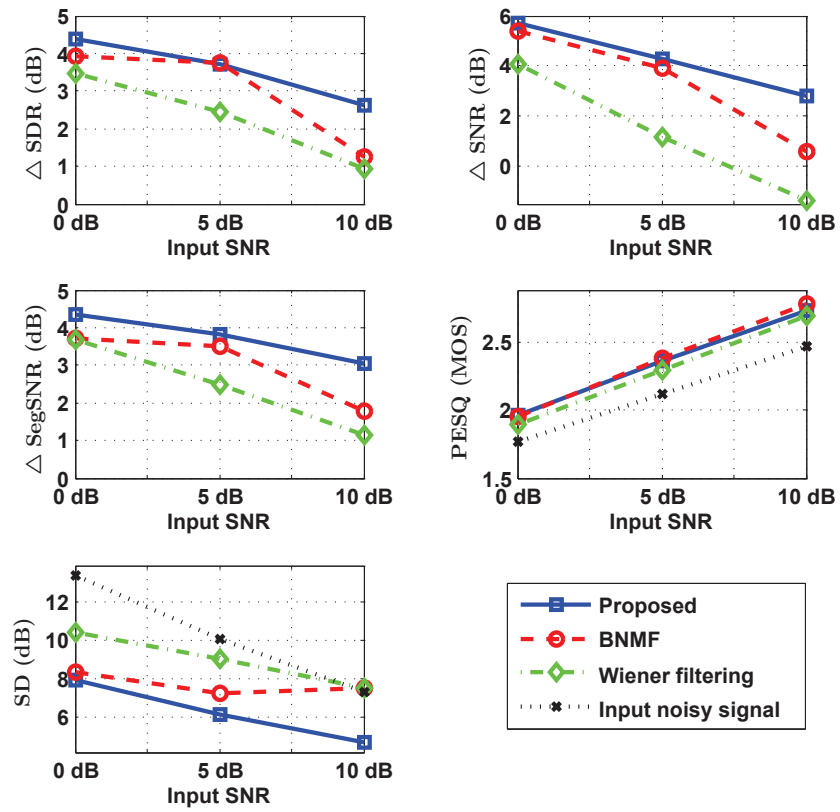


Figure 2: Objective evaluation of the application of the proposed babble model in noise reduction. Δ is used to show the improvements gained by the noise reduction algorithms, e.g., $\Delta\text{SDR} = \text{SDR}_{\text{enhanced}} - \text{SDR}_{\text{noisy}}$.

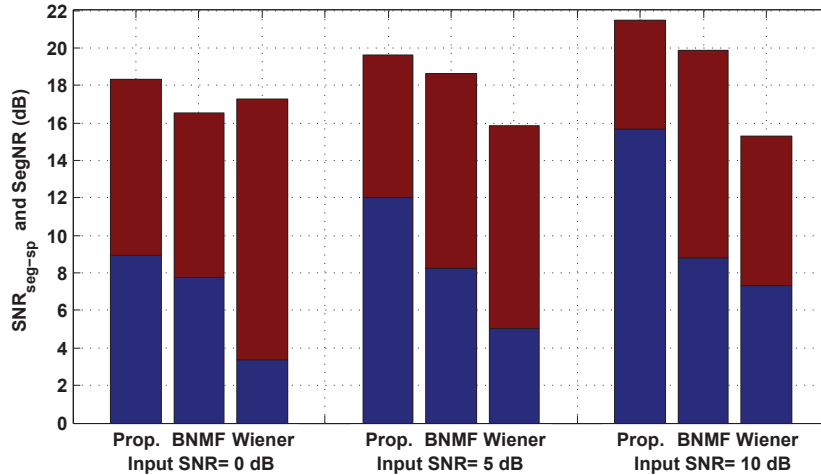


Figure 3: Stacked presentation of the segmental speech SNR ($\text{SNR}_{\text{seg-sp}}$, bottom), and the segmental noise reduction (SegNR, top).

filter was obtained using the input noisy signal (as it was done in 6.2.1), and it was applied to the clean speech and noise components of the input noisy signal, separately. The output speech and noise signals were compared to the corresponding inputs to compute these two measures. For both measures a high value is desired, and $\text{SNR}_{\text{seg-sp}}$ is inversely proportional to the speech distortion.

The results are shown in Figure 3. As it can be seen in the figure, the proposed system leads to a higher segmental speech SNR (less distortion) in all of the input SNRs. Also, the sum of the $\text{SNR}_{\text{seg-sp}}$ and SegNR is the highest for the proposed method.

6.2.3 Effect of the Number of Speakers in Babble

It is well known that the performance of the model-based noise reduction systems that are trained for a specific signal degrades when there is a mismatch between training and testing. Therefore, in the case of a mismatch, the standard Wiener filter might outperform the model-based approaches since it is not restricted to any specific noise type. In this part, we investigate the performance of the noise reduction algorithms as a function of the number of speakers in the babble. For the experiments, an artificial babble was generated by adding waveforms of different speakers from the TIMIT database, with equal speech level for all of the speakers. The number of speakers in generating babble were chosen as $M \in \{4, 6, 10, 20, 50, 100\}$.

Moreover, the babble noise from NOISEX-92 is also considered in the evaluation for comparison.

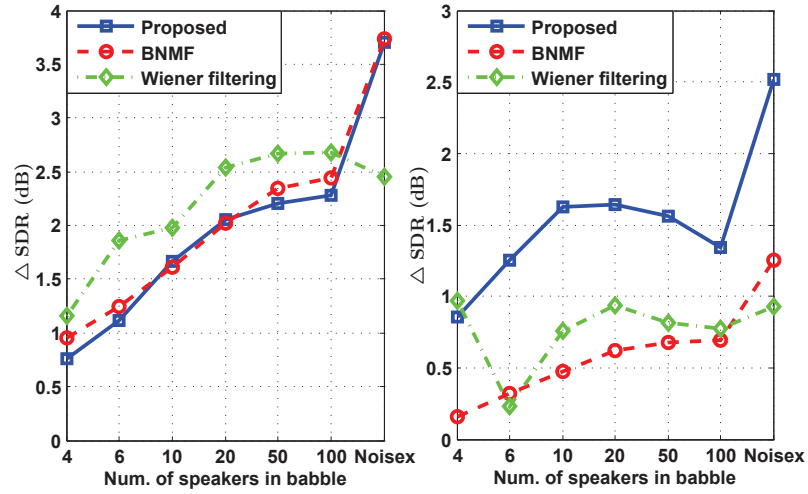
For the proposed method and BNMF, two babble models were trained using (1) only the NOISEX-92, (2) both the NOISEX-92 and 10-person babble noise (different from the test signal). Also, for the proposed system, we trained $\tilde{N} = 50$ states for the babble for both of the models since a pilot experiment indicated that $\tilde{N} = 10$ was insufficient in this case because of the high non-stationarity of the noise.

Improvements gained in the source to distortion ratio (Δ SDR) are shown in Figure 4 for two input SNRs, 5 and 10 dB. Figure 4a shows the results using the babble model that is trained on only the NOISEX-92. Looking at the 5 dB input SNR scenario (left-hand side of Figure 4a), it can be seen that even though the performance of the model-based approaches is much better when exposed to the NOISEX-92 babble noise, the ETSI Wiener filter gives a better result in the other types of babble noise. Figure 4b shows the results using the babble model that is trained using both the NOISEX-92 and 10-person artificial babble noise. Here, the performance of the model-based approaches is slightly reduced for the NOISEX-92 babble noise, but in general their performance is significantly improved (especially for the proposed method). This also implies that if the proposed method is combined with another system that estimates the number of the speakers from the observed babble signal (for example [5]), the performance might be improved further.

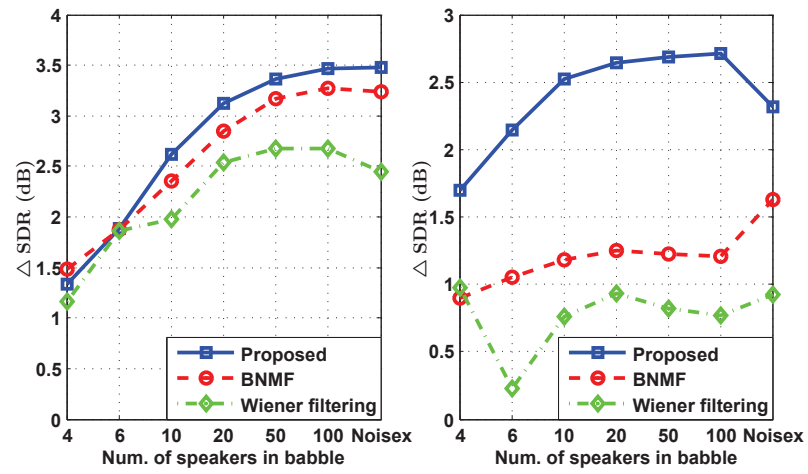
6.2.4 Cross-predictive Test for Model Fitting

A cross-predictive test was carried out in which both of the speech and babble models from the proposed and the BNMF frameworks were applied to the speech and babble signals (as separate inputs) in a predictive way. Here, the goal is to investigate whether the babble (speech) model fits to the babble (speech) signal better than the speech (babble) model. Two measures were computed to compare the input and the estimated signals. To compare the signals in the spectral domain, spectral distortion (SD) [42] was measured. To compare the input and estimated waveforms, segmental SNR (SegSNR) [41, ch. 10] was measured. To reconstruct the output waveforms, the NMF representations of the spectrograms together with the phase information from the input signal were fed into the inverse DFT. For the proposed method, the gamma-NHMM representation (similar to Subsection 2.2) was used to obtain the NMF approximation, and for the BNMF that was achieved by multiplying the mean values of the posterior distributions of the basis matrix and the coefficients matrix.

The results of this predictive test are shown in TABLE 1 in the form of confusion matrices. If a model is good then for each type of signal (each row in the table), the best result should be found in the element on the main



(a) Babble models trained using only NOISEX-92. The results are shown for two input SNRs 5 dB (left) and 10 dB (right).



(b) Babble models trained using both the NOISEX-92 and 10-person babble (different from the testing signals). The results are shown for two input SNRs 5 dB (left) and 10 dB (right).

Figure 4: Performance of the noise reduction algorithms as a function of the number of speakers in the babble.

Table 1: A cross-predictive test for the different models. The specified signal is fed as the input signal to the given model and the quality of the reconstructed signal is measured. The results are averaged over the test set explained in Subsection 6.1.

(a) Spectral distortion (SD, lower value is desired) in dB.

Proposed	Speech Model	Babble Model	BNMF	Speech Model	Babble Model
Speech Sig.	3.9	6.7	Speech Sig.	6.3	9.4
Babble Sig.	3.2	2.4	Babble Sig.	2.2	2.8

(b) Segmental SNR (SegSNR, higher value is desired) in dB.

Proposed	Speech Model	Babble Model	BNMF	Speech Model	Babble Model
Speech Sig.	3.2	-2.1	Speech Sig.	9.1	-2.2
Babble Sig.	4	5.3	Babble Sig.	9.2	6.5

diagonal. Both of the measures point in the same direction, and show that in the proposed framework a better score is obtained for the speech and babble signals using the speech and babble models, respectively. However, for the BNMF, the speech model gives a better score to the babble signal than the babble model itself (shown in a red color in the table). This is because the babble spectrogram can be approximated quite well by combining the speech basis vectors freely. The result of this test is another indication of the excellence of the proposed babble model, and provides an additional explanation for the achieved results in the previous subsections.

6.2.5 Subjective Evaluation of the Noise Reduction

To assess the subjective quality of the estimated speech signal, a subjective listening test was carried out. The test setup was similar to the ITU recommendation ITU-R BS.1534-1 MUSHRA [45]. Six experienced and four inexperienced listeners (ten in total) participated in the test. The subjective evaluation was performed for three input SNRs (0, 5, 10 dB), and for each SNR seven sentences from the core test set of the TIMIT database (4 males and 3 females) were presented to the listeners. In each of the 21 listening sessions, 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, 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 (as suggested in [10]). The listeners were allowed to play each sentence as many times as they wanted, and they always had

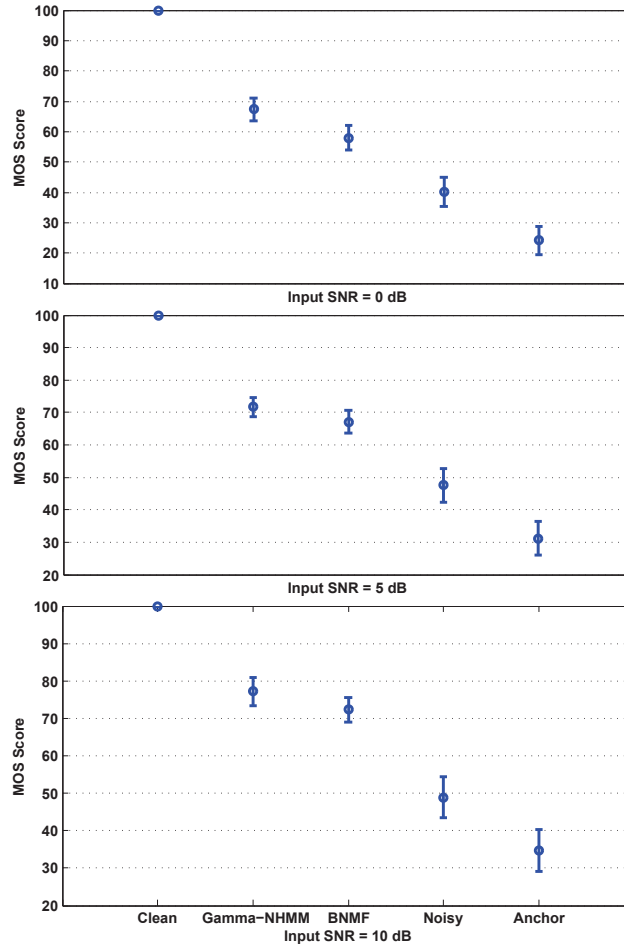


Figure 5: Results of the MUSHRA test at 3 input SNRs: 0 dB, 5 dB, 10 dB (top to down) with a 95% confidence interval.

access to the reference signal. They were asked to rate the signals based on the global speech quality. Also, some sample signals were presented, and the graphical user interface was introduced to the listeners prior to the test procedure. The order of the signals was randomized with respect to the algorithm and input SNR. Each listener took around 30 minutes on average to complete the listening test.

The results of this listening test, averaged over all of the participants, with a 95% confidence interval are presented in Figure 5. 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 100. Also, both of the algorithms were preferred over the noisy input signal by at least 20 units. 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.

To verify the statistical significance of the preference of the gamma-NHMM algorithm, a one-tailed t-test was performed. This statistical analysis shows that the gamma-NHMM leads to a significantly better performance than the BNMF at all three SNRs. For 0 dB, the significance level was $p \approx 9.10^{-5}$, for 5 dB it was $p \approx 0.013$, and finally for 10 dB we obtained $p \approx 0.014$.

7 Conclusion

As babble noise is generated by adding some different speech signals, improving the intelligibility and quality of the speech signal degraded by the babble noise has been a challenging task for a long time. In this paper, a gamma nonnegative HMM was proposed to model the normalized power spectra of babble noise in which the babble basis vectors were identical to the speech basis vectors. In the proposed models, the time-varying energy levels of speech and babble signals were modeled by gamma distributions whose scale parameters were estimated online.

The simulations show that the proposed system achieves better model recognition (i.e. babble signal gets a better score with the babble model rather than the speech model) compared to the Bayesian NMF approach. Also, the objective evaluations and the subjective MUSHRA listening test verify the excellence of the proposed noise reduction system. For instance, at 0 dB input SNR, the enhanced speech of the currently developed system was preferred by around 10 MOS units to the enhanced speech of the closest competing algorithm (Bayesian NMF) and by 27 to the input noisy signal in the scale of 100. Moreover, the simulations show that the proposed noise reduction scheme is less sensitive to a mismatch (varying number of speakers in babble) compared to the other competing model-based approach.

8 Appendix

8.1 MAP Estimate of the Gain Variables

Problem (17) is a MAP estimator that can be solved by the standard EM algorithm. Let the hidden variables for EM be $\mathbf{Z} = \{\mathbf{X}_t, \mathbf{V}_t\}$, and $\lambda = \{g'_t, h'_t\}$ be the parameters of interest. Thus, the EM help function is written to

$Q(\hat{\lambda}, \lambda) = E_{\mathbf{Z}|\mathbf{Y}_t, \lambda}(\log f(\mathbf{y}_t, \mathbf{x}_t, \mathbf{v}_t | \mathbf{s}_t, \hat{\lambda}) + \log f(\hat{\lambda}))$. The terms containing \hat{g}'_t can be gathered into

$$Q_{\hat{g}'_t}(\hat{\lambda}, \lambda) = E_{\mathbf{Z}|\mathbf{Y}_t, \lambda} \left(\log f(\mathbf{x}_t | \hat{g}'_t, \bar{S}_t = i) + \log f(\hat{g}'_t) \right). \quad (52)$$

Taking the derivative of (52) and setting it to zero yields the solution:

$$\hat{g}'_t = \frac{-(K\theta_t - \theta_t(\phi - 1)) + \sqrt{(K\theta_t - \theta_t(\phi - 1))^2 + 4\theta_t C_X}}{2}, \quad (53)$$

where K is the number of frequency bins, dimension of \mathbf{y}_t , and $C_X = \sum_{k=1}^K \frac{E(|X_{kt}|^2 | \mathbf{Y}_t, \lambda)}{\alpha_k b_{ki}}$. The posterior expected value of $|X_{kt}|^2$ is calculated using (24,25). The update rule for \hat{h}'_t is also given as:

$$\hat{h}'_t = \frac{-(K\gamma_t - \gamma_t(\psi - 1)) + \sqrt{(K\gamma_t - \gamma_t(\psi - 1))^2 + 4\gamma_t C_V}}{2}, \quad (54)$$

where $C_V = \sum_{k=1}^K \frac{E(|V_{kt}|^2 | \mathbf{Y}_t, \lambda)}{\beta_k [\mathbf{b}\bar{\mathbf{s}}'_t]_k}$. In very rare cases in practice, the above algorithm may get stuck at a non-maximum stationary point in which the EM algorithm has to be repeated from a different initial point to obtain a local maximum of the likelihood.

The negative Hessian matrix in (18) is defined as:

$$A_{\mathbf{s}_t}(1, 1) = -\frac{\partial^2 (\log f(\mathbf{y}_t, g_t, h_t | \mathbf{s}_t))}{\partial g_t \partial g_t} = \frac{(\phi - 1)}{g_t^2} - \sum_{k=1}^K \frac{(g_t \alpha_k \mathbf{b}_{ki})^2}{(g_t \alpha_k \mathbf{b}_{ki} + h_t \beta_k [\mathbf{b}\bar{\mathbf{s}}'_t]_k)^2} \left(1 - \frac{2|y_{kt}|^2}{(g_t \alpha_k \mathbf{b}_{ki} + h_t \beta_k [\mathbf{b}\bar{\mathbf{s}}'_t]_k)} \right), \quad (55)$$

$$A_{\mathbf{s}_t}(1, 2) = A_{\mathbf{s}_t}(2, 1) = -\frac{\partial^2 (\log f(\mathbf{y}_t, g_t, h_t | \mathbf{s}_t))}{\partial g_t \partial h_t} = -\sum_{k=1}^K \frac{(g_t \alpha_k \mathbf{b}_{ki}) (h_t \beta_k [\mathbf{b}\bar{\mathbf{s}}'_t]_k)}{(g_t \alpha_k \mathbf{b}_{ki} + h_t \beta_k [\mathbf{b}\bar{\mathbf{s}}'_t]_k)^2} \left(1 - \frac{2|y_{kt}|^2}{(g_t \alpha_k \mathbf{b}_{ki} + h_t \beta_k [\mathbf{b}\bar{\mathbf{s}}'_t]_k)} \right), \quad (56)$$

$$A_{\mathbf{s}_t}(2, 2) = -\frac{\partial^2 (\log f(\mathbf{y}_t, g_t, h_t | \mathbf{s}_t))}{\partial h_t \partial h_t} = \frac{(\psi - 1)}{h_t^2} - \sum_{k=1}^K \frac{(h_t \beta_k [\mathbf{b}\bar{\mathbf{s}}'_t]_k)^2}{(g_t \alpha_k \mathbf{b}_{ki} + h_t \beta_k [\mathbf{b}\bar{\mathbf{s}}'_t]_k)^2} \left(1 - \frac{2|y_{kt}|^2}{(g_t \alpha_k \mathbf{b}_{ki} + h_t \beta_k [\mathbf{b}\bar{\mathbf{s}}'_t]_k)} \right). \quad (57)$$

8.2 Posterior Distribution of the Gain Variables

The posterior distribution of the stochastic gain variable of the speech signal in Subsection 5.1, given the hidden Markov state and the observation is given by:

$$f(g_{r,t} | \bar{\mathbf{o}}_{r,t}, \bar{S}_{r,t} = i, \lambda) = \frac{f(\bar{\mathbf{o}}_{r,t} | g_{r,t}, \bar{S}_{r,t} = i, \lambda) f(g_{r,t} | \lambda)}{f(\bar{\mathbf{o}}_{r,t} | \bar{S}_{r,t} = i, \lambda)}, \quad (58)$$

where $\lambda = \{\bar{\mathbf{a}}, \mathbf{b}, \boldsymbol{\alpha}, \phi, \theta\}$ is the estimated parameters from the previous iteration of the Baum-Welch algorithm. Since the denominator is constant, using (3) and (2) we get:

$$\begin{aligned} \log f(g_{r,t} | \bar{\mathbf{o}}_{r,t}, \bar{S}_{r,t} = i, \lambda) &\propto \\ &\sum_{k=1}^K \left(-\alpha_k \log g_{r,t} - \frac{\bar{o}_{r,kt}}{g_{r,t} b_{k,i}} \right) + (\phi - 1) \log g_{r,t} - \frac{g_{r,t}}{\theta_r} = \\ &-\frac{1}{\theta_r} g_{r,t} + \left(\phi - 1 - \sum_{k=1}^K \alpha_k \right) \log g_{r,t} - \left(\sum_{k=1}^K \frac{\bar{o}_{r,kt}}{b_{k,i}} \right) \frac{1}{g_{r,t}}. \end{aligned} \quad (59)$$

Eq. (59) corresponds to a generalized inverse Gaussian (GIG) distribution [46] with parameters $\vartheta = \phi - \sum_{k=1}^K \alpha_k$, $\rho = \frac{1}{\theta_r}$, and $\tau = \sum_{k=1}^K \frac{\bar{o}_{r,kt}}{b_{k,i}}$. The GIG distribution is generally defined as:

$$\begin{aligned} \log \text{GIG}(g; \vartheta, \rho, \tau) &= -\rho g + (\vartheta - 1) \log g - \frac{\tau}{g} + \\ &\frac{\vartheta}{2} \log \rho - \log 2 - \frac{\vartheta}{2} \log \tau - \log \mathcal{K}_{\vartheta}(2\sqrt{\rho\tau}), \end{aligned} \quad (60)$$

for $g \geq 0, \rho \geq 0$, and $\tau \geq 0$. Here, $\mathcal{K}_{\vartheta}(\cdot)$ denotes a modified Bessel function of the second kind. The required expectations are given as [46]:

$$E(G) = \frac{\mathcal{K}_{\vartheta+1}(2\sqrt{\rho\tau}) \sqrt{\tau}}{\mathcal{K}_{\vartheta}(2\sqrt{\rho\tau}) \sqrt{\rho}}, \quad (61)$$

$$E(G^{-1}) = \frac{\mathcal{K}_{\vartheta-1}(2\sqrt{\rho\tau}) \sqrt{\rho}}{\mathcal{K}_{\vartheta}(2\sqrt{\rho\tau}) \sqrt{\tau}}, \quad (62)$$

$$E(\log G) = \frac{\partial \mathcal{K}_{\vartheta}(2\sqrt{\rho\tau})}{\partial \vartheta} \Big|_{\vartheta} + \log \sqrt{\frac{\tau}{\rho}}. \quad (63)$$

The posterior distribution of the stochastic gain variable of the noise signal can be obtained similarly.

8.3 Gradient and Hessian for Babble States

The gradient of P_2 evaluated at $\widehat{\mathbf{s}}'_i$, which is used in the CCCP procedure in Subsection 5.2, is simply given as:

$$\nabla P_2 \left(\widehat{\mathbf{s}}'_i \right) = \left[\frac{\partial P_2 \left(\widehat{\mathbf{s}}'_i \right)}{\partial \widehat{s}'_{mi}} \right] = \sum_{r,t,k} \omega_{t,r} \left(\widehat{\mathbf{s}}'_i \right) \mathbf{b}_k^\top \left(\frac{\beta_k}{\left[\mathbf{b} \widehat{\mathbf{s}}'_i \right]_k} \right), \quad (64)$$

where \mathbf{b}_k denotes the k^{th} row of the basis matrix \mathbf{b} , and $'\top'$ denotes the transpose. Denoting $C(\mathbf{x}) = P_1(\mathbf{x}) + \mathbf{x}^\top \nabla P_2(\widehat{\mathbf{s}}'_i)$, the gradient and the hessian of the cost function in (38) are also given as:

$$\begin{aligned} \nabla C(\mathbf{x}) &= \left[\frac{\partial C(\mathbf{x})}{\partial x_m} \right] = \nabla P_2 \left(\widehat{\mathbf{s}}'_i \right) - \\ &\sum_{r,t,k} \omega_{t,r} \left(\widehat{\mathbf{s}}'_i \right) \mathbf{b}_k^\top \left(\frac{\ddot{\omega}_{r,kt}}{\left[\mathbf{b} \mathbf{x} \right]_k^2} E_{H_{r,t} | \widehat{\mathbf{s}}_{r,t,\lambda}} \left(H_{r,t}^{-1} \right) \right), \end{aligned} \quad (65)$$

$$\begin{aligned} \nabla^2 C(\mathbf{x}) &= \left[\frac{\partial^2 C(\mathbf{x})}{\partial x_m \partial x_n} \right] = \\ &\sum_{r,t,k} \omega_{t,r} \left(\widehat{\mathbf{s}}'_i \right) \mathbf{b}_k^\top \mathbf{b}_k \left(\frac{2\ddot{\omega}_{r,kt}}{\left[\mathbf{b} \mathbf{x} \right]_k^3} E_{H_{r,t} | \widehat{\mathbf{s}}_{r,t,\lambda}} \left(H_{r,t}^{-1} \right) \right). \end{aligned} \quad (66)$$

Acknowledgment

Part of this work was supported by the EU Initial Training Network AUDIS (grant 2008-214699). The authors would like to thank W. Bastiaan Kleijn for a useful discussion about the babble model.

References

- [1] E. C. Cherry, "Some experiments on the recognition of speech, with one and two ears," *Journal of Acoustical Society of America (JASA)*, vol. 25, pp. 975–979, 1953.
- [2] B. Arons, "A review of the cocktail party effect," *Journal of Acoustical Society of America (JASA)*, vol. 12, pp. 35–50, 1992.
- [3] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol. 17, pp. 1875–1902, 2005.

-
- [4] S. A. Simpson and M. Cooke, "Consonant identification in N-talker babble is a nonmonotonic function of N," *Journal of Acoustical Society of America (JASA)*, vol. 118, no. 5, pp. 2775–2778, 2005.
- [5] N. Krishnamurthy and J. H. L. Hansen, "Babble noise: Modeling, analysis, and applications," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 7, pp. 1394–1407, sep. 2009.
- [6] 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.
- [7] X. Shen and L. Deng, "A dynamic system approach to speech enhancement using the H_∞ filtering algorithm," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 4, pp. 391–399, jul. 1999.
- [8] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," vol. 10, no. 3, pp. 173–185, mar. 2002.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] H. Levitt, "Noise reduction in hearing aids: An overview," *Journal of Rehabilitation Research and Development*, vol. 38, no. 1, pp. 111–121, 2001.
- [13] 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.
- [14] 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.

-
- [15] 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.
- [16] 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.
- [17] 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.
- [18] P. Smaragdis, "Convolutional 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.
- [19] 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.
- [20] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutional mixtures for audio source separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 3, pp. 550–563, mar. 2010.
- [21] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, 2008, pp. 4029–4032.
- [22] 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.
- [23] N. Mohammadiha, T. Gerkmann, and A. Leijon, "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.
- [24] —, "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.

- [25] 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.
- [26] D. M. Titterton, "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>
- [27] 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.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] I. Cohen, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *Signal Process.*, vol. 86, no. 4, pp. 698–709, apr. 2006.
- [32] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.
- [33] 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.
- [34] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, vol. 1, pp. 29–45, 1986.
- [35] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, pp. 915–936, 2003.
- [36] 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.

-
- [37] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [38] 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.
- [39] "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.
- [40] 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.
- [41] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 1st ed. Boca Raton, FL: CRC Press, 2007.
- [42] K. K. Paliwal and W. B. Kleijn, "Quantization of LPC parameters," in *Speech Coding and Synthesis*, W.B. Kleijn, K.K. Paliwal, Eds. New York: Elsevier, 1995, ch. 12, pp. 443–466.
- [43] 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.
- [44] 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.
- [45] *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>
- [46] T. Kawamura and K. Iwase, "Characterizations of the distributions of power inverse Gaussian and others based on the entropy maximization principle," *Journal of The Japan Statistical Society*, vol. 33, no. 1, pp. 95–104, 2003.

Paper D

Spectral Domain Speech Enhancement Using HMM State-dependent Super-Gaussian Priors

Nasser Mohammadiha, Rainer Martin, and Arne Leijon

Refereed article published in
IEEE Signal Processing Letters, vol. 20, no. 3, pp. 253–256, mar. 2013.

©2013 IEEE
Layout has been revised for thesis consistency

Spectral Domain Speech Enhancement Using HMM State-dependent Super-Gaussian Priors

Nasser Mohammadiha, Rainer Martin, and Arne Leijon

Abstract

The derivation of MMSE estimators for the DFT coefficients of speech signals, given an observed noisy signal and super-Gaussian prior distributions, has received a lot of interest recently. In this letter, we look at the distribution of the periodogram coefficients of different phonemes, and show that they have a gamma distribution with shape parameters less than one. This verifies that the DFT coefficients for not only the whole speech signal but also for individual phonemes have super-Gaussian distributions. We develop a spectral domain speech enhancement algorithm, and derive hidden Markov model (HMM) based MMSE estimators for speech periodogram coefficients under this gamma assumption in both a high uniform resolution and a reduced-resolution Mel domain. The simulations show that the performance is improved using a gamma distribution compared to the exponential case. Moreover, we show that, even though beneficial in some aspects, the Mel-domain processing does not lead to better results than the algorithms in the high-resolution domain.

1 Introduction

Time-frequency domain single-channel noise reduction approaches using super-Gaussian priors have received a lot of attention during recent years. The real and imaginary parts of the speech (and noise) DFT coefficients, for instance, are better modeled with super-Gaussian distributions, e.g. Laplacian and two-sided gamma distributions, than with a Gaussian distribution [1]. Several approaches have been proposed to derive MMSE estimators for the DFT coefficients of speech, given the noisy signal, using these super-Gaussian prior distributions [1–3]. In these works the super-Gaussianity

is considered for the long-term statistics of a speech signal and not conditioned on the phoneme type. Hence, an interesting question is whether this phenomenon depends on the phoneme type.

Moreover, it is important for signal processing algorithms to investigate the distribution of the speech and noise DFT coefficients given the so called “hidden state”, which can be considered as the phoneme type. This can be very beneficial in deriving better estimators in the HMM-based speech enhancement approaches [4–7]. Traditionally, HMM-based noise reduction schemes have been derived by assuming auto-regressive (AR) models for the speech and noise signals, and then the AR parameters are assumed to be Gaussian [4–6]. Recently, an HMM-based speech enhancement approach was proposed in [8] in which the DFT coefficients of the speech and noise signals were assumed to be complex Gaussian.

This letter proposes two main contributions:

1. We explore the distribution of the state-conditional speech DFT coefficients. Our experiments show that phoneme-dependent periodogram coefficients have a gamma (with shape parameters less than one) rather than an exponential distribution.

2. We extend the HMM-based speech enhancement algorithm from [8] and derive new MMSE estimators for the speech power spectral coefficients using super-Gaussian prior distributions, given the noisy signal. We assume 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 denoising algorithm is superior to algorithms using the exponential distribution. Hence, the results support the super-Gaussianity hypothesis. Furthermore, we compare the performance of the derived estimators in the high-resolution DFT domain and in the reduced-resolution Mel frequency domain.

2 Conditional Distribution of the Speech Power Spectral Coefficients

In this section, we look at the distribution of the speech power spectral coefficients (estimated using periodogram or magnitude-squared DFT coefficients) conditioned on the hidden state that can be seen as the phoneme type. We denote the random variables associated with the speech DFT coefficients and their realizations by \bar{O}_{mt} and \bar{o}_{mt} , respectively, where m is the frequency bin and t is the time-frame index. Moreover, let $|\cdot|^2$ represent the element-wise magnitude-square operator. Let us define the conditional

gamma distribution as:

$$f\left(|\bar{o}_{mt}|^2 \mid \bar{S}_t = i\right) = \frac{\left(|\bar{o}_{mt}|^2\right)^{\alpha_{mi}-1}}{\left(b_{mi}\right)^{\alpha_{mi}} \Gamma\left(\alpha_{mi}\right)} e^{-|\bar{o}_{mt}|^2/b_{mi}}, \quad (1)$$

where $\bar{S}_t = \bar{s}_t \in [1, \bar{N}]$ is the hidden state, α_{mi} and b_{mi} are the state-dependent shape and scale parameters, and $\Gamma(\cdot)$ is the complete Gamma function. If $\bar{N} = 50 \sim 60$, each state is identified roughly by one phoneme. For (1), we have: $E(|\bar{O}_{mt}|^2 \mid \bar{S}_t = i) = \alpha_{mi} b_{mi}$ and $\text{var}(|\bar{O}_{mt}|^2 \mid \bar{S}_t = i) = \alpha_{mi} b_{mi}^2$.

For $\alpha_{mi} = 1$, (1) reduces to an exponential distribution. This corresponds to assuming that real and imaginary parts of the DFT coefficients ($\text{Re}\{\bar{O}_{mt}\}$ and $\text{Im}\{\bar{O}_{mt}\}$) have a Gaussian distribution. For $\alpha_{mi} < 1$, however, the resulting distribution for DFT coefficients will be super-Gaussian, as shown next. Assuming that $\text{Re}\{\bar{O}_{mt}\}$ and $\text{Im}\{\bar{O}_{mt}\}$ are independent and identically distributed, Eq. (1) leads to a gamma distribution for $|\text{Re}\{\bar{O}_{mt}\}|^2$ and $|\text{Im}\{\bar{O}_{mt}\}|^2$ with shape parameters equal to $\alpha_{mi}/2$. This is because the sum of two independent gamma random variables (RV) with equal scales is a gamma RV. Then, it can be easily shown that $|\text{Re}\{\bar{O}_{mt}\}|$ and $|\text{Im}\{\bar{O}_{mt}\}|$ have generalized gamma distribution with $\nu = \alpha_{mi}/2, \gamma = 2, \beta = 1/b_{mi}$ (see [2] for definition of these parameters), or equivalently, $\text{Re}\{\bar{O}_{mt}\}$ and $\text{Im}\{\bar{O}_{mt}\}$ have two-sided generalized gamma distributions.

2.1 Experimental Data

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 waveform of each realization was normalized to have unit variance. To obtain the spectral coefficients, first, each waveform was windowed into short-time frames using a Hann window with a frame length of 20 ms and 50% overlap, and second, the DFT was applied to these short-time frames to obtain the periodogram.

The top panel of Figure 1 shows the shape parameters of the estimated gamma distributions for two phonemes, ‘‘ah’’ and ‘‘sh’’. The estimation of the shape and scale parameters of the gamma distributions was done using a standard maximum-likelihood approach independently for each frequency bin. As Figure 1 shows, the shape parameters for these two phonemes are less than one at all frequencies. In the bottom panel of Figure 1, 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 Figure 1 for comparison. As a result, we find that the power spectral coefficients will have gamma rather than exponential distributions even if we limit the speech data to come from a specific phoneme and normalize each realization. Therefore,

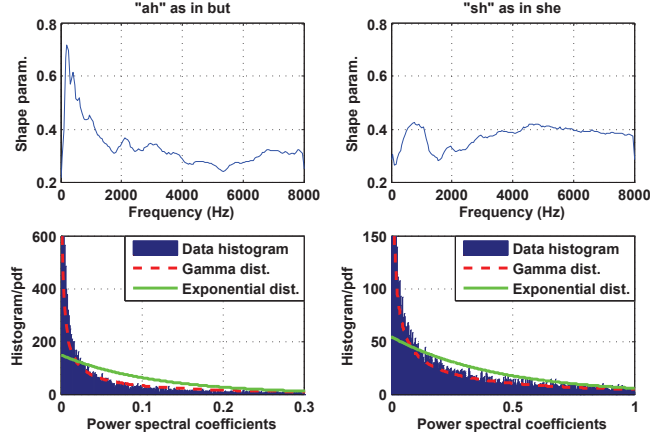


Figure 1: Experimental distribution of the speech power spectral coefficients for two speech sounds “ah” and “sh”. The top panel shows the estimated shape parameters of the fitted gamma distributions at different frequencies. The bottom panel shows the histogram and estimated distributions of spectral coefficients of “ah” at 2500 Hz (left), and of “sh” at 6000 Hz (right).

real and imaginary parts of the phoneme-conditioned speech DFT coefficients have super-Gaussian distributions. As the top-left panel of Figure 1 shows, there are distinct differences between phones: the shape parameters of gamma distributions corresponding to “ah” are higher at frequencies close to the main formants due to less variation in the signal energy in these frequencies. This can be generalized to other vowels as well.

3 HMM-based Speech Enhancement

3.1 Speech Model

Besides DFT coefficients, we also consider a more coarse resolution frequency as it reduces the number of model parameters and smoothes out the signals’ random variations before processing. In this case, the power at adjacent speech DFT bins is summed to obtain $\mathbf{X} = [X_{kt}]$, with elements representing the frame band power in analysis band k , as

$$X_{kt} = \sum_{m=m_i(k)}^{m_h(k)} w_m(k) |\bar{O}_{mt}|^2, \quad (2)$$

where w denotes a set of overlapped triangular filters that approximate the Mel-scale filter bank, m_l and m_h represent the band-dependent lowest and highest DFT indices to be summed, respectively. If $m_l = m_h$ and $w_m = 1$, we recover the original spectra as: $X_{kt} = |\bar{O}_{kt}|^2$. The state dependent conditional distribution of X_{kt} is now obtained by a slight modification of (1):

$$f(x_{kt} | \bar{S}_t = i, G_t = g_t) = \frac{(x_{kt})^{\alpha_{ki}-1}}{(g_t b_{ki})^{\alpha_{ki}} \Gamma(\alpha_{ki})} e^{-x_{kt}/(g_t b_{ki})}, \quad (3)$$

where G_t is a short-term stochastic gain parameter, which is assumed to have a gamma distribution as:

$$f(g_t) = \frac{g_t^{\phi-1}}{\theta_t^\phi \Gamma(\phi)} e^{-g_t/\theta_t}. \quad (4)$$

Here, ϕ is the shape parameter and θ_t is a time-varying scale parameter, which models the long-term speech level. Assuming conditional independence of the elements of \mathbf{X}_t [1,2], the HMM output density functions for a given state can be expressed as:

$$f(\mathbf{x}_t | \bar{S}_t = i, G_t = g_t) = \prod_{k=1}^K f(x_{kt} | \bar{S}_t = i, G_t = g_t). \quad (5)$$

The sequence of the speech hidden states are characterized by a fully connected first-order Markov model with transition probability matrix $\bar{\mathbf{a}}$, with elements $\bar{a}_{i'i} = f[\bar{S}_t = i | \bar{S}_{t-1} = i']$, and a time-invariant state probability mass vector $\bar{\mathbf{p}}$, with elements $\bar{p}_i = f[\bar{S}_t = i]$. The parameters of the speech model denoted by $\lambda = \{\bar{\mathbf{a}}, \mathbf{b}, \boldsymbol{\alpha}, \phi, \theta\}$ are obtained from training data using the EM algorithm [8]¹

3.2 Noise Model

Let $\ddot{\mathbf{O}} = [\ddot{O}_{mt}]$ denote the noise DFT coefficients. The noise band power spectral vectors, $\mathbf{V} = [V_{kt}]$, are obtained similarly to (2). The noise signal is modeled using an \ddot{N} -state HMM with hidden states denoted as \ddot{S}_t . The noise power spectral coefficients are assumed to have an Erlang distribution which includes the exponential distribution as a special case and provides a sufficiently accurate fit to the data:

$$f(v_{kt} | \ddot{S}_t = j, H_t = h_t) = \frac{(v_{kt})^{\beta_k-1} e^{-v_{kt}/(h_t c_{kj})}}{(h_t c_{kj})^{\beta_k} (\beta_k - 1)!}, \quad (6)$$

¹The update equation of the speech shape parameters has to be modified slightly to exclude the summation over the states since $\boldsymbol{\alpha}$ are state-dependent.

where β_k is the state-independent integer shape parameter, c_{kj} is the scale parameter, and "!" represents the factorial function. The short-term stochastic gain parameter of the noise is also assumed to have a gamma distribution as:

$$f(h_t) = \frac{h_t^{\psi-1}}{\gamma_t^\psi \Gamma(\psi)} e^{-h_t/\gamma_t}. \quad (7)$$

The noise Markov chain construction and parameter estimation is done similarly to speech model (Subsection 3.1). The only difference is that after each iteration of the EM algorithm, the shape parameters are rounded to the closest integer numbers.

3.3 Speech Estimation: Complex Gaussian Case

This subsection presents a speech enhancement algorithm in the DFT domain, i.e., a special case of (2) where $X_{mt} = |\bar{O}_{mt}|^2$ and $V_{mt} = |\bar{O}_{mt}|^2$. Assuming that the DFT coefficients of the clean speech and noise signals are complex Gaussian ($\alpha_{ki} = \beta_k = 1$ in (3), (6)), DFT coefficients of the mixed signal \mathbf{O} , $\mathbf{O}_t = \bar{\mathbf{O}}_t + \check{\mathbf{O}}_t$, will also have complex Gaussian distributions. Let us represent the composite hidden state of the mixed signal by S_t with realizations s_t that can take one of the $\bar{N}\check{N}$ possible outcomes. Let $\sigma_{O_{mt}}^2 = E(X_{mt} | \bar{s}_t, g_t) + E(V_{mt} | \check{s}_t, h_t)$, which is calculated considering (3) and (6). We have:

$$f(o_{mt} | g_t, h_t, s_t) = \frac{1}{\pi \sigma_{O_{mt}}^2} e^{-\frac{|o_{mt}|^2}{\sigma_{O_{mt}}^2}}, \quad (8)$$

$$f(\mathbf{o}_t | g_t, h_t, s_t) = \prod_m f(o_{mt} | g_t, h_t, s_t). \quad (9)$$

To prevent the numerical problems, (9) is computed in the logarithmic domain. We approximate the state-conditional distribution of the mixed signal by taking a point estimate for the gain parameters (see [6, 8]), as:

$$f(\mathbf{o}_t | s_t) = \int \int f(\mathbf{o}_t, g_t, h_t | s_t) dg_t dh_t \approx f(\mathbf{o}_t | g'_t, h'_t, s_t). \quad (10)$$

In this letter, we use the mean values of the gain distributions as the point estimates, $g'_t = \phi\theta_t$, and $h'_t = \psi\gamma_t$. θ_t and γ_t represent the long-term speech and noise levels, respectively. As it is shown in [8], the MMSE estimator of the speech DFT coefficients is given by

$$E(\bar{\mathbf{O}}_t | \mathbf{o}_1^t) = \frac{\sum_{s_t} \zeta_t(s_t, \mathbf{o}_t) E(\bar{\mathbf{O}}_t | \mathbf{o}_t, g'_t, h'_t, s_t)}{\sum_{s_t} \zeta_t(s_t, \mathbf{o}_t)}, \quad (11)$$

where $\mathbf{o}_1^t = \{\mathbf{o}_1, \dots, \mathbf{o}_t\}$, and $\zeta_t(s_t, \mathbf{o}_t) = f(s_t | \mathbf{o}_1^{t-1})f(\mathbf{o}_t | g'_t, h'_t, s_t)$. Also,

$$f(s_t | \mathbf{o}_1^{t-1}) = \sum_{s_{t-1}} a_{s_{t-1}, s_t} f(s_{t-1} | \mathbf{o}_1^{t-1}), \quad (12)$$

with $a_{s_{t-1}, s_t} = \bar{a}_{\bar{s}_{t-1}, \bar{s}_t} \ddot{a}_{\ddot{s}_{t-1}, \ddot{s}_t}$, and $f(s_{t-1} | \mathbf{o}_1^{t-1})$ is the scaled forward variable. Due to the Gaussian assumptions, the state-conditional estimates of the speech DFT coefficients are obtained using a Wiener filter as:

$$E(\bar{O}_{mt} | \mathbf{o}_t, g'_t, h'_t, s_t) = \frac{E(X_{mt} | g'_t, \bar{s}_t) o_{mt}}{E(X_{mt} | g'_t, \bar{s}_t) + E(V_{mt} | h'_t, \ddot{s}_t)}, \quad (13)$$

in which (3) and (6) are used to compute the expected values. Although different functions of the speech DFT coefficients can also be estimated within the HMM framework [4], we used (13) here since it is a widely used reference method.

3.4 Speech Estimation: Erlang-Gamma Case

This subsection presents one of the main contributions of this letter where we derive new MMSE estimators using super-Gaussian prior distributions. We assume that the speech and noise band powers are additive, i.e. $\mathbf{Y}_t = \mathbf{X}_t + \mathbf{V}_t$, where the band powers are obtained similarly to (2), and $\mathbf{X}_t, \mathbf{V}_t$ are assumed to be independent. The additivity assumption is widely used in the literature to circumvent the difficulty of phase modeling. Here, we derive an MMSE estimator for \mathbf{X}_t given that both \mathbf{X} and \mathbf{V} are modeled using an HMM with gamma and Erlang output distributions, respectively. Again, denote the composite hidden state of the mixed signal \mathbf{Y} by \mathbf{S} with $S_t = s_t \in [1, \bar{N}\bar{N}]$. Although the conditional distribution $f(y_{kt} | g_t, h_t, s_t)$ is not exactly gamma, still a gamma distribution would be flexible enough to describe Y_{kt} practically, and we continue with this approximation for simplicity. Therefore, we follow a standard moment matching algorithm—up to second order, and considering that $E(Y_{kt} | g_t, h_t, s_t) = E(X_{kt} | g_t, \bar{s}_t) + E(V_{kt} | h_t, \ddot{s}_t)$, and $\text{var}(Y_{kt} | g_t, h_t, s_t) = \text{var}(X_{kt} | g_t, \bar{s}_t) + \text{var}(V_{kt} | h_t, \ddot{s}_t)$ —to obtain a gamma distribution to describe $f(y_{kt} | g_t, h_t, s_t)$. Then, the state-conditional distribution of the mixed signal is obtained using similar assumptions exploited in (9) and (10).

The MMSE estimate of the speech band powers can now be obtained similarly to (11). Since different speech band-powers are assumed to be conditionally independent, we now focus on obtaining $E(X_{kt} | y_{kt}, g'_t, h'_t, s_t)$. First, note that

$$f(y_{kt} | x_{kt}, h_t, s_t) = \begin{cases} f(V_{kt} = y_{kt} - x_{kt} | h_t, \ddot{s}_t) & y_{kt} \geq x_{kt} \\ 0 & y_{kt} < x_{kt} \end{cases} \quad (14)$$

Using Bayes rule, the MMSE estimate of X_{kt} is given as:

$$\hat{x}_{kt} = E(X_{kt} | y_{kt}, g'_t, h'_t, s_t) = \frac{\int_0^{y_{kt}} x_{kt} f(y_{kt} | x_{kt}, h'_t, s_t) f(x_{kt} | \bar{s}_t, g'_t) dx_{kt}}{\int_0^{y_{kt}} f(y_{kt} | x_{kt}, h'_t, s_t) f(x_{kt} | \bar{s}_t, g'_t) dx_{kt}}. \quad (15)$$

Exploiting (3), (6), and (14) in (15) yields:

$$\hat{x}_{kt} = \frac{\int_0^{y_{kt}} x_{kt}^{\alpha_{ki}} (y_{kt} - x_{kt})^{\beta_k - 1} e^{-\left(\frac{y_{kt} - x_{kt}}{h'_t c_{kj}} + \frac{x_{kt}}{g'_t b_{ki}}\right)} dx_{kt}}{\int_0^{y_{kt}} x_{kt}^{\alpha_{ki} - 1} (y_{kt} - x_{kt})^{\beta_k - 1} e^{-\left(\frac{y_{kt} - x_{kt}}{h'_t c_{kj}} + \frac{x_{kt}}{g'_t b_{ki}}\right)} dx_{kt}}, \quad (16)$$

where we have set $\bar{s}_t = i$ and $\bar{s}_t = j$ to keep notations uncluttered. Since V_{kt} is assumed to have an Erlang distribution, β_k is integer. Using the binomial theorem, we can write:

$$(y_{kt} - x_{kt})^{\beta_k - 1} = \sum_{l=0}^{\beta_k - 1} \binom{\beta_k - 1}{l} y_{kt}^{\beta_k - 1 - l} (-x_{kt})^l, \quad (17)$$

in which $\binom{\beta_k - 1}{l}$ is the binomial coefficient. Define $z_{k,ij} = 1/(g'_t b_{ki}) - 1/(h'_t c_{kj})$ and $\mathbf{a}_{kl} = (-1)^l \binom{\beta_k - 1}{l} y_{kt}^{\beta_k - 1 - l}$. Since the integration and summation are interchangeable, inserting (17) into (16) yields:

$$\hat{x}_{kt} = \frac{\sum_{l=0}^{\beta_k - 1} \mathbf{a}_{kl} \int_0^{y_{kt}} x_{kt}^{\alpha_{ki} + l} e^{-z_{k,ij} x_{kt}} dx_{kt}}{\sum_{l=0}^{\beta_k - 1} \mathbf{a}_{kl} \int_0^{y_{kt}} x_{kt}^{\alpha_{ki} + l - 1} e^{-z_{k,ij} x_{kt}} dx_{kt}}. \quad (18)$$

In the following, we discuss two special cases for which the integrals in (18) can be solved analytically. First, for positive $z_{k,ij}$, we obtain the subsequent closed-form expression:

$$\begin{aligned} \int_0^{y_{kt}} x_{kt}^{\alpha_{ki} + l} e^{-z_{k,ij} x_{kt}} dx_{kt} &= z_{k,ij}^{-(\alpha_{ki} + l + 1)} \int_0^{z_{k,ij} y_{kt}} u^{\alpha_{ki} + l} e^{-u} du \\ &= z_{k,ij}^{-(\alpha_{ki} + l + 1)} \gamma(\alpha_{ki} + l + 1, z_{k,ij} y_{kt}), \end{aligned} \quad (19)$$

where we have defined $u = z_{k,ij} x_{kt}$ and $\gamma(a, y)$ is the incomplete gamma function [9, eq. 8.350].

Second, when the speech shape parameters α_{ki} are integer-valued, we use [9, eq. 2.323] to get the following closed-form solution for the required integrations in (18):

$$\begin{aligned} \int_0^{y_{kt}} x_{kt}^{\alpha_{ki} + l} e^{-z_{k,ij} x_{kt}} dx_{kt} &= \frac{e^{-z_{k,ij} y_{kt}}}{-z_{k,ij}} \sum_{q=0}^{\alpha_{ki} + l} (-1)^q \frac{P^{(q)}(y_{kt})}{(-z_{k,ij})^q} \\ &\quad + \frac{1}{z_{k,ij}} (z_{k,ij})^{-l - \alpha_{ki}} P^{(\alpha_{ki} + l)}(0), \end{aligned} \quad (20)$$

where $P^q(y_{kt})$ is the q^{th} derivative of $x_{kt}^{\alpha_{kt}+l}$ with respect to x_{kt} , evaluated at y_{kt} .

If neither (19) nor (20) can be used to calculate (18), the integrals can be tabulated, or they can be computed online using the stochastic integrations.

The derived algorithm in this subsection provides an MMSE estimator for the speech band powers, \hat{x}_{kt} . To obtain an estimate of the speech spectral vectors in the original DFT resolution, we first obtain the gain function at the central frequencies of the bands as $\kappa_{kt} = \hat{x}_{kt}/y_{kt}$, and then interpolate this gain values to obtain the high resolution gain vector $\bar{\kappa}_{mt}$, and then speech DFT coefficients are estimated as $\bar{\kappa}_{mt}o_{mt}$.

4 Experiments and Results

The proposed speech enhancement strategies are evaluated and compared at different input signal to noise ratios (SNR) for different interfering noise types including babble, factory and highway traffic noises. The speech models are trained using the training data from the TIMIT database while babble and factory noises were taken from NOISEX-92, and highway traffic noise was taken from Sound-Ideas database. All of the signals were down-sampled to 16-kHz. The core test set of the TIMIT database (192 sentences) was exploited for the noise reduction evaluation, and the train and test segments of noises were disjoint. The signal synthesis was performed using the overlap-and-add procedure using a frame length of 320 samples with 50% overlapped windows and a Hann window. For the speech model $\bar{N} = 55$ states and for each noise type $\bar{N} = 10$ states were trained.

Two objective measures including source to distortion ratio (SDR) and perceptual evaluation of speech quality (PESQ) were considered for the evaluation. The SDR and PESQ improvements are averaged over all of the three noise types and the final scores are shown in Figure 2. Three algorithms are considered for comparison. Two algorithms are implemented directly in the high-resolution spectral domain, which are referred as: complex Gaussian (13) and exp-gamma (18 with $\beta_k = 1$). To evaluate (18), we used either (19) or (20) whenever possible, and if none of them were applicable, we calculated the integrals using the stochastic integrations. The other algorithm, referred as Erlang-gamma, is implemented in reduced resolution domain, for which (18) is used.

The presented results in Figure 2 show that the exp-gamma algorithm is clearly better than the complex Gaussian, in terms of both SDR and PESQ. Thus, the simulation results verify the observation from Section 2, and imply that the real and imaginary parts of the speech DFT coefficients are modeled better with super-Gaussian than with Gaussian distributions.

The results of the Mel-domain Erlang-gamma algorithm and the DFT domain exp-gamma algorithm are very close in the sense of PESQ, but exp-

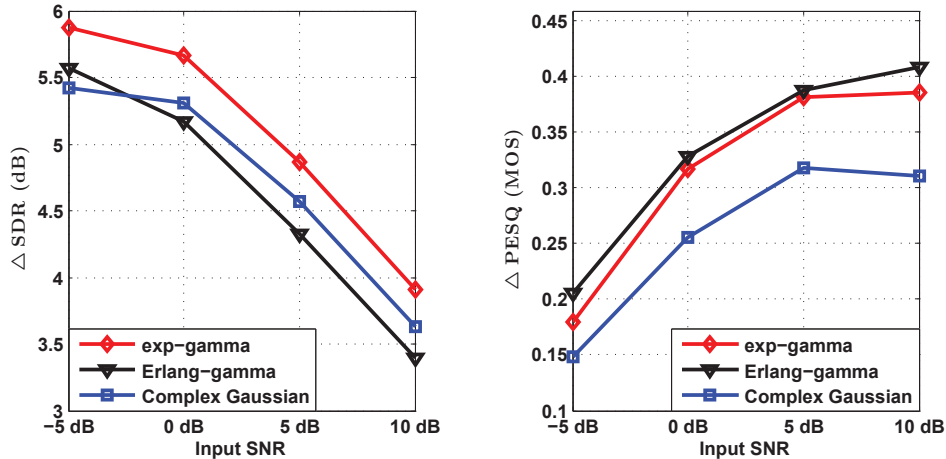


Figure 2: Performance of the proposed noise reduction algorithms averaged over different noise types.

gamma is superior considering SDR. The benefit of Mel-domain algorithms is that the random speech and noise fluctuations at different frequency bins are reduced and smoother signals are fed into the models. Also, the assumption of additive speech and noise band powers is more justified in this case. On the other hand, due to the reduced resolution, the filter estimation is less accurate. Informal listening test results were consistent with these objective results.

5 Conclusion

In this letter, we aim to investigate the distribution of the phoneme-conditioned speech power spectral coefficients. We looked at the empirical distribution of the periodogram coefficients for different phonemes, and also we derived new HMM-based speech spectral enhancement algorithms. The empirical distributions together with the simulation results of the denoising algorithms support our hypothesis that the power spectral coefficients will rather have gamma distributions with shape parameters less than one even at the scale of individual phones. For example, using a gamma assumption the source to distortion ratio was increased up to 0.8 dB compared to the exponential assumption. We also showed that this finding can be equivalently expressed as the super-Gaussianity of the DFT coefficients for different phonemes.

References

- [1] 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.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed., A. Jeffrey and D. Zwillinger, Eds. Academic Press, feb. 2007.

Paper E

Prediction Based Filtering and Smoothing to Exploit Temporal Dependencies in NMF

Nasser Mohammadiha, Paris Smaragdis, and Arne Leijon

Refereed paper published in
*Proceedings of IEEE International Conference on Acoustics, Speech, and
Signal Processing (ICASSP)*, may 2013, pp. 873–877.

©2013 IEEE
Layout has been revised for thesis consistency

Prediction Based Filtering and Smoothing to Exploit Temporal Dependencies in NMF

Nasser Mohammadiha, Paris Smaragdis, and Arne Leijon

Abstract

Nonnegative matrix factorization is an appealing technique for many audio applications. However, in its basic form it does not use temporal structure, which is an important source of information in speech processing. In this paper, we propose NMF-based filtering and smoothing algorithms that are related to Kalman filtering and smoothing. While our prediction step is similar to that of Kalman filtering, we develop a multiplicative update step which is more convenient for nonnegative data analysis and in line with existing NMF literature. The proposed smoothing approach introduces an unavoidable processing delay, but the filtering algorithm does not and can be readily used for on-line applications. Our experiments using the proposed algorithms show a significant improvement over the baseline NMF approaches. In the case of speech denoising with factory noise at 0 dB input SNR, the smoothing algorithm outperforms NMF with 3.2 dB in SDR and around 0.5 MOS in PESQ, likewise source separation experiments result in improved performance due to taking advantage of the temporal regularities in speech.

1 Introduction

Nonnegative matrix factorization (NMF) [1] is a technique that decomposes a nonnegative matrix into a product of two nonnegative matrices such that one contains basis vectors and the other contains activations. NMF can be seen as a feature extraction method that discovers a low-dimensional representation in terms of a set of basis vectors. When applied to speech or music spectrograms, NMF has been shown to produce promising results in different applications [2–4].

Since the basic NMF model ignores temporal correlations, different approaches have been used in the past to enhance the decomposition to model

time dependencies for audio signals. For example, Virtanen [2] used a regularization term in NMF, motivated by the temporal dependencies of speech signals, to develop a monaural sound source separation algorithm. A regularized NMF was also used in [5] where a heuristic regulation term was added to the NMF cost function that enforced temporal constraints as part of a noise reduction scheme.

In a recently developed class of approaches, NMF and the hidden Markov model (HMM) are combined to model the temporal aspects in the NMF [6–8]. In order to develop a blind source separation or speech enhancement algorithm in this case, the models for the two considered signals should be combined to form a factorial HMM. Therefore, even though these approaches are quite successful in modeling temporal dependencies, they are too computationally expensive for an on-line algorithm. Moreover, the temporal modeling in these methods cannot go beyond the first order Markov chain because of computational issues.

Bayesian NMF approaches can also provide an alternative way to derive more meaningful factorizations for audio signals. In [9], an on-line speech enhancement algorithm was proposed in which temporal aspects of the data were used to obtain informative prior distributions to be applied in a Bayesian NMF framework.

In this paper, we propose filtering and smoothing algorithms for NMF strategies that are motivated by Kalman filtering and smoothing. We assume that the NMF coefficients are stochastic processes, and that they evolve through a vector autoregressive (VAR) model over time. Therefore, in addition to the basis matrix, there will be some regression parameters associated with each signal. The proposed algorithm (for both filtering and smoothing) has two steps. First, we predict the current frame’s NMF coefficients given either past observations (in filtering) or both past and future observations (in smoothing), and second, we update the estimates given the current observation. We propose a multiplicative update step of the estimates that can be interpreted using the HMM terminology. The proposed scheme introduces a new way of thinking about the problem that has not been considered in the current literature. We demonstrate the strength of our method using both synthetic examples and real applications including denoising and speech source separation.

2 Proposed Method

In this section, we present the proposed approach for a probabilistic NMF in the context of probabilistic latent component analysis (PLCA) [10]. In Subsection 2.1, we review the basic PLCA model and define the required notations. The proposed approach is given in Subsection 2.2 for the filtering and in 2.3 for the smoothing problems, and finally, Subsection 2.4 illustrates

how we can process a mixed signal with these techniques.

2.1 Background

PLCA is a probabilistic formulation of NMF in which the distribution of an input vector is approximated as a convex combination of some weighted marginal distributions. A latent variable is defined to refer to the index of the underlying mixture component that has generated an observation, and the probabilities of different outcomes of this latent variable determine the weights in the mixture.

We denote the magnitude spectrogram of the speech by a random matrix \mathbf{X} with elements X_{ft} where f is the frequency index and t is the time index, and a particular realization by $\mathbf{x} = [x_{ft}]$. Also, we refer to the t -th column of \mathbf{X} by \mathbf{X}_t . The random vector \mathbf{X}_t is assumed to be distributed according to a multinomial distribution [11] whose parameter vector is denoted by $\boldsymbol{\theta}_t$, with the expected value given as: $E(\mathbf{X}_t) = \gamma_t \boldsymbol{\theta}_t$. Here, $\gamma_t = \sum_f x_{ft}$ is the total number of draws from the distribution at time t . The f -th element of $\boldsymbol{\theta}_t$ (θ_{ft}) indicates the probability that f -th row of \mathbf{X}_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 F possible frequency indices $f = 1, \dots, F$ as its outcome. The f -th element of $\boldsymbol{\theta}_t$ is now given as: $\theta_{ft} = p(\Phi_t = f)$. Also, let V_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, $p(\Phi_t = f)$ is given by

$$\theta_{ft} = p(\Phi_t = f) = \sum_{i=1}^I p(\Phi_t = f | V_t = i) p(V_t = i). \quad (1)$$

We define a coefficient matrix \mathbf{v} with elements $v_{it} = p(V_t = i)$, and a basis matrix \mathbf{b} with elements $b_{fi} = p(\Phi_t = f | V_t = i)$. In principle, \mathbf{b} is time-invariant and includes the possible spectral structures of the speech signal. Eq. (1) is now equivalently written as: $\boldsymbol{\theta}_t = \mathbf{b}\mathbf{v}_t$.

An observed spectrogram \mathbf{x} can be approximated as the expected value of the underlying multinomial distribution as $\mathbf{x}_t \approx E(\mathbf{X}_t) = \gamma_t \boldsymbol{\theta}_t$. Consequently, the nonnegative factorization is written as: $\mathbf{x}_t \approx \gamma_t \mathbf{b}\mathbf{v}_t$ or $\mathbf{x}_t = \gamma_t \mathbf{b}\mathbf{v}_t + \mathbf{w}_t$ where \mathbf{w}_t is an additive noise.

The basis and coefficient matrices (\mathbf{b} and \mathbf{v}) can be estimated using the expectation-maximization (EM) algorithm [11]. The iterative update rules are given by:

$$v_{it} \leftarrow \frac{v_{it} \sum_f b_{fi}(x_{ft}/\hat{x}_{ft})}{\sum_i v_{it} \sum_f b_{fi}(x_{ft}/\hat{x}_{ft})}, \quad (2)$$

$$b_{fi} \leftarrow \frac{b_{fi} \sum_t v_{it}(x_{ft}/\hat{x}_{ft})}{\sum_f b_{fi} \sum_t v_{it}(x_{ft}/\hat{x}_{ft})}, \quad (3)$$

where $\hat{\mathbf{x}}_t = \gamma_t \mathbf{b} \mathbf{v}_t$ is the model approximation that is updated after each iteration. Note that, given the basis matrix \mathbf{b} in (2), the update equation of \mathbf{v}_t is independent of all the other time instances. Therefore, the time dependencies can not be modeled using (2).

2.2 Filtering

The goal of the proposed filtering approach is to develop an on-line algorithm to estimate a coefficient vector \mathbf{v}_t given all the current and past observations, which are denoted by $\mathbf{x}_1^t = \{\mathbf{x}_1, \dots, \mathbf{x}_t\}$. Here, we assume that the basis matrix \mathbf{b} is obtained using some training data and is kept fixed thereafter. We assume that the coefficient vectors are modeled by an M -th order vector autoregressive (VAR) model as:

$$\mathbf{v}_t = \sum_{m=1}^M A_m \mathbf{v}_{t-m} + \mathbf{u}_t, \quad (4)$$

$$\mathbf{x}_t = \gamma_t \mathbf{b} \mathbf{v}_t + \mathbf{w}_t, \quad (5)$$

where A_m is the $I \times I$ autoregressive coefficient matrix associated with m -th lag, \mathbf{u}_t is the process noise, and \mathbf{w}_t is the observation noise in the model.

Even though (4) and (5) represent a complete state-space model that can be easily converted to a first order VAR model, nonnegativity of \mathbf{v}_t and \mathbf{x}_t prohibits the direct application of Kalman filtering. Following, we present an alternative approach that has a prediction and an update step as with Kalman filtering. The prediction of the coefficient vector \mathbf{v}_t , given \mathbf{x}_1^{t-1} , is denoted by $\hat{\mathbf{v}}_{t|t-1}$ and is simply obtained as:

$$\hat{\mathbf{v}}_{t|t-1} = \sum_{m=1}^M A_m \hat{\mathbf{v}}_{t-m|t-m}, \quad (6)$$

where $\hat{\mathbf{v}}_{t-m|t-m}$ is the updated estimate of \mathbf{v}_{t-m} given \mathbf{x}_1^{t-m} . To obtain the correction term, the basic PLCA model is applied to find $\tilde{\mathbf{v}}_t$ by iterating (2). Then, the updated estimate of \mathbf{v}_t is given by

$$\hat{\mathbf{v}}_{t|t} = \frac{(\hat{\mathbf{v}}_{t|t-1})^\beta \odot \tilde{\mathbf{v}}_t}{\sum_i (\hat{\mathbf{v}}_{t|t-1})^\beta \odot \tilde{\mathbf{v}}_t}, \quad (7)$$

where $(\cdot)^\beta$ and \odot denote element-wise power and product operators, respectively, β is the prior strength and might not be equal one, and the normalization is performed to ensure that $\hat{\mathbf{v}}_{t|t}$ is a probability vector. $\tilde{\mathbf{v}}_t$ is a probability vector where each of its elements is proportional to the similarity between the corresponding basis vector and the observation \mathbf{x}_t . The multiplicative update in (7) is similar to the forward algorithm in an

HMM, where the observation likelihood is replaced with $\tilde{\mathbf{v}}_t$. Therefore, $\hat{\mathbf{v}}_{t|t}$ can also be seen as the posterior probability of the latent variables (hidden states in the HMM).

The VAR coefficients $A_m, m = 1, \dots, M$, can be estimated in different ways (e.g., [12, ch. 11]). In this paper, we carry out a sub-optimal approach to estimate these matrices for simplicity. Let $\mathbf{v}^{(m)}$ denote the matrix \mathbf{v} , in which the columns are shifted by m , i.e. $v_{i,t}^{(m)} = v_{i,t+m}$. Then, A_m is estimated as $A_m = \mathbf{v}^{(m)} \mathbf{v}^\top$ where \top represents the matrix transpose. The columns of A_m are then normalized to sum to one, and hence, A_m^\top can also be interpreted as a transition matrix in a multimatrix mixture transition distribution (MTD) model [13].

2.3 Smoothing

The smoothing problem arises when we want to estimate a coefficient vector \mathbf{v}_t given both past and future data, i.e. $\mathbf{x}_1^T = \{\mathbf{x}_1, \dots, \mathbf{x}_t, \mathbf{x}_{t+1}, \dots, \mathbf{x}_T\}$, where T is the total number of observations. This estimate is referred to as $\hat{\mathbf{v}}_{t|T}$ (in contrast, the estimate using filtering was denoted by $\hat{\mathbf{v}}_{t|t}$ in (7)).

For this purpose, the PLCA algorithm is applied to \mathbf{x}_1^T to find the coefficient matrix $\tilde{\mathbf{v}}$. Then, a forward prediction matrix with columns given by $\hat{\mathbf{v}}_{t|t-1}$ and a backward prediction matrix with columns given by $\hat{\lambda}_{t|T}$ are obtained as:

$$\hat{\mathbf{v}}_{t|t-1} = \sum_{m=1}^M A_m \tilde{\mathbf{v}}_{t-m}, \quad (8)$$

$$\hat{\lambda}_{t|T} = \sum_{m=1}^M A_m^\top \tilde{\mathbf{v}}_{t+m}. \quad (9)$$

In principle, to evaluate (8) and (9) it suffices to have access to observations from $t - M$ through $t + M$. Therefore, the algorithm will introduce a delay of M short time frames. Since our estimation approach of the VAR model parameters makes A_m^\top similar to a transition matrix, (9) can be seen as an adaption of the HMM backward algorithm [14]. The updated estimate of \mathbf{v}_t is now given as:

$$\hat{\mathbf{v}}_{t|T} = \frac{\left(\hat{\lambda}_{t|T} \odot \hat{\mathbf{v}}_{t|t-1}\right)^\beta \odot \tilde{\mathbf{v}}_t}{\sum_i \left(\hat{\lambda}_{t|T} \odot \hat{\mathbf{v}}_{t|t-1}\right)^\beta \odot \tilde{\mathbf{v}}_t}. \quad (10)$$

2.4 Source Separation Using the Proposed Method

To separate unknown sources from a given mixture, we can learn the basis matrices and VAR coefficient matrices for all the involved sources off-line, and then concatenate them properly to model the mixed signal.

Denote the coefficient vector of the mixed signal by \mathbf{v}_t , which is estimated using (7) or (10). Let $\mathbf{x}_t \approx \sum_k \mathbf{s}_{k,t}$ be the observed mixture, where $\mathbf{s}_{k,t}$ represents the t -th column of the k -th source's spectrogram. The spectrogram of each source is estimated by

$$\hat{\mathbf{s}}_{k,t} = \frac{\mathbf{b}_{s_k} \mathbf{v}_{k,t}}{\sum_k \mathbf{b}_{s_k} \mathbf{v}_{k,t}} \odot \mathbf{x}_t, \quad (11)$$

where division is performed element-wise, \mathbf{b}_{s_k} is the basis matrix of the k -th source, and $\mathbf{v}_{k,t}$ is a coefficient vector that includes a subset of the elements of \mathbf{v}_t that are associated with \mathbf{b}_{s_k} . Eq. (11) is known as the Wiener reconstruction and is widely used with NMF-based source separation (e.g., [4]).

3 Experiments and Results

The proposed filtering/smoothing and the basic PLCA algorithms were applied to three different problems. In this section, we present the results and discuss the effect of different model parameters on the performance. We used the magnitude spectrogram of speech and noise signals as the input to the algorithms. The separated/enhanced time-domain signals were obtained using the phase of the mixed input signal and the overlap-add procedure. In our experiments here we consider three tasks: the separation of structured speech signals, speech denoising, and source separation.

3.1 Separation of Speech and Its Time-reversed Version

We applied the smoothing algorithm (10) to a mixed signal where the mixture was obtained as the sum of a temporally structured speech signal (see Figure 1) and its time-reversed version at a sampling rate of 8 kHz. The discrete Fourier transform (DFT) with a frame length of 128 ms, 75% overlap, and a Hann window was applied to obtain the magnitude spectrogram of the signals as the input to the NMF algorithms. 60 basis vectors were trained for each source and were used in PLCA and the proposed algorithm.

The top panels of Figure 1 show the spectrogram of the original signals. Since the basis matrices for the two source signals are effectively similar, basic PLCA or any other standard NMF algorithm will not be able to separate the sources. We see that by observing the separated sources which unfortunately closely resemble the mixture signal (see second row panels in Figure 1). The bottom panels of Figure 1 show the extracted source spectrograms using (10), which are obtained using parameters $M = 4$ and $\beta = 1$. Because there is a specific temporal structure that the two sources have (either ascending or descending pitch), we can tell the two sounds apart

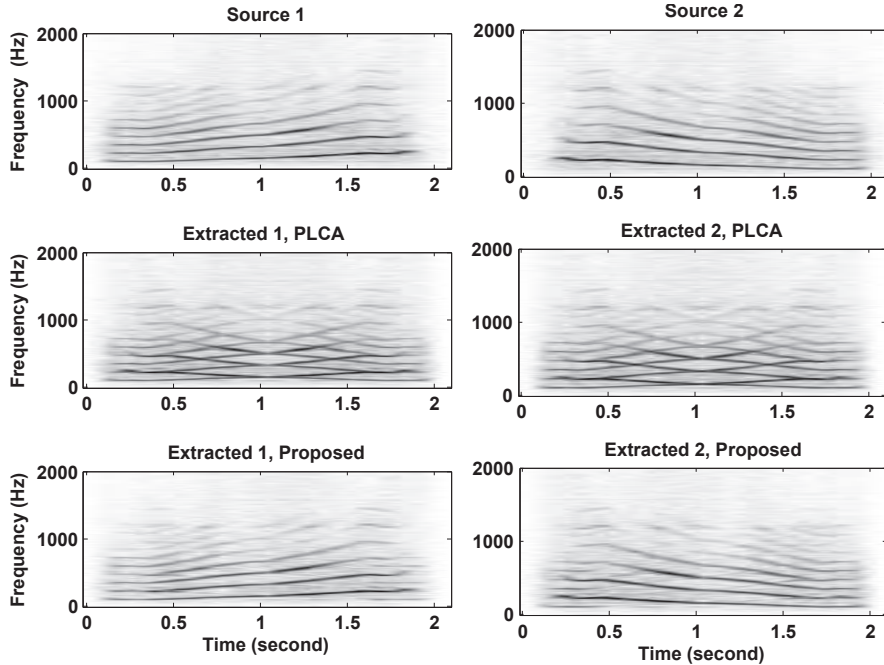


Figure 1: Magnitude spectrograms of the original inputs (top row), the separated sources using PLCA (middle row) and the separated sources using the proposed algorithm (bottom row). For legibility reasons we only show the frequency range 0 ~ 2 kHz.

despite the fact that they have spectrally identical basis matrices. This experiment verifies the benefit of temporal modeling in a difficult separation task. The separation performance in this case is around 11 dB improvement in source to distortion ratio (SDR) [15], while the basic PLCA leads to only 0.5 dB improvement, which is effectively no separation.

3.2 Speech Denoising

We consider a noise reduction application where the desired speech signal is corrupted by an additive noise. A speaker-dependent approach is followed here in which a separate basis matrix is trained for each speaker and each noise type beforehand. The experiment was done for 100 randomly chosen speakers with different genders from the TIMIT database [16], where 9 out of the 10 available sentences were used for training speech model and the other sentence was used for testing.

The denoising algorithms were evaluated for two babble and factory

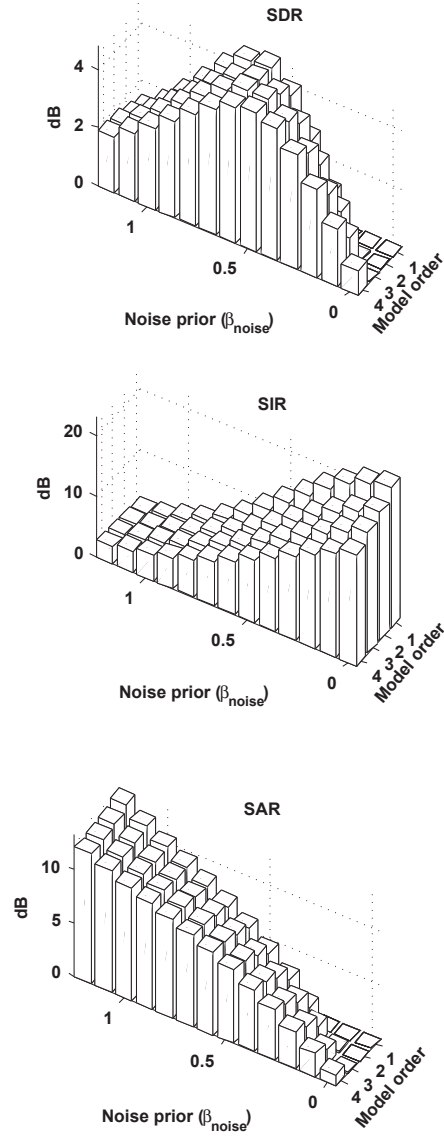


Figure 2: Effect of the VAR model order and noise prior strength on the performance of speech denoising with the smoothing algorithm.

noises taken from the NOISEX-92 database [17]. All the signals were down-sampled to 16 kHz. The frame length and overlap length in the DFT analysis were set to 64 and 60 ms, respectively. We learned 60 basis vectors for speech and 20 and 30 basis vectors for babble and factory noises, respectively.

First, we start by presenting an overall result of the denoising performances for both the smoothing and filtering algorithms. Since speech and noise signals have different temporal characteristics, we chose to use different powers (β) in (7) or (10) for speech (β_{speech}) and noise (β_{noise}) coefficients. These should be set experimentally, and we will discuss it shortly using Figure 2. The performance is measured using SDR, source to interference ratio (SIR), and source to artifact ratio (SAR) [15]. We also evaluated the perceptual quality of the enhanced speech using PESQ [18]. Figure 3 presents the results for a noisy signal at a 0 dB input signal to noise ratio (SNR), where we have used $M = 1$, $\beta_{\text{speech}} = 0.5$, $\beta_{\text{noise}} = 0.2$ for filtering, and $\beta_{\text{speech}} = 0.9$, $\beta_{\text{noise}} = 0.6$ for smoothing.

The results show a significant improvement in SDR, which results in better overall quality of demixed 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. Specifically in the case of the factory noise and with the smoothing algorithm, PESQ is improved by around 0.5 MOS compared to the baseline. Additionally, the figure illustrates that the smoothing algorithm has produced slightly better SDR and PESQ values than the filtering approach.

Finally, let us consider the smoothing approach applied to the babble case and study the effect of the model order (M) and prior strength (β) on the performance. Figure 2 shows three objective measures as functions of the model order ($M = 1, 2, 3, 4$) and noise prior strength (β_{noise}) while $\beta_{\text{speech}} = 0.9$. As the figure shows, increasing the model order from 1 to 4 has not changed the peak performance. However, it has made the algorithm more robust to the value of β_{noise} . Also, the previously used $\beta_{\text{noise}} = 0.6$ falls into the optimal range of β_{noise} .

3.3 Speech Source Separation

The last application we consider here is monaural speech source separation. We applied the proposed algorithms to 50 mixture signals for randomly-chosen different-gender speaker pair 0dB mixtures from the TIMIT database. The DFT analysis and the setting of model parameters including the number of speech basis vectors, M , and β_{speech} were done as described in Subsection 3.2.

Table 1 summarizes the results in terms of BSS-EVAL measures [15].

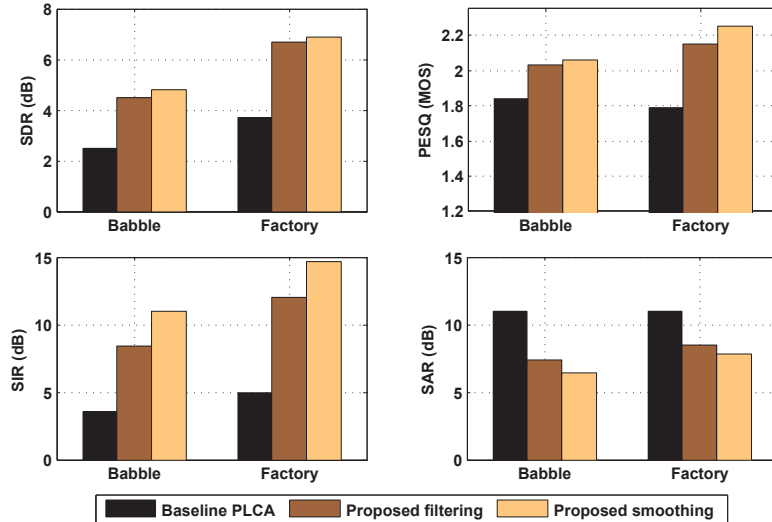


Figure 3: Performance of denoising algorithms for a noisy signal at a 0 dB input SNR.

Table 1: Performance of the algorithms for speech source separation.

Algorithm	SDR (dB)	SIR (dB)	SAR (dB)
Baseline PLCA	4.8	8.5	8.2
Filtering	5.5	11	7.8
Smoothing	5.7	12.5	7.5

Including the temporal dynamics has increased SIR but reduced SAR compared to the baseline. This is consistent with what was also observed in noise reduction in 3.2. In this case, the reduction in SAR is small and almost negligible while the SIR improvement is significant. Considering the SDR as a measure of overall speech quality, the evaluation shows that the performance has increased up to 0.9 dB due to the smoothing algorithm.

4 Conclusion

In this paper we introduced an approach to take advantage of temporal dependencies of sounds when performing NMF-style denoising and separation. Although we developed the algorithm using the PLCA terminology, adaptation of the scheme to NMF and its variants is straightforward. The

proposed two-step estimation approach for the NMF coefficients makes use of both temporal continuity and fidelity of an observation at a given time instant. We demonstrated the improvements that we obtained by the developed method in various applications using experimental means. Noticeably, we showed that our method can lead to improved results in source separation even when the basis matrices of the two underlying sources are practically the same. This allows us to attack mixture problems with sources that can be very similar in spectral characteristics and discernible only through their temporal structure.

References

- [1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Process. Systems (NIPS)*. MIT Press, 2000, pp. 556–562.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, 2008, pp. 4029–4032.
- [6] 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.
- [7] 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.

-
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] J. D. Hamilton, *Time Series Analysis*. New Jersey: Princeton University Press, 1994.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, no. 3, pp. 247–251, jul. 1993.
- [18] 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.

Paper F

Low-artifact Source Separation Using Probabilistic Latent Component Analysis

Nasser Mohammadiha, Paris Smaragdis, and Arne Leijon

Refereed paper to appear in
*Proceedings of IEEE Workshop on Applications of Signal Processing to
Audio and Acoustics (WASPAA)*, oct. 2013.

©2013 IEEE
Layout has been revised for thesis consistency

Low-artifact Source Separation Using Probabilistic Latent Component Analysis

Nasser Mohammadiha, Paris Smaragdis, and Arne Leijon

Abstract

We propose a method based on the probabilistic latent component analysis (PLCA) in which we use exponential distributions as priors to decrease the activity level of a given basis vector. A straightforward application of this method is when we try to extract a desired source from a mixture with low artifacts. For this purpose, we propose a maximum a posteriori (MAP) approach to identify the common basis vectors between two sources. A low-artifact estimate can now be obtained by using a constraint such that the common basis vectors in the interfering signal's dictionary tend to remain inactive. We discuss applications of this method in source separation with similar-gender speakers and in enhancing a speech signal that is contaminated with babble noise. Our simulations show that the proposed method not only reduces the artifacts but also increases the overall quality of the estimated signal.

1 Introduction

A popular class of dictionary learning approaches is nonnegative matrix factorization (NMF) in which nonnegative dictionaries are learned from the magnitude or power spectrogram of the speech signals denoted by \mathbf{X} . This factorization is written as $\mathbf{X} \approx \mathbf{B}\mathbf{V}$, where \mathbf{B} and \mathbf{V} are usually referred to as the basis and activation matrices. Since the basic NMF has many degrees of freedom, researchers have used different constraints to obtain more semantic factorizations with a better performance in a considered application [1–4].

In the probabilistic formulations of NMF, some prior distributions are considered over the basis or the activation matrices. These prior distributions may be motivated, e.g., by the temporal dependencies of the audio

signals. The goal of this prior information is to guide NMF by making some combination of the basis vectors more likely. For instance, to employ the time correlation of the audio signals, a constraint or a prior distribution is usually designed to govern the activations, e.g., [3]. Thus, having the result of the factorization at the current time instance, we put a prior over the activations for the next time instance that encourages the same pattern of activities as in the current time.

In this paper, we consider a source separation problem in which the underlying sources have some common basis vectors, i.e., some of the basis vectors are shared between two sources. 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 [5]. Consequently, these problems are among the hardest ones in the NMF-based approaches. As a result of having a common set of basis vectors, NMF can not correctly separate the sources and depending on the initial conditions one of the sources will steal some parts of the other one. This will lead to artifacts in the separated signals. In a denoising problem, we may prefer to reduce the noise as far as it does not introduce artifacts in the speech. Similarly for a source separation problem, we may have a preference over one of the speakers and then our goal would be to separate one of the sources with low artifacts. One approach to achieve this is to learn the basis matrix of the interfering signal such that its similarity with the known basis matrix of the target signal is minimized [6].

Another 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. To the best of our knowledge, there is not any work in the NMF community that investigates this solution. In this paper, we consider probabilistic latent component analysis (PLCA) [7] and propose an algorithm that can be used for this purpose.

In this paper, we argue 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 propose to use an exponential distribution as the prior and show that it can be used to force some basis vectors to be *inactive*. Moreover, we derive a MAP approach to identify a set of the common basis vectors and use that to separate a desired source with arbitrarily low artifacts. We demonstrate the application of this method in a simple toy example and also in speech denoising and speech source separation for speakers with same and different genders. Our experiments show that the presented approach leads to a higher quality for the estimated signal by reducing the artifacts.

2 Proposed Solution

In the following we first describe the basic PLCA approach. Then, we present our algorithm in which we use exponential distributions as priors for the activations. We discuss how this approach can be used to prevent (reduce) the activity of a given subset of the basis vectors. Additionally, we describe an approach to find a set of the common basis vectors between two underlying sources in Section 2.4. This information is then combined with the algorithm from Section 2.2 to design a source separation or speech enhancement algorithm in which we can recover a source with as low artifacts as desired.

2.1 PLCA: A Review

PLCA is a probabilistic nonnegative matrix factorization in which the speech magnitude spectrogram is modeled as a count data and is assumed to have a multinomial distribution:

$$\mathbf{x}_t \sim \text{Mult}(\boldsymbol{\theta}_t), \quad (1)$$

$$\theta_{ft} = p_t(f) = \sum_{z=1}^I p(f|z)p_t(z), \quad (2)$$

where \mathbf{x}_t is the vector of the DFT magnitudes at time frame t , f is the frequency index, θ_{ft} is the f -th element of $\boldsymbol{\theta}_t$, and z is the hidden variable that can take an integer value from $\{1 \dots I\}$. An NMF approximation of \mathbf{x}_t can be obtained as the expected value of its distribution:

$$\mathbf{x}_t \approx \hat{\mathbf{x}}_t = g_t \boldsymbol{\theta}_t, \quad (3)$$

where $g_t = \sum_f x_{ft}$. The set of the I probability vectors $p(f|z)$ are the basis vectors and can be found using an expectation-maximization approach. In the E step of the algorithm, the posterior probabilities of the hidden variables (z) are computed as:

$$p_t(z|f) = \frac{p(f|z)p_t(z)}{\sum_{z'=1}^I p(f|z')p_t(z')}. \quad (4)$$

In the M step of the algorithm the basis vectors and the weights are updated as:

$$p_t(z) = \frac{\sum_f x_{ft} p_t(z|f)}{\sum_{f,z'} x_{ft} p_t(z'|f)}, \quad (5)$$

$$p(f|z) = \frac{\sum_t x_{ft} p_t(z|f)}{\sum_{f',t} x_{f't} p_t(z|f')}. \quad (6)$$

2.2 PLCA with Exponential Priors

We can impose constraints on PLCA to use our a-priori knowledge. In this paper, we focus on prior distributions over the activations. Since Dirichlet distribution is the conjugate prior of the multinomial, we first give the update rules for this case. Let β_t be an I -dimensional vector with elements $\beta_{zt} = p_t(z)$. The Dirichlet prior for β_t is given as:

$$p(\beta_t) \propto \prod_{z=1}^I p_t(z)^{\alpha_z - 1}, \quad (7)$$

where $\alpha_z > 0$, $z \in \{1 \dots I\}$ are the parameters of the Dirichlet distribution. The E step of the algorithm (4) and the update rule of the basis vectors (6) remain the same as before. The update of the weights however changes to:

$$p_t(z) = \frac{\sum_f x_{ft} p_t(z | f) + \alpha_z - 1}{\lambda_t}, \quad (8)$$

where λ_t is a Lagrange multiplier and is used to ensure that $p_t(z)$ is a probability vector. Computing λ_t is trivial in this case and is given as the sum of the numerator of (8) over z .

The problem of the Dirichlet prior is that it does not naturally fit to the estimation of the nonnegative elements $p_t(z)$ and can lead to a negative value in the right hand side of (8). One way to avoid this problem is to put a threshold on (8) such that its minimum value is limited to be a very small positive number. Here, we propose to use an exponential distribution as the prior that does not suffer from this problem, and at the same time provides a single parameter to control the activity of each basis vector individually. The form of this prior is given by:

$$p(\beta_t) \propto \prod_{z=1}^I e^{-p_t(z)/\alpha_z}, \quad (9)$$

where $\alpha = \{\alpha_z\}$ is the vector of scale or inverse rate parameters. The update rule of the activations can be obtained by using the EM algorithm in which the M step is given by:

$$p_t(z) = \frac{\sum_f x_{ft} p_t(z | f)}{\lambda_t + 1/\alpha_z}. \quad (10)$$

Computation of the Lagrange multiplier λ_t is not trivial in (10) because the denominator is not the same for different latent components z as it was in (8). However, since λ_t is a scalar variable we can use a simple iterative algorithm, e.g., Newton's method, to find its optimal value.

In contrast to (8), (10) leads to nonnegative estimates for the activations for any value of the hyperparameters α . Now consider a simple problem

where we are interested to force a certain basis vector to be inactive. To do this using the Dirichlet priors, we have to boost the activity of all the other basis vectors since to ensure nonnegativity we should avoid choosing $\alpha_z < 1$. Using the exponential priors, we only need to use a small hyperparameter in the prior distribution corresponding to the given basis vector. In this case α_z can approach to 0 without making any theoretical problem.

2.3 Example: Separation of Sources with One Common Basis Vector

We consider a source separation problem to illustrate how the exponential distribution can be used to avoid certain type of activations. In this toy example, we generate 3-d nonnegative data for two sources. Let \mathbf{e}_i denote a 3-d indicator vector whose i -th element is 1 and the rest of its elements are zero. We considered two basis vectors per each source from which one is shared between the sources: $\mathbf{B}^{(1)} = [\mathbf{e}_1 \ \mathbf{e}_2]$ and $\mathbf{B}^{(2)} = [\mathbf{e}_2 \ \mathbf{e}_3]$. We also added small nonnegative random noise to the basis matrices. Data was generated by multiplying the bases by an activation vector with elements sampled from a uniform distribution in the interval $[0, 1]$. For our example, this procedure yielded to:

$$\mathbf{x}^{(1)} = \begin{bmatrix} 0.30 \\ 0.63 \\ 0.07 \end{bmatrix}, \mathbf{x}^{(2)} = \begin{bmatrix} 0.03 \\ 0.32 \\ 0.65 \end{bmatrix}. \quad (11)$$

These vectors together with the mixture

$$\mathbf{x} = \mathbf{x}^{(1)} + \mathbf{x}^{(2)} = \begin{bmatrix} 0.33 \\ 0.95 \\ 0.72 \end{bmatrix}, \quad (12)$$

are shown in Figure 1. Note that to be on the 2-d simplex, all of these vectors are normalized to sum to one. Applying our proposed method with $\boldsymbol{\alpha} = [1 \ 1 \ 0.5 \ 1]^1$ leads to the estimates which are shown in the figure. Numerically, we got:

$$\hat{\mathbf{x}}^{(1)} = \begin{bmatrix} 0.3 \\ 0.9 \\ 0.08 \end{bmatrix}, \hat{\mathbf{x}}^{(2)} = \begin{bmatrix} 0.03 \\ 0.05 \\ 0.64 \end{bmatrix}. \quad (13)$$

As we can see in Figure 1 also, first source has taken over and the second dimension of its estimate (0.93) is very close to the corresponding element

¹Each element in $\boldsymbol{\alpha}$ reflects our preference of having this basis vector active. If we set an element to a value smaller than the average of $\boldsymbol{\alpha}$, that basis vector is encouraged to be inactive. The choice of 0.5 was arbitrary in this example.

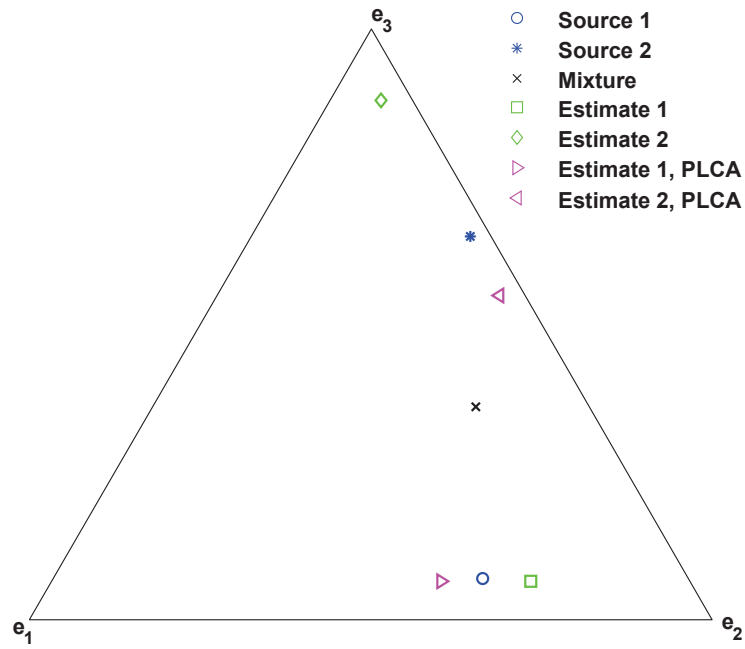


Figure 1: This example shows the original sources, the mixture and the estimated sources on a 2-d simplex. To be on the simplex, all the vectors are normalized to sum to one. e_i is a 3-d indicator vector whose i -th element is 1. Each source has two basis vectors from which e_2 is shared between them. Since the shared basis vector will introduce artifacts in the estimate of the desired source (estimate 1), a prior is constructed such that a big portion of the mixture's second element is taken as the corresponding component in "Estimate 1".

in the mixture (0.95). If we use PLCA here (with the same initial value for the activation of the similar bases), the second dimension will be divided almost equally between two estimates (see Figure 1), which means that we lose some part of the desired source.

2.4 Identifying Common Bases

We need to know which basis vectors are shared between sources to use the algorithm given in Section 2.2. In the following, we describe a maximum a-posteriori (MAP) approach to get this information. Let us assume that we have trained I_1 and I_2 basis vectors for the desired source (source 1) and the interfering source (source 2), respectively, using some training data. Our goal in this section is to develop an approach to identify a subset of basis vectors that belong to the interfering source and can also explain the desired source with a given accuracy. This subset can be actually seen as the common set of bases between two sources. By having this information, we can construct the vector α to automatically prevent the activity of this subset. This will reduce artifacts in the estimate of the desired source.

We start by concatenating small development sets of both of the sources (clean signals) as: $\mathbf{X} = [\mathbf{X}^{(1)} \mathbf{X}^{(2)}]$, where we have T_1 and T_2 observations (columns) in $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, respectively. Also, we concatenate the basis vectors of the two sources to obtain a larger basis matrix to explain both of the sources. We now apply PLCA to the concatenated signal \mathbf{X} with the basis matrix being fixed. The probability of choosing a basis vector, given the source, can be written as:

$$\begin{aligned} p(z | s^{(j)}) &= \sum_{t=1}^{T_1+T_2} p(z, t | s^{(j)}) \\ &= \sum_{t=1}^{T_1+T_2} p(z | t, s^{(j)}) \frac{p(s^{(j)} | t) p(t)}{p(s^{(j)})} \\ &= \frac{1}{T_j} \sum_{t \in T_{s^{(j)}}} p_t(z), z \in \{1 \dots I_1 + I_2\}, \end{aligned} \quad (14)$$

where $T_{s^{(j)}}$ includes the indices of the observations from $s^{(j)}$ and for these observations we have: $p(s^{(j)} | t) = 1$. To get the last line we have used a uniform distribution for t as $p(t) = 1/(T_1 + T_2)$, and also we have $p(s^{(j)}) = T_j/(T_1 + T_2)$.

To get a MAP classifier, we can use the Bayes' theorem (with a flat prior over the sources) to obtain the probability of each source given a basis vector. For $j = 1$, this results to:

$$p(s^{(1)} | z) = \frac{\sum_{t \in T_{s^{(1)}}} p_t(z) / T_1}{\sum_{t \in T_{s^{(1)}}} p_t(z) / T_1 + \sum_{t \in T_{s^{(2)}}} p_t(z) / T_2}. \quad (15)$$

To identify the basis vectors from the dictionary of $s^{(2)}$ that can also explain $s^{(1)}$, we now compare $p(s^{(1)} | z)$, $z \in \{I_1 + 1 \dots I_1 + I_2\}$ with a given threshold $0 \leq \gamma \leq 1$. If $p(s^{(1)} | z)$ was larger than γ , it means that this basis vector can also explain $s^{(1)}$ good enough. We should avoid the activity of this basis vector in a given mixture so that its similar basis vector that belongs to $s^{(1)}$ takes over and explains the mixture. $\gamma = 1$ recovers the basic PLCA, and $\gamma = 0$ corresponds to outputting the mixture signal as the estimate of the desired source. Thus, $\gamma = 0$ will neither suppress the interfering signal nor will introduce any artifacts in the final estimate.

3 Experiments Using Speech Data

We consider two problems to demonstrate the application of the proposed algorithm. In our experiments with the source separation and noise reduction, we used speech signals from the TIMIT and babble noise from the NOISEX-92 databases. Here, we considered instantaneous mixtures of sources where the mixed signal is obtained by adding the speech and noise waveforms, or by adding the speech signals of two speakers. All the signals were down-sampled to 16 kHz. The discrete Fourier transform (DFT) with a frame length of 64 ms, 50% overlap, and a Hann window was used in our simulations.

3.1 Source Separation

We learned 30 basis vectors using a set of training sentences for each speaker. The results presented here are averaged over 40 pairs of randomly-selected speakers. We aim to separate the speech signal from the first source with low artifacts. Hence, we first find a subset of the basis vectors belonging to the second source that can also explain the first source, and then we set the corresponding elements in α to 0.5. All the other elements are given a value of 1.

We study the performance of the algorithm for the same gender (male-male or female-female) and different gender scenarios. The performance of the separation is measured using the source to distortion ratio (SDR), source to interference ratio (SIR) and source to artifact ratio (SAR) [8]. The results are shown in Figure 2. A high value of SAR corresponds to a low-artifact estimate.

Our simulations show that by reducing the threshold (γ) we get a higher SAR in all scenarios. In fact, the lower we set the threshold, the more number of basis vectors (from the interfering source) are recognized as the common bases and we put a prior that motivates these basis vectors to remain inactive. As a result, the corresponding parts of the mixture signal

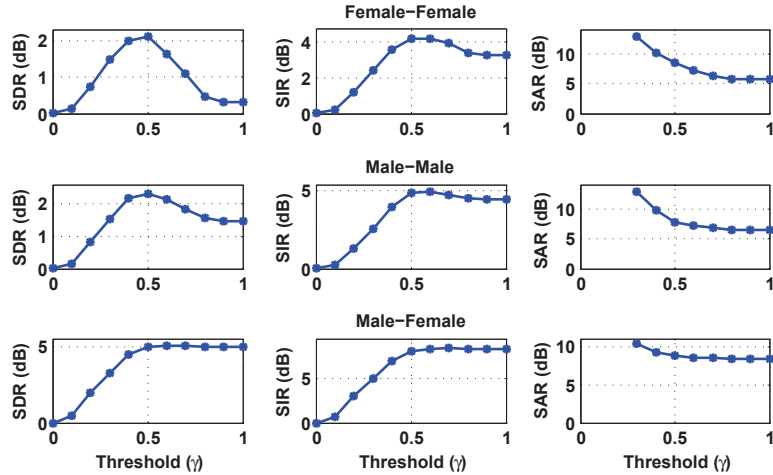


Figure 2: Results of source separation for female-female (top panel), male-male (middle panel), male-female (bottom panel) speakers. $\gamma = 1$ corresponds to the basic PLCA. Results show that by reducing the threshold a lower-artifact estimate is obtained for the desired source.

are taken in the favor of the desired source and we lose less and less of the desired signal.

Figure 2 shows that the performance is maximized in terms of SDR and SIR at a threshold equal to 0.5. For SIR, by increasing threshold we first get a higher suppression of the interfering signal. But when we set a very high value for the threshold, we get lower suppression. This might be explained by noting that with a proper value of γ , some shared basis vectors (of the interference signal) are inactive while the other basis vectors get higher activations, which results in a stronger suppression. Considering SDR, we again see that we get the best quality for $\gamma = 0.5$.

Another interesting result that can be seen in Figure 2 is that for the Male-Female configuration we do not get any improvement in SDR by using our algorithm. However, we can recover the desired source with lower artifacts. This is intuitive since we do not expect many common basis vectors in this case.

3.2 Reducing Babble Noise

As our second experiment, we consider a noise reduction problem where a speech signal is degraded by a babble noise. As discussed earlier, we expect the two sources to share some basis vectors. So we apply our method to

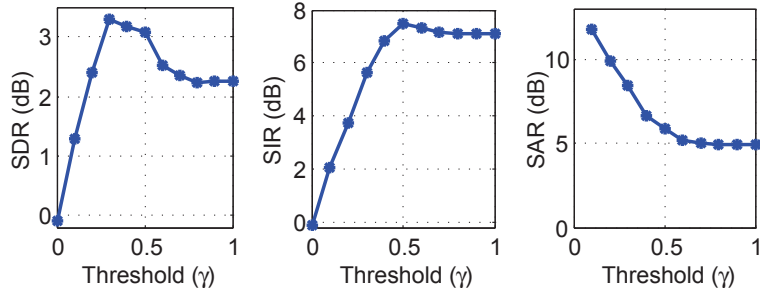


Figure 3: Speech denoising results for the babble noise. Similar to Figure 2, the lower the threshold is, the lower the artifacts are. A trade-off between noise suppression and artifact absence is obtained for $\gamma = 0.3$ where the SDR is maximum. Input SNR is 0 dB.

reconstruct the speech signal with low artifacts and better quality. Here, we learned 30 and 50 basis vectors for the speech and babble signals, respectively. The results are averaged over 40 speech signals from different speakers and are shown in Figure 3.

The experimental results are similar to the ones in Figure 2. Again, we see that the SAR value is reducing as a monotonic function of the threshold γ while SIR and SDR exhibit a maximum around $\gamma = 0.3 \sim 0.5$. Our informal listening tests were consistent with these results.

4 Conclusions

In this paper, we discussed a source separation problem in which the sources share some basis vectors. We proposed a PLCA-based approach to extract a desired source with an arbitrarily low artifacts. This was achieved by keeping the common basis vectors from the interfering source’s dictionary inactive. We developed a MAP approach to automatically detect the similar basis vectors. We considered applications of the proposed method in speech source separation and noise reduction. Our simulations show that when the underlying speakers have the same gender or the speech is contaminated with babble noise (for which we expect to see a sufficient number of common basis vectors) the proposed method can be used to reduce the artifacts and increase the quality in the estimate of the desired source.

References

- [1] 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.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] K. Yagi, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, “Music signal separation by orthogonality and maximum-distance constrained nonnegative matrix factorization with target signal information,” in *Proc. Audio Engineering Society Int. Conf.*, mar. 2012.
- [7] 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.
- [8] 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.