Online Parametric NMF for Speech Enhancement

Kavalekalam, Mathew Shaji; Nielsen, Jesper Kjær; Shi, Liming; Christensen, Mads Græsbøll; Boldt, Jesper

Published in:
2018 26th European Signal Processing Conference (EUSIPCO)

DOI (link to publication from Publisher):
10.23919/EUSIPCO.2018.8553039

Publication date:
2018

Document Version
Accepted author manuscript, peer reviewed version

Link to publication from Aalborg University

Citation for published version (APA):
Kavalekalam, M. S., Nielsen, J. K., Shi, L., Christensen, M. G., & Boldt, J. (2018). Online Parametric NMF for Speech Enhancement. In *2018 26th European Signal Processing Conference (EUSIPCO)* [8553039] IEEE. Proceedings of the European Signal Processing Conference https://doi.org/10.23919/EUSIPCO.2018.8553039

Online Parametric NMF for Speech Enhancement

Mathew Shaji Kavalekalam, Jesper Kjær Nielsen,
Liming Shi and Mads Græsbøll Christensen
Audio Analysis Lab, CREATE
Aalborg University
Aalborg, Denmark
{msk, jkn, ls, mgc}@create.aau.dk

Jesper Boldt
GN Hearing
Ballerup, Denmark
jboldt@gnresound.com

***Abstract*—In this paper, we propose a speech enhancement method based on non-negative matrix factorization (NMF) techniques. NMF techniques allow us to approximate the power spectral density (PSD) of the noisy signal as a weighted linear combination of trained speech and noise basis vectors arranged as the columns of a matrix. In this work, we propose to use basis vectors that are parameterised by autoregressive (AR) coefficients. Parametric representation of the spectral basis is beneficial as it can encompass the signal characteristics like, e.g. the speech production model. It is observed that the parametric representation of basis vectors is beneficial while performing online speech enhancement in low delay scenarios.**

***Index Terms*—autoregressive modelling, speech enhancement, NMF**

I. Introduction

A healthy human auditory system is capable of focusing on desired signal from a target source while ignoring background noise in a complex noisy environment. In comparison to a healthy auditory system, the auditory system of a hearing impaired person lacks this ability, leading to degradation in speech intelligibility. In such scenarios, a hearing impaired person often relies on speech enhancement algorithms present in a hearing aid. However, the performance of the current hearing aid technology in this aspect is limited [1]. Speech enhancement algorithms that have been developed can be mainly categorised into supervised and unsupervised methods. Some of the existing unsupervised methods are spectral subtraction methods [2], statistical model based methods [3] and subspace based methods [4]. Supervised methods generally use some amount of training data to estimate the model parameters corresponding to speech and noise. The model parameters are subsequently used for enhancement. Examples of supervised enhancement methods include codebook based methods [5], [6], NMF methods [7]–[9], hidden Markov model based methods [10], [11].

In this paper, we propose a speech enhancement method based on non-negative matrix factorization (NMF) techniques. NMF for source separation and speech enhancement has been previously proposed [7], [8]. NMF techniques allow us to approximate the power spectrum or the magnitude spectrum of the noisy signal as a weighted linear combination of trained speech and noise basis vectors arranged as the columns of a matrix. Generally the basis vectors used in NMF based speech enhancement are not constrained by any parameters. Parameterisation of the basis vectors in the field of music processing has been previously done in [12]. In [12], harmonic combs parametrised by the fundamental frequency was used as the basis vectors. This parametrisation was found to efficiently represent the music signal in comparison to the non parametric counterpart.

In this work, we propose to use basis vectors that are parametrised by autoregressive (AR) coefficients. This parametrisation allows representation of power spectral density (PSD) using a small set of parameters. Parametrisation by AR coefficients is motivated by the source filter model of speech production. This model describes speech components as a combination of a sound source (excitation signal produced by the vocal chords) and an AR filter which models the vocal tract. In this work, we show that if we model the observed data in the time domain as a sum of AR processes, the maximisation of the likelihood corresponds to performing NMF of the observed data into a basis matrix and activation coefficients, using Itakura-Saito (IS) divergence as the optimisation criterion. The IS divergence has been extensively used in speech and music processing due to its similarity to perceptual distance. The basis matrix here consists of AR spectral envelopes parameterised by AR coefficients, and the activation coefficients can be physically interpreted as the excitation variance of the noise that excites the AR filter parametrised by the AR coefficients. A benefit of parametrically representing the spectral basis, is that, it can be represented by a small set of parameters, which means that fewer parameters have to be trained a priori for performing on-line speech enhancement.

The remainder of this paper is organised as follows. Section II explains the signal model and formulates the problem mathematically. Training of the speech and noise spectral bases is explained in Section III. Section IV explains the on-line estimation of the activation coefficients corresponding to the spectral bases followed by the enhancement procedure using the Wiener filter. Sections V and VI give the experimental results and the conclusion respectively.

This work was supported by Innovations fund Denmark (Grant no: 99-2014-1).

II. Mathematical formulation

This section explains the signal model and mathematically formulates the problem. The noisy signal is expressed as

$$x(n) = s(n) + w(n) \tag{1}$$

where $s(n)$ is the clean speech and $w(n)$ is the noise signal. The objective of a speech enhancement system is to obtain an estimate of the clean speech signal from the noisy signal. A very popular method for estimating the clean speech signal is by applying a Wiener filter onto the noisy signal. Wiener filtering requires the knowledge of the speech and noise statistics. Since there is no direct access to either speech or noise in practical scenarios, these statistics have to be estimated from the noisy observation. As the speech and noise properties change over time, these statistics are generally time varying. The majority of the speech processing algorithms consider these statistics to be quasi-stationary. Thus, these statistics are assumed to be constant for short segments of time ($\approx$ 25 ms).

We now, explain the signal model used in the estimation of the statistics from a frame of noisy signal. It is assumed that a frame of noisy signal $\mathbf{x} = [x(0), \dots x(N-1)]^T$ can be represented as a sum of $U = U_s + U_w$ AR processes $\mathbf{c}_u$. This is mathematically written as

$$\mathbf{x} = \sum_{u=1}^{U} \mathbf{c}_u \; = \sum_{u=1}^{U_s} \mathbf{c}_u + \sum_{u=U_s+1}^{U} \mathbf{c}_u, \tag{2}$$

where the first U_s AR processes correspond to the speech signal and the remaining U_w AR processes correspond to the noise signal. Each of the AR process is expressed as a multivariate Gaussian [6] as shown below

$$\mathbf{c}_u \sim \mathcal{N}(\mathbf{0}, \, \sigma_u^2 \mathbf{Q}_u). \tag{3}$$

The gain normalised covariance matrix, $\mathbf{Q}_u$ can be asymptotically approximated as a circulant matrix which can be diagonalised using the Fourier transform as [13]

$$\mathbf{Q}_u = \mathbf{F}\mathbf{D}_u\mathbf{F}^H \tag{4}$$

where $\mathbf{F}$ is the DFT matrix defined as $[\mathbf{F}]_{k,n} = \frac{1}{\sqrt{N}} \exp(j2\pi nk/N), \quad n, k = 0 \dots N-1$ and

$$\mathbf{D}_u = (\mathbf{\Lambda}_u^H \mathbf{\Lambda}_u)^{-1}, \qquad \mathbf{\Lambda}_u = \mathrm{diag}(\sqrt{N}\mathbf{F}^H \begin{bmatrix} \mathbf{a}_u \\ \mathbf{0} \end{bmatrix}) \tag{5}$$

where $\mathbf{a}_u = [1, \, a_u(1) \dots a_u(P)]^T$ represents the vector of AR coefficients corresponding to u^{th} basis vector and P is the AR order. The likelihood as a function of U excitation variances and AR spectral envelopes are expressed as

$$p(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{D}) \sim \mathcal{N}(\mathbf{0}, \sum_{u=1}^{U} \sigma_u^2 \mathbf{Q}_u) \tag{6}$$

where $\boldsymbol{\sigma}$ represents the excitation variances corresponding to the U AR processes and $\mathbf{D}$ represents AR spectral envelopes corresponding to the U AR processes. In this paper, we are interested in the maximum likelihood (ML) estimation of activation coefficients $\boldsymbol{\sigma}$ given the noisy signal $\mathbf{x}$. Since, we are performing supervised enhancement here, we assume that the spectral basis are trained a priori, which is explained in Section III. Thus, in this work we only estimate the activation coefficients online while the basis vectors are assumed known. This is expressed mathematically as,

$$\boldsymbol{\sigma}_{est} = \underset{\boldsymbol{\sigma} \geq 0}{\arg\max} \quad p(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{D}). \tag{7}$$

To solve this, the logarithm of likelihood in (6) is written as

$$\begin{aligned} \ln p(\mathbf{x}|\boldsymbol{\sigma}, \; \mathbf{D}) = &-\frac{N}{2}\ln 2\pi + \ln\Big|\sum_{u=1}^{U} \sigma_u^2 \mathbf{F}\mathbf{D}_u\mathbf{F}^H\Big|^{-\frac{1}{2}} \\ &-\frac{1}{2}\mathbf{x}^T[\sum_{u=1}^{U} \sigma_u^2 \mathbf{F}\mathbf{D}_u\mathbf{F}^H]^{-1}\mathbf{x}. \end{aligned} \tag{8}$$

This is further simplified as

$$\begin{aligned} \ln p(\mathbf{x}|\boldsymbol{\sigma}, \; \mathbf{D}) = &-\frac{K}{2}\ln 2\pi + \ln\prod_{k=1}^{K}\Big(\sum_{u=1}^{U} \sigma_u^2 d_u(k)\Big)^{-\frac{1}{2}} \\ &-\frac{1}{2}\mathbf{x}^T\mathbf{F}[\sum_{u=1}^{U} \sigma_u^2 \mathbf{D}_u]^{-1}\mathbf{F}^H\mathbf{x} \end{aligned} \tag{9}$$

where $d_u(k)$ represents the k^{th} diagonal element of $\mathbf{D}_u$ and number of frequency indices $K = N$. Further simplifying,

$$\begin{aligned} \ln p(\mathbf{x}|\boldsymbol{\sigma}, \; \mathbf{D}) = &-\frac{K}{2}\ln 2\pi + \ln\prod_{k=1}^{K}\Big(\sum_{u=1}^{U} \hat{\Phi}_u(k)\Big)^{-\frac{1}{2}} \\ &-\frac{1}{2}\sum_{k=1}^{K}\frac{\Phi(k)}{\sum_{u=1}^{U}\hat{\Phi}_u(k)} \end{aligned} \tag{10}$$

where $\hat{\Phi}_u(k) = \sigma_u^2 d_u(k)$, $\Phi(k) = |X(k)|^2$ and $X(k) = \frac{1}{\sqrt{N}}\sum_{n=0}^{N-1} x(n)\exp(-j2\pi nk/N)$. Log-likelihood is then written as

$$\ln p(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{D}) = -\frac{K}{2}\ln 2\pi - \frac{1}{2}\sum_{k=1}^{K}\left(\frac{\Phi(k)}{\sum_{u=1}^{U}\hat{\Phi}_u(k)} + \ln\sum_{u=1}^{U}\hat{\Phi}_u(k)\right) \tag{11}$$

where

$$\sum_{u=1}^{U}\hat{\Phi}_u(k) = \sum_{u=1}^{U}\sigma_u^2 d_u(k) = \mathbf{d}_k\boldsymbol{\sigma} \tag{12}$$

where $\mathbf{d}_k = [d_1(k) \dots d_U(k)]$ and $\boldsymbol{\sigma} = [\sigma_1^2 \dots \sigma_U^2]^T$. Thus maximising the likelihood is equivalent to minimising the IS divergence between $\boldsymbol{\phi} = [\Phi(1) \dots \Phi(K)]^T$ and $\mathbf{D}\boldsymbol{\sigma}$ subject to $\Phi(k) > 0 \;\forall k$ where $\mathbf{D} = [\mathbf{d}_1^T \dots \mathbf{d}_K^T]^T$. In case we observe $V > 1$ frames, this corresponds to performing NMF of $\mathbf{\Phi} = [\boldsymbol{\phi}_1 \dots \boldsymbol{\phi}_v \dots \boldsymbol{\phi}_V]$ (where $\boldsymbol{\phi}_v = [\Phi_v(1) \dots \Phi_v(K)]^T$ contains the periodogram of the v^{th} frame) as

$$\mathbf{\Phi} \approx \underbrace{\begin{bmatrix} d_1(1) & \dots & d_U(1) \\ \vdots & \ddots & \vdots \\ d_1(K) & \dots & d_U(K) \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} \sigma_1^2(1) & \dots & \sigma_1^2(V) \\ \vdots & \ddots & \vdots \\ \sigma_U^2(1) & \dots & \sigma_U^2(V) \end{bmatrix}}_{\mathbf{\Sigma}}. \tag{13}$$

The first U_s columns of $\mathbf{D}$ corresponds to the spectral basis corresponding to the speech and the remaining U_w columns of $\mathbf{D}$ correspond to noise signal. The first U_s rows of $\boldsymbol{\Sigma}$ correspond to the activation coefficients for speech and the remaining U_w rows of $\boldsymbol{\Sigma}$ correspond to the activation coefficients corresponding to the noise signal, which leads to (13) being rewritten as,

$$\boldsymbol{\Phi} \approx [\mathbf{D}_s \, \mathbf{D}_w] \begin{bmatrix} \boldsymbol{\Sigma}_s \\ \boldsymbol{\Sigma}_w \end{bmatrix} = \mathbf{D}\boldsymbol{\Sigma}. \tag{14}$$

III. Training the Spectral Bases

This section explains the training of the basis vectors used for the construction of the basis matrix $\mathbf{D}$. In this work we use a parametric representation of the PSD [14] where the u^{th} spectral basis $\mathbf{d}_u = [d_u(1)...d_u(k)...d_u(K)]^T$ is represented as

$$d_u(k) = \frac{1}{\left|1 + \sum_{p=1}^{P} a_u(p) \exp(\frac{-j2\pi pk}{N})\right|^2}, \tag{15}$$

where $\{a_u(p)\}_{p=1}^{P}$ is the set of AR coefficients corresponding to the u^{th} basis vector. During the training stage, a speech and noise codebook is first computed using the generalised Lloyd algorithm [15] [16] [6]. The speech codebook and noise codebooks contain AR coefficients corresponding to the spectral envelopes of speech and noise. During the training process linear prediction coefficients (converted into line spectral frequency coefficients) are extracted from windowed frames, obtained from the training signal and passed as input to the vector quantiser[1]. Once the speech codebook and noise codebooks are created, the spectral envelopes corresponding to the speech AR coefficients ($\{\mathbf{a}_u\}_{u=1}^{U_s}$) and noise AR coefficients ($\{\mathbf{a}_u\}_{u=U_s+1}^{U}$) are computed using (15), and arranged as columns of $\mathbf{D}$. The spectral envelopes generated here are gain normalised, so they do not include the excitation variance. Fig. 1 shows a few examples of the trained speech and noise spectral envelopes.

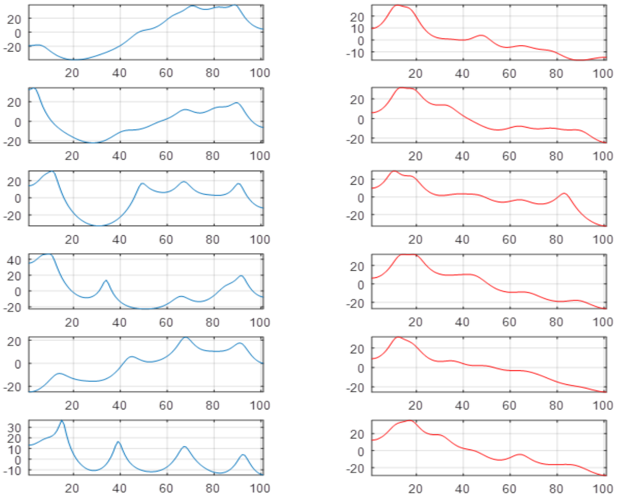
IV. Enhancement - Multiplicative Update

This section describes the estimation of speech and noise PSDs using the signal model explained in Section II. Since we are interested in on-line processing of the noisy signal, we here assume that only a frame of noisy signal is available at particular time for enhancement. The method considered here assumes that

$$\boldsymbol{\phi} \approx \mathbf{D}\boldsymbol{\sigma} \tag{16}$$

where $\boldsymbol{\phi}$ is a $K \times 1$ vector containing the noisy PSD, $\mathbf{D}$ is $K \times U$ basis matrix and $\boldsymbol{\sigma}$ is $U \times 1$ vector containing the activation coefficients. The objective here, is to estimate $\boldsymbol{\sigma}$ given the noisy periodogram $\boldsymbol{\phi}$ and $\mathbf{D}$. As explained in Section II, this is done by minimising the IS divergence as

$$\boldsymbol{\sigma}_{est} = [\boldsymbol{\sigma}_{s_{est}}^T \, \boldsymbol{\sigma}_{w_{est}}^T]^T = \underset{\boldsymbol{\sigma} \geq 0}{\arg\min} \quad d_{\text{IS}}(\boldsymbol{\phi}|\mathbf{D}\boldsymbol{\sigma}). \tag{17}$$



(a) (b)

Fig. 1: Figure showing a set of (a) trained speech spectral envelopes and (b) noise spectral envelopes.

In this work, a multiplicative update (MU) method is used to estimate the activation coefficients which are calculated as [8], [17]

$$\boldsymbol{\sigma}_{est} \leftarrow \boldsymbol{\sigma}_{est} \frac{\mathbf{D}^T((\mathbf{D}\boldsymbol{\sigma}_{est})^{[-2]}.\boldsymbol{\phi})}{\mathbf{D}^T(\mathbf{D}\boldsymbol{\sigma}_{est})^{[-1]}}. \tag{18}$$

Once the gains are estimated, a Wiener filter can be constructed to extract the speech/noise components. The estimated clean speech PSD is obtained as $\mathbf{D}_s\boldsymbol{\sigma}_{s_{est}}$ and the estimated noise PSD is obtained as $\mathbf{D}_w\boldsymbol{\sigma}_{w_{est}}$. The Wiener filter vector constructed to extract the speech component is denoted as

$$\mathbf{g}_{est} = \frac{\mathbf{D}_s\boldsymbol{\sigma}_{s_{est}}}{\mathbf{D}_s\boldsymbol{\sigma}_{s_{est}} + \mathbf{D}_w\boldsymbol{\sigma}_{w_{est}}}, \tag{19}$$

where the division is an element wise division.

V. Experiments

A. Implementation Details

This section explains the experiments that have been carried out to evaluate the proposed enhancement framework. The test signals used here consist of sentences taken from the GRID database [18]. The speech and noise PSD parameters are estimated (as explained in Section IV) for a segment of 25 ms with 50 percent overlap. The parameters used for the experiments are summarised in table I. For our experiments, we have used both a speaker-specific codebook and a general speech codebook. A speaker-specific codebook of 64 entries was trained using a training sample of 5 minutes of speech from the specific speaker of interest. A general speech codebook of 64 entries was trained from a training sample of approximately 150 minutes of speech from 30 different speakers. It should be noted that the sentences used for training the codebook were not included for testing. The

[1] The code for training the speech and noise codebooks will be available at `https://tinyurl.com/mskcreatevbn`

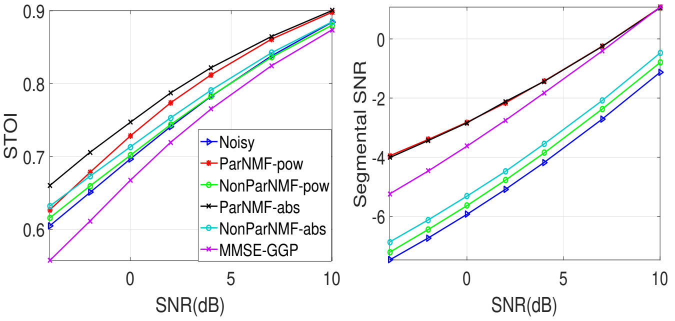
proposed enhancement framework was tested on three different types of commonly encountered background noise: babble, restaurant and exhibition noise taken from the NOIZEUS database [19]. We have performed experiments for a noise specific codebook as well as general noise codebook. A noise-specific codebook of 8 entries was trained on the specific noise type of interest. For creating a general noise codebook, a noise codebook of 4 entries was trained for each noise type. While testing for a particular noise scenario, the noise codebook entries corresponding to that scenario are not used for the estimation of noise PSD. For example, while testing in the babble noise scenario, the noise codebook consists a total of 8 entries formed by concatenating the entries trained for restaurant and exhibition scenarios. After obtaining the speech and noise codebooks, the spectral basis matrix is constructed as explained in Section III. The estimated PSD parameters are then used to create a Wiener filter for speech enhancement. Wiener filter is applied in the frequency domain and time-domain enhanced signal is synthesised using overlap-add.

B. Results

We have used the objective measures such as STOI and Segmental SNR to evaluate the proposed algorithm. We will denote the proposed parametric NMF as ParNMF. We have compared the performance of the proposed method to non parametric NMF where there is no parametrisation involved in the creation of the basis vectors. We will denote this method as NonParNMF. It should be noted that we have used the same training set for ParNMF and NonParNMF. We have also used the speech enhancement method proposed in [20] for comparison purposes, which we denote as MMSE-GGP. Traditionally, NMF methods for speech enhancement generally try to approximate the magnitude spectrum than the power spectrum. Even though, this is not theoretically well formulated, this has been observed to give better performance [21]. Thus, here we evaluated the performance of the proposed algorithm for both the cases, which we denote as ParNMF-abs while approximating the magnitude spectrum and ParNMF-pow while approximating the power spectrum. We do the same evaluation in the case of NonParNMF. Figures 2-4 show these measures for different methods in different commonly encountered background noises while using a speaker specific codebook and a noise specific codebook. It can be seen that NMF based methods perform better than MMSE-GGP in terms of STOI. When comparing the ParNMF and NonParNMF, it is demonstrated that the former performs better in terms of STOI and Segmental SNR measures. We have also performed experiments when having an access to a general speech codebook and a general noise codebook. Figures 5-7 shows the objective measures obtained for this case. It can be seen that performance in this case degrades in comparison to figures 2-4 due to the mismatch in training and testing conditions. Even though there is a degradation in the performance, the proposed method is able to increase the STOI measure significantly over the conventional method.

TABLE I: Parameters used for the experiments

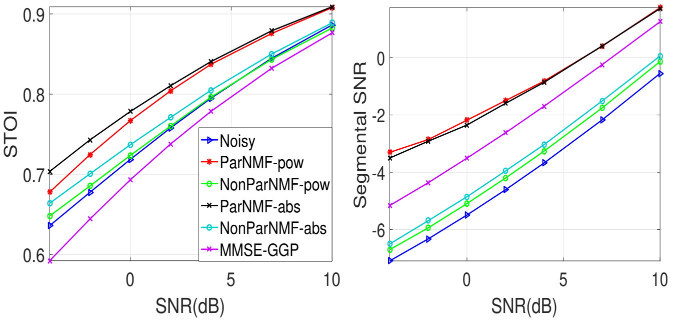
Parameters	
sampling frequency	8000 Hz
Frame Size	200
Frame Overlap	50%
Speech AR order	14
Noise AR order	14
U_s	64
U_w	8
MU iterations	50



(a) STOI

(b) Segmental SNR

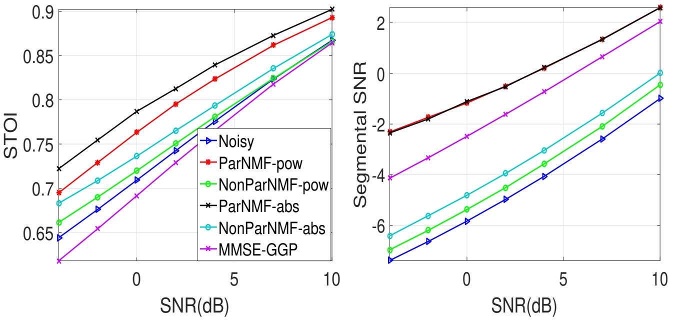
Fig. 2: Objective measures for babble noise when using speaker-specific codebook and a noise-specific codebook.



(a) STOI

(b) Segmental SNR

Fig. 3: Objective measures for restaurant noise when using speaker-specific codebook and a noise-specific codebook.



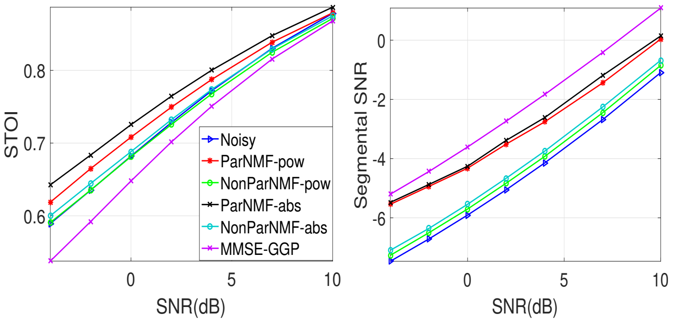
(a) STOI

(b) Segmental SNR

Fig. 4: Objective measures for exhibition noise when using speaker-specific codebook and a noise-specific codebook.

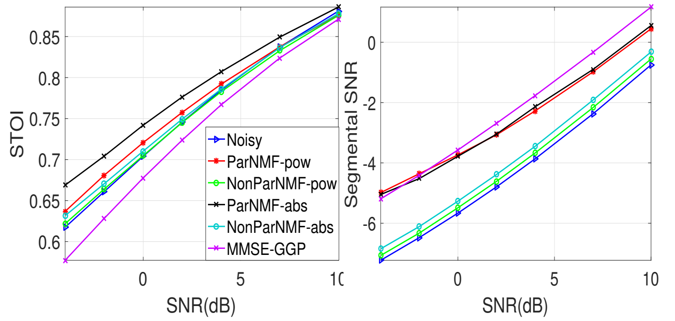
VI. Conclusion

In this paper, we have proposed an NMF based speech enhancement method where the basis vectors are parametrised using AR coefficients. Parametrisation of the spectral basis vectors helps in encompassing the signal characterestics. We have demonstrated, through objective measures, that the proposed parametric NMF based speech enhancement out performs its non-parametric counterpart in some of the typically encountered background noises.



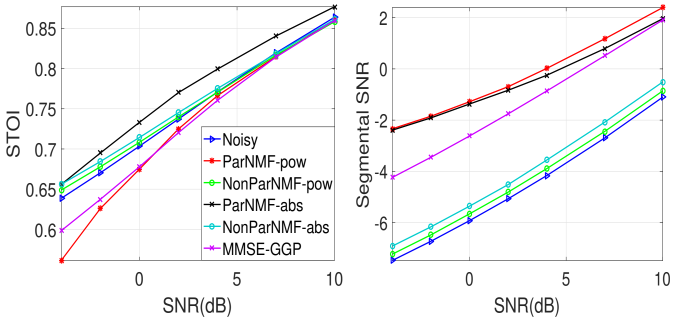
(a) STOI (b) Segmental SNR

Fig. 5: Objective measures for babble noise when using general speech codebook and a general noise codebook.



(a) STOI (b) Segmental SNR

Fig. 6: Objective measures for restaurant noise when using general speech codebook and a general noise codebook.



(a) STOI (b) Segmental SNR

Fig. 7: Objective measures for exhibition noise when using general speech codebook and a general noise codebook.

References

[1] S. Kochkin, "10-year customer satisfaction trends in the US hearing instrument market," *Hearing Review*, vol. 9, no. 10, pp. 14–25, 2002.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," vol. 27, no. 2, pp. 113–120, 1979.

[3] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

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

[5] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," vol. 14, no. 1, pp. 163–176, 2006.

[6] ——, "Codebook-based bayesian speech enhancement for nonstationary environments," vol. 15, no. 2, pp. 441–452, 2007.

[7] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[8] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[9] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low snr conditions via noise estimation using sparse and low-rank nmf with kullback–leibler divergence," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 7, pp. 1233–1242, 2015.

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

[11] D. Y. Zhao and W. B. Kleijn, "Hmm-based gain modeling for enhancement of speech in noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 882–892, 2007.

[12] R. Hennequin, R. Badeau, and B. David, "Time-dependent parametric and harmonic templates in non-negative matrix factorization," in *Proc. of the 13th International Conference on Digital Audio Effects (DAFx)*, 2010.

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

[14] P. Stoica, R. L. Moses *et al.*, *Spectral analysis of signals*. Pearson Prentice Hall Upper Saddle River, NJ, 2005, vol. 452.

[15] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[16] M. S. Kavalekalam, M. G. Christensen, F. Gran, and J. B. Boldt, "Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 191–195.

[17] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[18] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," *IEEE 2015 Automatic Speech Recognition and Understanding Workshop*, 2015.

[19] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech communication*, vol. 49, no. 7, pp. 588–601, 2007.

[20] 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 Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, 2007.

[21] A. Liutkus and R. Badeau, "Generalized wiener filtering with fractional power spectrograms," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 266–270.