NMF-Based Speech Enhancement Using Multitaper Spectrum Estimation

Yazid Attabi, Hanwook Chung, Benoit Champagne
Dept. of Electrical and Computer Engineering
McGill University
Montreal, Quebec, Canada H3A 0E9
yattabi@hotmail.com, hanwook.chung@mail.mcgill.ca
benoit.champagne@mcgill.ca

Wei-Ping Zhu
Dept. of Electrical and Computer Engineering
Concordia University
Montreal, Quebec, Canada H3G 1M8
weiping@ece.concordia.ca

***Abstract*—In the nonnegative matrix factorization (NMF) method, the clean speech dictionary, noise dictionary, and the activation matrix represent the key elements that greatly affect the speech enhancement performance. Current implementation of this framework employs the magnitude (or the power) spectrum coefficients computed from the short-time Fourier transform (STFT) as input features, which are characterized by an increased estimator variance. In this paper, we investigate the effects of using multitaper spectral estimation in the learning of the NMF speech/noise dictionary and activation matrix, on the quality of the Wiener filter used to predict the clean speech spectrum. The evaluation of the proposed method for various types of noise and input SNR show substantial speech quality improvements in terms of PESQ, SDR and SSNR measures compared to the NMF system based on conventional windowed periodogram estimation. We also find that the noise dictionary and the activation matrix of the noisy speech are the most important elements that benefit from the multitapering approach in the NMF system.**

Keywords—Single-channel speech enhancement, non-negative matrix factorization, multitaper, low-variance spectrum estimate

I. INTRODUCTION

Speech enhancement aims to isolate clean speech from a noisy background when only noisy speech is available, in order to improve its quality and/or intelligibility. Several algorithms for single-channel speech enhancement have been proposed in the past [1]-[4]. Recently the nonnegative matrix factorization (NMF) approach has been successfully applied to source separation [5] and speech enhancement [6]. NMF is typically used as a dimensionality reduction tool, which decomposes a given matrix into the product of a basis matrix (also known as dictionary) and an activation matrix (or encoding) with non-negative elements constraint [7]-[8]. In audio and speech applications, the magnitude or power spectrum of the signal is interpreted as a linear combination of the basis vectors, which play a key role in the enhancement or separation process. The NMF method is known for its capability to recover clean speech from noisy observations without the stationarity assumption on the nature of the noise [9][10].

Notable advances have been made more recently regarding different modules of the NMF-based speech enhancement system, such as : (i) its feature extractor module, e.g., the segmentation of the full-band speech into sub-bands [14], (ii) its structure, e.g., the use of a mixture of local dictionaries [11] and the introduction of the concept of deep NMF architecture [12], (iii) its training algorithm, e.g., the use of discriminative training criteria to simultaneously learn the basis vectors of the clean speech and noise sources [13], (iv) its enhancement algorithm, e.g., on-line update of the speech and noise bases [9] to overcome the shortcoming of the time-varying noise environments with NMF-based gain function and, (v) its use as feature extractor into a DNN framework [15]-[16]. These propositions led to important improvements in speech quality of the corrupted speech signal. Given the various studies that have shown a link between the quality of the enhanced speech and the spectral estimation [19], we can, however, expect that the sound quality can be further improved if more robust features are extracted to feed the NMF system.

The most common features used for NMF are the magnitude or power spectrum coefficients computed from the short-time Fourier transform (STFT) of overlapped and windowed speech frames. Despite having low bias, a consequence of the windowing is an increased estimator variance [17]. It should be recalled that musical noise, which is introduced by most of the speech enhancement algorithms, is due to the inaccurate and large-variance estimates of the spectra of the noise and noisy signals [18]. Furthermore, a small decrease in the bias and variance of the estimator can greatly reduce the occurrence of musical noise and distortion in the recovered speech [19]-[20].

A convenient way for reducing the spectral variance is to replace a windowed periodogram estimate with a multiple windowed (or multitaper) spectrum estimate [21]-[22]. In this method, a set of orthogonal tapers is applied to the short-time speech signal and the resulting spectral estimates are averaged, which reduces the spectral variance. The multitapering method was shown to have small bias and variance particularly for spectra with high dynamic range and/or rapid variations [21]. The multitaper method has been widely used in geophysical applications and more recently, in speaker [23], speech [24], and emotion [25] recognition to extract more robust features, where it has been shown to outperform the windowed periodogram. For the speech enhancement task, the multitaper spectral estimation method has been mainly used in combination with the wavelet thresholding technique [18], [26]-[27] and, to the best of our knowledge, multitapering has not been yet investigated for NMF method.

Funding for this work was provided by a CRD grant from NSERC (Govt. of Canada) with sponsorhip from Microsemi (Ottawa, Canada).

978-1-5386-5689-1/18/$31.00 ©2018 IEEE

In this study, we explain how the multitaper spectrum estimator can be incorporated into the NMF framework. We also show that the Wiener filter used in the clean speech spectrum prediction, can be better estimated in the NMF context, when the noise dictionary and the activation matrix are learned using multitaper power spectral estimator. The computed Wiener filter based on multitaper yields a better estimation of clean speech when compared to a single windowed periodogram method, under a wide variety of noise types and SNR conditions as shown by speech enhancement performances evaluated on TSP [28] and NOISEX [29] datasets using three objective measures.

II. NMF-BASED SPEECH ENHANCEMENT

In single-channel speech enhancement, the time-domain noisy speech $y(j)$ is composed of the clean speech signal $s(j)$ and the additive noise signal $n(j)$, that is,

$$y(j) = s(j) + n(j) \tag{1}$$

where j is the sample index. The noisy speech spectrum, obtained via STFT, can be expressed as $Y(k,l) = S(k,l) + N(k,l)$, where l represents the frame index, $k = 0, \cdots, K'-1$, the frequency bin index[1], $K' = K/2$, and, K the frame size. In NMF-based speech enhancement, we assume in practice that the magnitude spectrum of the noisy speech, obtained via STFT, can be approximated by the sum of the clean speech and noise magnitude spectra[2], i.e., $|Y(k,l)|^\nu \approx |S(k,l)|^\nu + |N(k,l)|^\nu$ with $\nu = 1$ [5]-[6]. For a nonnegative matrix $\mathbf{V} = [v_{kl}] \in \mathbb{R}_+^{K' \times L}$, NMF aims to find a local optimal decomposition of $\mathbf{V} = \mathbf{WH}$, where $\mathbf{W} = [w_{km}] \in \mathbb{R}_+^{K' \times M}$ is a basis matrix, $\mathbf{H} = [h_{ml}] \in \mathbb{R}_+^{M \times L}$ is an activation matrix, $\mathbb{R}_+$ denotes the set of nonnegative real numbers, M is the number of basis vectors and L is the number of consecutive and overlapping frames. The factorization is obtained by minimizing the reconstruction error between the observation matrix $\mathbf{V}$ and the model $\mathbf{WH}$ using a cost function, such as the Kullback-Leibler (KL) divergence, while constraining the matrices to be entry-wise nonnegative. The solutions can be obtained iteratively using the multiplicative update rules [7],

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{(\mathbf{V}/\mathbf{WH})\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T}, \mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T(\mathbf{V}/\mathbf{WH})}{\mathbf{W}^T\mathbf{1}} \tag{2}$$

where the operation $\otimes$ denotes element-wise multiplication, / and the quotient line are element-wise division, $\mathbf{1}$ is a $K' \times L$ matrix with ones, and the superscript T is the matrix transpose. In this work, $\mathbf{V} = [v_{kl}]$ contains the magnitude spectrum values of either one of the noisy speech, clean speech and noise, as indicated by subscripts or superscripts Y, S, and N, respectively.

In a supervised framework, the $\mathbf{W}$ matrices of clean speech and noise, denoted as $\mathbf{W}_S$ and $\mathbf{W}_N$ respectively, are obtained first during the training stage, by applying (2) to the training data $\mathbf{V}_S$ and $\mathbf{V}_N$. In the enhancement stage, the activation matrix $\mathbf{H}_Y = [\mathbf{H}_S^T \mathbf{H}_N^T]^T$ is estimated by applying the activation update to $\mathbf{V}_Y$, while fixing $\mathbf{W}_Y = [\mathbf{W}_S \mathbf{W}_N]$. Then, the clean speech spectrum can be estimated using a Wiener filter as [8]-[9],

$$\hat{S}_{kl} = \frac{\hat{P}_{kl}^S}{\hat{P}_{kl}^S + \hat{P}_{kl}^N} Y_{kl} \tag{3}$$

where $\hat{P}_{kl}^S$ and $\hat{P}_{kl}^N$ denote the estimated power spectral densities (PSD) of the clean speech and noise. The latter are obtained via temporal smoothing of the NMF-based periodograms as [9]

$$\begin{aligned} \hat{P}_{kl}^S &= \tau_s \hat{P}_{k,l-1}^S + (1-\tau_s)([\mathbf{W}_S\mathbf{H}_S]_{kl})^2 \\ \hat{P}_{kl}^N &= \tau_N \hat{P}_{k,l-1}^N + (1-\tau_N)([\mathbf{W}_N\mathbf{H}_N]_{kl})^2 \end{aligned} \tag{4}$$

where τ_s and τ_N are the smoothing factors for the speech and noise, and $[\cdot]_{kl}$ denotes the $(k,l)th$ entry of its matrix argument.

III. MULTITAPER SPECTRUM ESTIMATION

The power spectrum is often estimated using a windowed spectrum estimator applied to the time-domain signal $x(j)$ (x represents either the clean speech s, the noise n, or the noisy speech y). For the l-th frame and k-th frequency bin, an estimate of the windowed periodogram (called also single-taper) can be formulated as,

$$\hat{P}_{kl}^X = \left| \sum_{j=0}^{K-1} w(j)x(j)e^{-\frac{i2\pi jk}{K}} \right|^2 \tag{5}$$

where $w(j)$ is the windowing function such as the *Hann* or *Hamming* window.

Windowing reduces the bias but does not reduce the variance of the spectral estimate [28]. A set of P orthogonal tapers is applied to the short-time signal to smooth the spectral estimates for the reduction of spectral variance [18], [21]-[22]. At best, the variance of the multitaper estimate will be smaller than the variance of each spectral estimate by a factor of $1/P$ [18]. The multitaper spectrum estimator $\check{P}_{kl}^X$, which uses P orthogonal window functions can be expressed as,

$$\check{P}_{kl}^X = \frac{1}{P} \sum_{p=1}^{P} \left| \sum_{j=0}^{K-1} w_p(j)x(j)e^{-\frac{i2\pi jk}{K}} \right|^2 \tag{6}$$

where $w_p(\cdot)$ is the pth data taper, $p = 1, \cdots, P$. The set of tapers $w_p(\cdot)$, can be selected from a family of orthogonal tapers, such as the sine tapers. The latter tapers were shown in [22] to produce smaller local bias than the Slepian tapers, with roughly the same spectral concentration [18]. The sine tapers family is formulated as,

$$w_p(j) = \sqrt{\frac{2}{J+1}} \sin\left(\frac{\pi p\,(j+1)}{J+1}\right), j = 0,1,\cdots,J-1 \tag{7}$$

where the multiplicative constant makes the tapers orthogonal. Here, the number of tapers P required to represent the signal will depend on the background noise and will be determined empirically.

[1] Only half of coefficients are useful because the audio signals are real and their spectral coefficients are conjugate symmetric.

[2] Through independent experiments with the NMF method, we could verify that the choice of $\nu = 1$ provides the best enhancement results overall.

978-1-5386-5689-1/18/$31.00 ©2018 IEEE

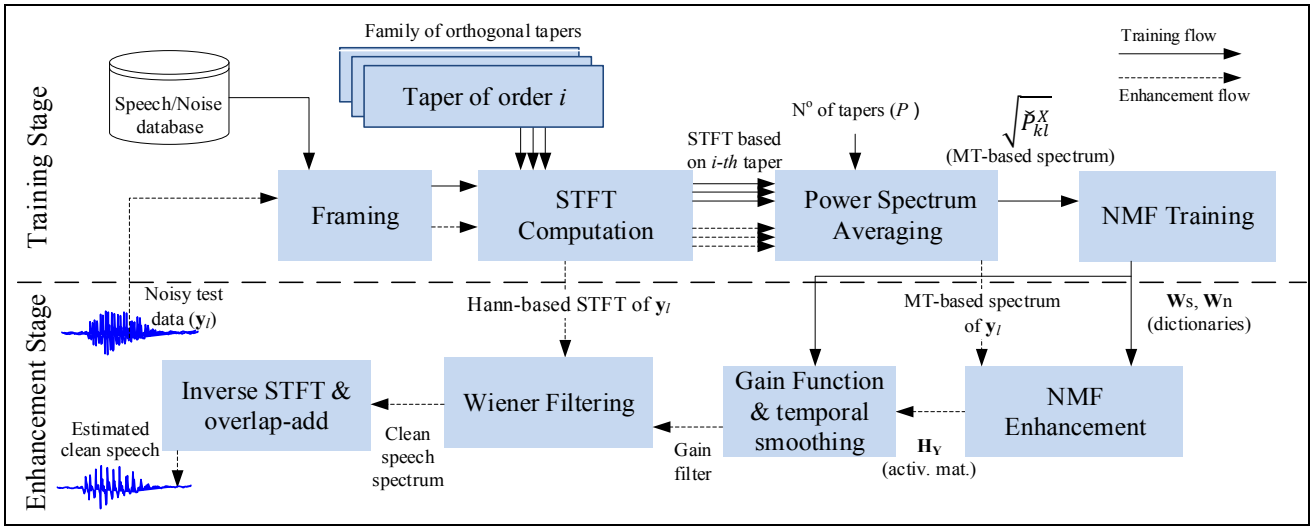

Figure 1. Block diagram of the NMF-based speech enhancement system using on multitaper spectrum estimation.

IV. Integration of Multitapering into NMF

In the previous speech enhancement studies [18], [26]-[27], where the multitaper spectrum estimate is used along with the wavelet thresholding technique, the multitapering method is applicable only in the enhancement stage. In the present context with focus on NMF, we propose to employ the multitapering in both training and enhancement stages in order to estimate the gain function of the Wiener filter. Figure 1 illustrates how multitapering is integrated into the two stages of the NMF framework. For convenience, let $\mathbf{y}_l$ be the vector of noisy test data, as obtained in the time domain for the lth frame, $\mathbf{y}_l = [y(j), \dots, y(j+K-1)]$, and $Y(k,l)$ its corresponding STFT in the spectral domain as used in Figure 1. In the training stage, the magnitude spectrum based on multitapering, i.e., where each entry of $\mathbf{V}$ represents the square root of the power spectrum computed using (6), is used in the estimation of more accurate noise dictionary $\mathbf{W}_N$ and eventually to learn the clean speech dictionary $\mathbf{W}_S$. Note that $\mathbf{W}_S$ can be estimated using the windowed periodogram rather than multitapering if the large-variance spectra issue does not significantly affect the estimation of the clean speech spectra. In the enhancement stage, the multitaper-based magnitude spectrum estimate is used a single time to compute the activation matrix $\mathbf{H}_Y$ of the test data $\mathbf{y}_l$ based on the clean speech and noise dictionaries learned in training stage. Once the gain function is estimated, the clean speech spectrum is obtained by multiplying the gain function by the STFT, $Y(k,l)$ of $\mathbf{y}_l$, computed, this time, using the conventional windowing periodogram, namely, the Hann window. The time-domain enhanced speech signal is then obtained via inverse STFT followed by the overlap-add method. Note that the multitaper power spectrum estimate could be used jointly with Itakura-Saito (IS) [8] divergence as goodness-of-fit criterion in the learning of NMF matrices instead of the multitaper magnitude spectrum together with KL divergence employed here. However, the preliminary results have shown that the latter configuration outperforms the former.

V. Experiments

A. Expimental set-up

The performance of the proposed system is evaluated using the clean speech TSP corpus [28] and noise from the NOISEX dataset [29]. For the clean speech, all adult speakers (11 males and 12 females) from the TSP corpus are selected. For the noise, we selected a subset of the NOISEX corpus, denoted as S_n and consisting of buccaneer 1, HF Channel, babble, factory 1, and pink noises. Each of clean speech and noise signal was divided into three partitions: i) training data, used in the NMF learning stage, ii) validation data, used for tuning the number of data tapers parameter, and iii) test data, used for final performance evaluation. Specifically, the training data consisted of approximately 2 minutes (50 sentences) of speech segments for each speaker from the TSP corpus, as well as 3 minutes segment for the noises. The validation data consisted of 11.5 seconds (5 sentences) of speech for each speaker from the TSP corpus, and 30 seconds of noise from the NOISEX database. The same durations were used for the test partition. The noisy speech was generated by adding the noise to the clean speech to obtain input SNR of 0, 5, and 10 dB. We used $M = 80$ basis vectors for the clean speech and all noise types. Temporal smoothing factors were selected as $(\tau_S, \tau_N) = (0.4, 0.9)$. The sampling rate of TSP and NOISEX signals is adjusted to 16 kHz. For the STFT analysis, we use a window of $K = 512$ samples with 75 % overlap. Regarding the implementation of the proposed system, we considered two noise dictionary estimation approaches: noise-dependent (ND) and noise-independent (NI). In the ND application, a specific noise dictionary, $\mathbf{W}_N^i$, is estimated for each noise type using the underlying noise training data, where $i \in S_n$. In the NI case, we estimate a single universal noise dictionary covering all types of noise. Furthermore, we considered the speaker-independent application where one universal basis matrix covering all speakers is estimated for the proposed system, rather than using a specific clean speech

dictionary for each speaker. We used PESQ (Perceptual Evaluation of Speech Quality), SDR (Signal-to-Distortion Ratio), and SSNR as the objective measures, where a higher value indicates a better speech quality.

B. *Effect of multitapering on NMF components*

In this section, we investigate the impact of using the multitapering approach in the estimation of each of the three main NMF components, namely, the clean speech and noise dictionaries and the activation matrix, on the speech enhancement performance. For this aim, we designed the following NMF systems: i) in *single taper* system (the reference system), only a single data taper is used in the power spectrum estimation of the three components, ii) in the (*NS*) system, the multitapering is used only in the enhancement stage to estimate the activation matrix of the noisy speech, iii) in (*CS+NS*) system, the multitapering is used to estimate both the clean speech dictionary and the activation matrix, iv) in (*N+NS*) system, the multitapering is used for the noise dictionary and activation matrix, and finally v) in the (*CS+N+NS*) system, the multitapering is used in the estimation of the three NMF components. All these systems are evaluated using validation data for both NI and ND applications in terms of PESQ, SDR and SSNR measures. Only PESQ results are reported in this section given that the same trends were observed for the remaining measures. The results reported in Figure 2 show, on the one hand, that the multitapering approach helps to improve SE performance for both ND and NI applications, and on the other hand, not all multitaper-based NMF components are equally important in the achievement of this improvement. Specifically, the noise dictionary and the activation matrix of the noisy speech are the most important components that benefit from the multitapering method. Furthermore, we note that the use of multitapering in the estimation of the clean speech dictionary, not only, does not improve speech enhancement performance as observed for NI application, but can also cause slight performance degradation for ND application. This observation confirms that the problem of large-variance estimates of the spectra affects more the noise and noisy speech than the clean speech.

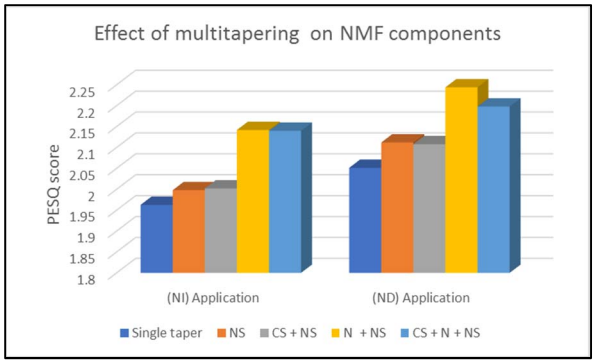

Figure 2. Effect of multitapering NMF components on the improvement of speech quality.

C. *Overlapping of the clean speech and noise dictionaries*

In the context of the NMF method, the observed performance degradation can also be explained by the increasing overlap that might occur between the clean speech and noise dictionaries when the multitapering is used. As the number of tapers used in the power spectrum estimation of the clean speech increases, the fast-varying signal, probably more similar to the noise characteristics, will also be captured and coded in the clean speech dictionary. As a result, the clean speech dictionary will be closer to the noise dictionary, and the estimate of power spectrum of clean speech in equation (4) will contain residual signals from the noise which will degrade the separation performance. To verify this hypothesis, we estimate the amount of similar (or dissimilar) characteristics shared by the clean speech and noise dictionaries using the symmetrized KL divergence $\mathcal{J}(\mathbf{W}_S, \mathbf{W}_N)$, defined as,

$$\mathcal{J}(\mathbf{W}_S, \mathbf{W}_N) = \frac{1}{2}\mathcal{D}_{KL}(\mathbf{W}_S, \mathbf{W}_N) + \frac{1}{2}\mathcal{D}_{KL}(\mathbf{W}_N, \mathbf{W}_S) \quad (8)$$

where the KL divergence between matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ is

$$\mathcal{D}_{KL}(\mathbf{W}_1, \mathbf{W}_2) \triangleq \sum_{k=1}^{K}\sum_{l=1}^{L}\left([\mathbf{W}_1]_{kl} \ln \frac{[\mathbf{W}_1]_{kl}}{[\mathbf{W}_2]_{kl}}\right) \quad (9)$$

The curves of Figure 3 give the sym. KL divergence scores between clean speech and noise dictionaries with respect to the number of data tapers used to estimate either only clean speech dictionary (CS curve), only noise dictionary (N curve), or both (Both curve) dictionaries in NI application. Interesting observations can be made from the three curves. First, when only the clean speech dictionary is estimated using multitapering, we observe that clean speech and noise dictionaries become more similar (the divergence decreases sharply) as the number of tapers used increases, confirming our hypothesis about the reason of speech enhancement performance degradation. Second, we note from the (N) curve, that the overlap between the clean speech and noise dictionaries is substantially less important when the higher orders of data tapers are used to estimate only the noise dictionary compared to the previous case (CS curve). This observation suggests that the noise signal is characterized by faster variations than clean speech signal. Finally, we observe that the use of multitapering in the estimation of both dictionaries does not help to reduce the overlap between them. Thus, nearly the same KL divergence scores are obtained for (N) and (Both) systems, which explain the similar speech enhancement performance achieved for the underlying cases reported in Figure 3.

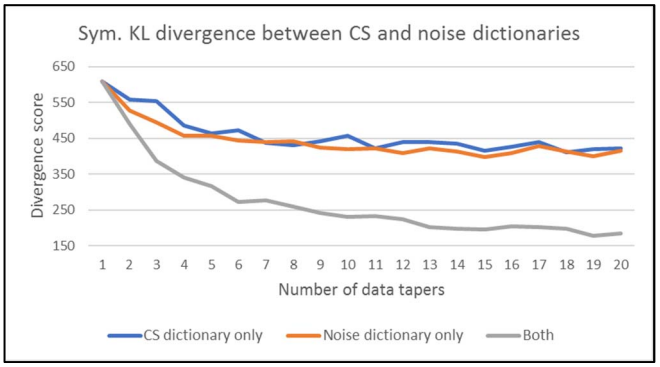

Figure 3. Symmetrized KL divergence between clean speech and noise dictionaries computed for noise-independent application.

D. *Effect of the number of the data tapers*

Beside the audio type, e.g., clean speech vs. noise, the number of data tapers required to better estimate the power spectrum of the audio signal may also depend on the noise type. Thus, we carried out several experiments that optimize PESQ results according to the number of the tapers used for each specific type of noise, based on the validation dataset as illustrated in Figure 4. The number of data tapers used in multitapering for noise dictionary and activation matrix ranges from one to 20. We note from Figure 4 that multitapering method used for the noise dictionary improves PESQ scores for all types of noises, in both NI and ND applications.

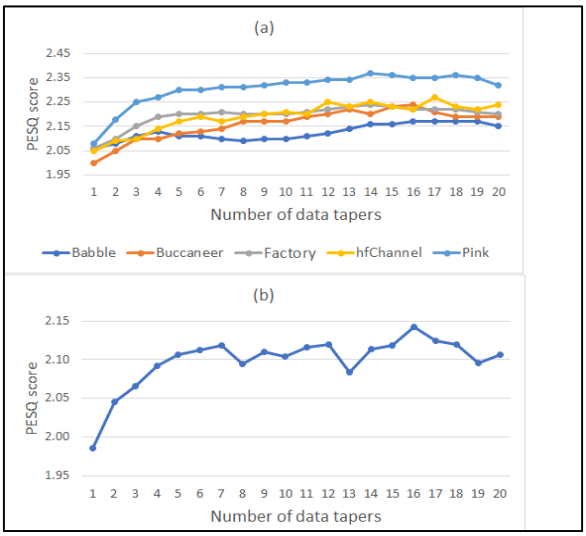

Figure 4. Impact of the number of data tapers at 5dB input SNR, on the PESQ scores in (a) ND and (b) NI applications.

For NI, a combination of 16 data tapers gives the best result, and for ND, the number of tapers ranges from 14 to 17 depending on the type of noise (14 for pink noise and factory 1, 16 for buccaneer 1 and babble, and 17 for HF Channel). The same conclusion is also made from the SDR and SSNR scores which are not reported in this section. The number of tapers optimizing SE performances obtained with validation data will be used for the final performance evaluation.

E. *Results*

In this section, we report the performance evaluation of the NMF system based on multitaper spectrum estimation (NMF-MT system) compared to the NMF reference system based on Hann window spectrum estimate (NMF-Hann), using TSP test data. The PESQ, SDR and SSNR results for each noise type at 0, 5, and 10 dB input SNR for ND and NI are shown in Table 1 and 2 respectively. In ND application, we observe that considerable improvements in PSQ, SDR and SSNR are achieved by multitaper-based NMF system compared to single taper-based one for all noise types and under all input noise conditions. On average, absolute improvements of 0.16, 1.49 dB and 1.97 dB are achieved for PESQ, SDR and SSNR respectively.

For NI application, better results are also obtained by the proposed system for all noise types except for babble noise where comparable PESQ results are achieved for both systems. On average, absolute improvements of 0.09, 1.36 dB and 1.51 dB are obtained for PESQ, SDR and SSNR respectively. These results on test data confirm that reducing the variance in the spectral estimation used for the noise dictionary and the noisy speech activation matrix, via a multitapering approach, has a positive impact on speech enhancement performance.

Table 1 Results achieved on TSP test data in ND application.

Noise Type	Input SNR	NMF-Hann			NMF-MT		
		PESQ	SDR	SSNR	PESQ	SDR	SSNR
Buccaneer 1	0 dB	1.72	4.36	-3.71	**1.87**	**7.15**	**-0.26**
	5 dB	2.03	8.99	0.49	**2.28**	**10.99**	**3.08**
	10 dB	2.44	13.62	5.10	**2.61**	**14.40**	**6.27**
Hfchannel	0 dB	1.68	5.69	-2.61	**1.82**	**7.80**	**0.33**
	5 dB	2.05	10.23	1.66	**2.25**	**11.90**	**4.22**
	10 dB	2.38	14.11	5.50	**2.60**	**15.53**	**7.89**
Babble	0 dB	1.78	2.90	-4.07	**1.85**	**5.05**	**-1.68**
	5 dB	2.17	8.14	0.49	**2.23**	**9.35**	**1.77**
	10 dB	2.50	12.13	4.02	**2.57**	**12.58**	**4.79**
Factory 1	0 dB	1.63	3.82	-3.69	**1.77**	**5.46**	**-0.85**
	5 dB	2.07	9.09	1.16	**2.17**	**9.93**	**2.42**
	10 dB	2.45	13.28	5.25	**2.52**	**13.57**	**5.38**
Pink	0 dB	1.69	5.14	-3.05	**2.00**	**7.87**	**0.30**
	5 dB	2.11	9.80	1.41	**2.39**	**11.67**	**3.45**
	10 dB	2.54	14.66	6.06	**2.70**	**15.08**	**6.45**

Table 2. Results achieved on TSP test data in NI application.

Noise Type	Input SNR	NMF-Hann			NMF-MT		
		PESQ	SDR	SSNR	PESQ	SDR	SSNR
Buccaneer 1	0 dB	1.65	3.97	-3.65	**1.80**	**6.44**	**-0.90**
	5 dB	2.01	8.72	0.58	**2.19**	**10.62**	**2.64**
	10 dB	2.38	12.76	4.34	**2.53**	**14.05**	**5.81**
Hfchannel	0 dB	1.59	4.61	-3.22	**1.63**	**5.67**	**-1.80**
	5 dB	1.93	9.33	1.06	**1.99**	**10.17**	**1.93**
	10 dB	2.29	13.25	4.83	**2.36**	**13.81**	**5.21**
Babble	0 dB	1.79	1.85	-5.49	**1.82**	**3.26**	**-4.12**
	5 dB	**2.12**	6.87	-0.95	**2.12**	**8.27**	**0.33**
	10 dB	**2.44**	11.41	3.20	2.42	**12.77**	**4.48**
Factory 1	0 dB	1.70	3.54	-3.81	**1.79**	**4.98**	**-1.72**
	5 dB	2.07	8.38	0.47	**2.18**	**9.69**	**2.13**
	10 dB	2.42	12.49	4.23	**2.51**	**13.65**	**5.60**
Pink	0 dB	1.74	4.83	-2.84	**1.87**	**6.63**	**-0.77**
	5 dB	2.14	9.51	1.33	**2.27**	**10.89**	**2.81**
	10 dB	2.50	13.30	4.93	**2.60**	**14.36**	**6.03**

VI. CONCLUSION

In this study, we have investigated the use of the multitaper spectrum estimation approach for NMF-based speech enhancement system. When compared to the NMF system based on conventional single tapering, the results obtained with TSP corpus show notable improvements in PESQ, SDR and SSNR scores in all input SNR conditions for both NI and ND applications and for almost all types of noise, except for Babble in NI application where single taper and multitaper approaches have comparable PESQ results. We have also found that the power spectrum estimates of noise and noisy speech greatly benefit from the multitapering approach in contrast to the clean speech where the conventional window periodogram is sufficiently accurate.

978-1-5386-5689-1/18/$31.00 ©2018 IEEE

REFERENCES

[1] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," IEEE Trans. Speech Audio Process., vol. 7, no. 2, pp. 126–137, Mar. 1999.

[2] E. Plourde and B. Champagne, "Auditory-based spectral amplitude estimators for speech enhancement," IEEE Trans. Audio Speech Lang. Process., vol. 16, no. 8, pp. 1614–1623, Nov. 2008.

[3] 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.

[4] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7-19, 2015.

[5] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness constraint," IEEE Trans. Audio Speech Lang. Process., vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[6] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," IEEE Trans. Audio Speech Lang. Process., vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[7] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in Proc. Neural Inf. Process. Syst. (NIPS), pp. 556–562, May 2001.

[8] C. Fevotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," Neural Comput., vol. 21, no. 3, pp. 793–830, Mar. 2009.

[9] K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based speech enhancement using bases update," IEEE Signal Process. Lett., vol. 22, no. 4, pp. 450–454, Apr. 2015.

[10] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based speech enhancement incorporating deep neural network," in INTERSPEECH, pp. 2843-2846, Sep. 2014.

[11] M. Kim and P. Smaragdis, "Mixtures of Local Dictionaries for Unsupervised Speech Enhancement," IEEE Signal Processing Letters, vol. 22, no. 3, pp. 293-297, 2015.

[12] J. L. Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in IEEE ICASSP, pp. 66-70, Apr. 2015.

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

[14] H. T. Fan, J. w. Hung, X. Lu, S. S. Wang, and Y. Tsao, "Speech enhancement using segmental nonnegative matrix factorization," in IEEE ICASSP, pp. 4483-4487, May 2014.

[15] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in IEEE ICASSP, pp. 3734-3738, May 2014.

[16] H. W. Tseng, M. Hong, and Z. Q. Luo, "Combining sparse NMF with deep neural network: A new classification-based approach for speech enhancement," in IEEE ICASSP, pp. 2145-2149, Apr. 2015.

[17] A. T. Walden, D. B. Percival, and E. J. McCoy, "Spectrum estimation by wavelet thresholding of multitaper estimators," IEEE Transactions on Signal Processing, vol. 46, no. 12, pp. 3153-3165, 1998.

[18] Y. Hu and P. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," IEEE Trans. on Speech and Audio Processing, vol. 12, no. 1, pp. 59–67, Jan. 2004.

[19] W. Charoenruengkit and N. Erdol, "The effect of spectral estimation on speech enhancement performance," in IEEE Trans. on Audio, Speech, and Language Processing, vol. 19, no. 5, pp. 1170-1179, July 2011.

[20] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," IEEE Trans. Audio, Speech, and Language Processing, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[21] D. J. Thomson, "Spectrum estimation and harmonic analysis," IEEE proceedings, vol. 70, no. 9, pp. 1055–1096, 1982.

[22] K. S. Riedel and A. Sidorenko, "Minimum bias multiple taper spectral estimation," IEEE Trans. on Signal Processing, 43(1), 188–195, Jan. 1995.

[23] T. Kinnunen, R. Saeidi, F. Sedlak, K.A. Lee, J. Sandberg, M. Hansson-Sandsten, H. Li, "Low-variance multitaper MFCC features: a case study in robust speaker verification," IEEE Transactions on Audio, Speech and Language Processing, vol. 20(7), pp. 1990-2001, 2012.

[24] J. Alam, P. Kenny, and D. O'Shaughnessy, "Low-variance multitaper mel-frequency cepstral coefficient features for speech and speaker recognition systems," Springer Cognitive Computation Journal, vol. 5, no. 4, pp. 533-544, 2013.

[25] Y. Attabi, Md J. Alam, P. Dumouchel, P. Kenny, and D. O'Shaughnessy, "Multiple windowed spectral features for emotion recognition." In IEEE ICASSP, pp. 7527-7531, May 2013.

[26] Y. Ma and A. Nishihara, "A modified Wiener filtering method combined with wavelet thresholding multitaper spectrum for speech enhancement," EURASIP Journal on Audio, Speech, and Music Processing, vol. 22, no. 1, p. 32, 2014.

[27] E. Jayakumar and P. Sathidevi, "Speech enhancement based on noise type and wavelet thresholding the multitaper spectrum," in Advances in Machine Learning and Signal Processing: Springer, pp. 187-200, 2016.

[28] P. Kabal, "TSP Speech Database," McGill Univ., Montreal, QC, Canada, Tech. Rep. 09-02, 2002.

[29] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition. II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Commun., vol. 12, no. 3, pp. 247–251, Jul. 1993.

[30] S. M. Kay, Modern Spectral Estimation. Signal Processing Series, Englewood Cliffs, NJ: Prentice-Hall, 1988.