ARCHIVES OF ACOUSTICS
Vol. **45**, No. 4, pp. 565–572 (2020)
DOI: 10.24425/aoa.2020.134072

Research Paper

Speech Enhancement Based on Discrete Wavelet Packet Transform and Itakura-Saito Nonnegative Matrix Factorisation

Houguang LIU, Wenbo WANG*, Lin XUE, Jianhua YANG
Zhihua WANG, Chunli HUA

School of Mechatronic Engineering
China University of Mining and Technology
Xuzhou 221116, China
*Corresponding Author e-mail: wangwenbo@cumt.edu.cn

(*received August 22, 2019; accepted August 26, 2020*)

Nonnegative matrix factorization (NMF) is one of the most popular machine learning tools for speech enhancement (SE). However, there are two problems reducing the performance of the traditional NMF-based SE algorithms. One is related to the overlap-and-add operation used in the short time Fourier transform (STFT) based signal reconstruction, and the other is the Euclidean distance used commonly as an objective function; these methods can cause distortion in the SE process. In order to get over these shortcomings, we propose a novel SE joint framework which combines the discrete wavelet packet transform (DWPT) and the Itakura-Saito nonnegative matrix factorisation (ISNMF). In this approach, the speech signal was first split into a series of subband signals using the DWPT. Then, the ISNMF was used to enhance the speech for each subband signal. Finally, the inverse DWPT (IDWT) was utilised to reconstruct these enhanced speech subband signals. The experimental results show that the proposed joint framework effectively enhances the performance of speech enhancement and performs better in the unseen noise case compared to the traditional NMF methods.

Keywords: speech enhancement; discrete wavelet packet transform; nonnegative matrix factorisation; Itakura-Saito divergence.

1. Introduction

Speech quality and intelligibility are degraded due to the presence of the environmental and background noises. Therefore, speech enhancement (SE) is a necessary for obtaining the original speech signal from contaminated one, in order to improve the speech quality and intelligibility. It is a key component in many speech applications including hearing aids, mobile communications, and speech recognition (Lai *et al.*, 2016; Li *et al.*, 2011; Wang, Chen, 2018; Wang, Hansen, 2018).

Major challenges for the SE stem from the underdetermined mixing systems, reverberant environments, and the presence of noise and non-stationarity of speech. Monaural speech separation is more challenging than the SE using multiple microphones, and it can hardly improve speech intelligibility (Krawczyk-Becker, Gerkmann, 2016; Luts *et al.*, 2010). In order to solve these problems, SE has been implemented using many methods, which can be divided into unsupervised and supervised ones. The researchers working on the unsupervised methods have proposed several different algorithms, such as Wiener filtering algorithms (Scalart, Filho, 1996), principal component analysis (Bavkar, Sahare, 2013; Saleem *et al.*, 2018), spectral subtraction (Boll, 1979), and Kalman filtering (Grancharov *et al.*, 2006). The supervised methods have developed rapidly in last several years. They involve nonnegative matrix factorisation (NMF) (Lee *et al.*, 2017; Varshney *et al.*, 2017; Wang *et al.*, 2018a; 2018b), Hidden Markov Model (Veisi *et al.*, 2015), deep neural networks (Nie *et al.*, 2018; Saleem *et al.*, 2019), and deep denoising auto-encoder (Wang *et al.*, 2015). Since the unsupervised methods assumed that the noise is stationary or slowly changing, it is difficult to apply to the real environment. But for the supervised methods, it contains training and enhancement phases for SE. In the training phase, an enhanced model which characterises the speech and noise correlation is trained. Then, this trained-enhanced model is used for SE during the enhancement phase. These supervised methods do not make any assumptions like unsupervised methods, so they achieve better results in SE.

As a supervised method, NMF has been successfully realised into SE. In the past two decades, a lot of work has been reported on NMF in the field of SE. Lee and Seung (1999) discovered that this method can learn the holistic features of the signal, which performs well over principal components analysis and vector quantisation. Mohammadiha *et al.* (2013) proposed an online Bayesian formulation of NMF to enhance the noisy signal. Their results showed that the system outperforms the competing algorithms substantially. Févotte *et al.* (2013) used the NMF to train sparse nonnegative dynamical model on speech data and the results showed that the model can capture the dynamics of speech in a useful way. Chien and Yang (2015) proposed a variational Bayesian NMF that learns the variational parameters and model parameters; the proposed method outperforms the conventional NMF.

Although the above NMF methods have been confirmed as being effective for SE, there are still some problems reducing its performance. Firstly, NMF methods are mostly estimated by using the short time Fourier transform (STFT) to analyse the spectrogram. However, there may be some distortion in the STFT process because of the segmentation, the windowing processes, and the noisy phase (Islam *et al.*, 2019; Mowlaee, Saeidi, 2014). Secondly, the NMF methods commonly optimise their cost function defined by Euclidean distance (Li *et al.*, 2017). But the Euclidean distance can cause relatively large reconstruction errors since it tends to overemphasise the reconstruction accuracy of large values; therefore the Euclidean distance is not suitable for processing speech signals.

Accordingly, in this paper, we propose a novel SE method which combines the discrete wavelet packet transform (DWPT) and the Itakura-Saito nonnegative matrix factorisation (ISNMF) to enhance the speech enhancement performance in different noise environments.

2. The proposed algorithm

2.1. Nonnegative matrix factorisation

As mentioned before, NMF as a sound source separation technology has been widely used in monaural speech mixtures. It is a technique for projecting any nonnegative matrices into space, including a nonnegative basis matrix and a nonnegative weight matrix. For example, given a data set $\mathbf{V} \in \mathrm{R}^{M\times N}$, NMF is used to calculate a nonnegative basis matrix $\mathbf{W} \in \mathrm{R}^{M\times r}$ and a weight matrix $\mathbf{H} \in \mathrm{R}^{r\times N}$, the size of r should be less or equal to min (M, N), such that:

$$\mathbf{V} \approx \mathbf{WH}, \tag{1}$$

where $\mathbf{V}$ usually represents the magnitude or power spectrogram of speech, and $\mathbf{W}$ is called a dictionary of spectral templates, $\mathbf{H}$ is a matrix of temporal activations in the process of speech signal processing.

In order to approximate the nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$, a cost function is used to penalise the error between $\mathbf{V}$ and $\mathbf{WH}$, such that:

$$\min_{\mathbf{W},\,\mathbf{H}\geq 0} C(\mathbf{V}|\mathbf{WH}), \tag{2}$$

where $C(\mathbf{V}|\mathbf{WH})$ represents the cost function defined by

$$C(\mathbf{V}|\mathbf{WH}) = \sum_{m=1}^{M}\sum_{n=1}^{N} d\left([\mathrm{V}]_{mn}|\,[\mathrm{WH}]_{mn}\right), \tag{3}$$

where $d(x|y)$ is a scalar cost function. There are many cost functions such as Euclidean distance or Kullback-Leibler (KL) and Itakura-Saito (IS) divergences. In this paper, we choose the IS divergence because it has been shown relevant for audio applications that we here define as

$$d_{IS}(x,y) = \frac{x}{y} - \log\frac{x}{y} - 1. \tag{4}$$

The matrices, $\mathbf{W}$ and $\mathbf{H}$, are expressed by applying multiplicative iterative updating rules as described in Lee and Seung (1999) and the update rules are given as (Magron, Virtanen, 2018; Nakano *et al.*, 2010):

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \left(\frac{\left([\mathbf{WH}]^{-2}\otimes\mathbf{V}\right)\mathbf{H}^{\mathrm{T}}}{[\mathbf{WH}]^{-1}\,\mathbf{H}^{\mathrm{T}}}\right)^{0.5}, \tag{5}$$

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \left(\frac{\mathbf{W}^{T}\left([\mathbf{WH}]^{-2}\otimes\mathbf{V}\right)}{\mathbf{W}^{\mathrm{T}}\,[\mathbf{WH}]^{-1}}\right)^{0.5}, \tag{6}$$

where $\mathbf{A}\otimes\mathbf{B}$ and $\frac{\mathbf{A}}{\mathbf{B}}$ represent respectively the element-wise multiplication and division of matrices $\mathbf{A}$ and $\mathbf{B}$. The $[\mathbf{A}]^{\beta}$ represents the element-wise exponentiation of matrices $\mathbf{A}$. The superscript T is the matrix transpose, and for the initialisations of $\mathbf{W}$ and $\mathbf{H}$, positive random numbers are often used.

2.2. Discrete wavelet packet transform

The discrete wavelet packet transform method is a generalisation of wavelet decomposition which was first proposed by Coifman and Wickerhauser (1992). Compared to wavelet analysis with a fixed decomposition structure, DWPT decomposes the signal into different subband signals. It decomposes not only the low frequency subband, but also the high frequency one. Therefore, the transformation can make the low-frequency components of the signal easy to be distinguished. Moreover, it provides more details for the signal at high frequencies. In addition, because the experimental observation found that the human ear is like a filter bank, the decomposition scheme for obtaining DWPT coefficients is very similar to the frequency analysis characteristics of the human ear and auditory perception (Bouzid, Ellouze, 2016; Gokhale, Khanduja, 2010; Mavaddaty *et al.*, 2017; Sun, Qin, 2016).

The tree structure of the three-level wavelet packet transform is shown in Fig. 1. In wavelet analysis, the signal is decomposed into approximate and detailed coefficients by a recursion of filter-decimation operations. The approximate and detailed coefficients will continue to be decomposed by the filter-decimation operations, and this process is repeated until the three-level wavelet packet transform is decomposed. As shown in Fig. 1, each node is represented by (E,n), where E indicates the level of decomposition and n is the subband index. The root of the tree $(E,n) = (0,0)$ is the time representation of the signal. The left and right branches represent the low pass and high pass filters, respectively. The bottom level of the tree is the frequency representation of the signal, and the corresponding node is (3, 0), (3, 1), ..., (3, 7). The detailed information of the DWPT (IDWPT) used for speech was provided in the following sections.

2.3. *DWPT-ISNMF based speech enhancement system*

This paper proposes a novel SE system that combines the discrete wavelet packet transform (DWPT) and the Itakura-Saito nonnegative matrix factorisation (ISNMF) to improve the speech enhancement performance. The overall methodology of the presented SE framework is shown in Fig. 2. The noise speech passes through the DWPT to produce a set of subband signals $\{\mathbf{s}_b^J\}$, each of the subband signal is individually enhanced by ISNMF. Then, the inverse DWPT (IDWPT) is used to reconstruct these enhanced speech subband signals. Finally, the enhanced speech is obtained.

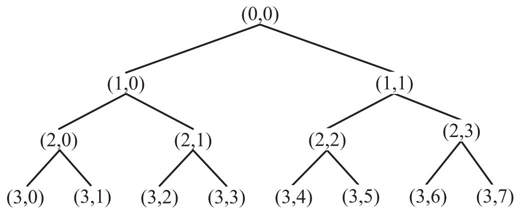

Fig. 1. Tree structure of the three-level discrete wavelet packet transform.

The DWPT can be implemented by two-channel filter banks which are filtering signals with a low-pass $h(k)$ and a high-pass $g(k)$ filters. The analysis of a signal is carried out first by decomposing the signal into two subband signals, carrying information of low- and high-frequency components. At the next level j, the scheme is iterated successively on both the subbands. The decomposition of the signal into different frequency bands with different resolutions is therefore obtained by successive high and low pass filtering of the signal.

The decomposition of DWPT and reconstruction of DWPT (IDWPT) can be considered as a tree-structured filter bank, as shown in Fig. 3a, the symbols $\downarrow 2$ and $\uparrow 2$ in the rectangles indicate the operation of down- and upsampling by 2, respectively. Downsampling by 2 means discarding all the odd or even samples of wavelet coefficients, whereas upsampling by 2 means adding zeros between the samples of wavelet coefficients. Figures 3b and 3c illustrate this operational process. For the left side of Fig. 3a, the time signal $\mathbf{f}$ is first decomposed into two subband signals

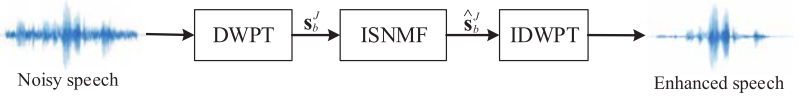

Fig. 2. Overall methodology of the presented SE framework.

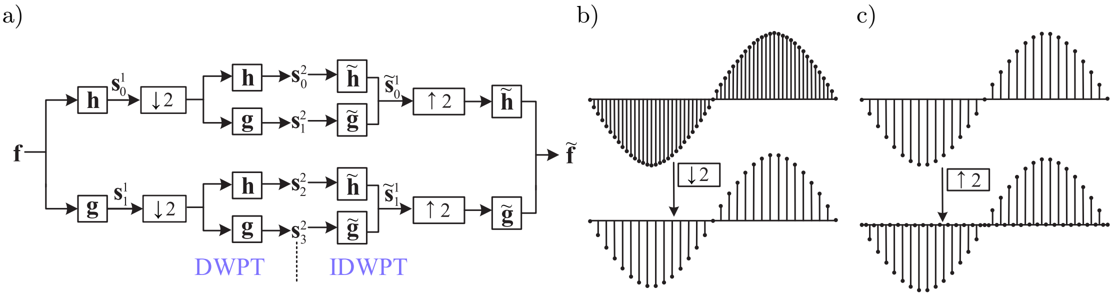

Fig. 3. (a) Flowchart of the two-level decomposition DWPT and IDWT. Here **h** and **g** denote the frequency responses of the low and high pass decomposition filters, respectively. (b) The operation of downsampling. (c) The operation of upsampling. The down arrows represent decimation by 2.

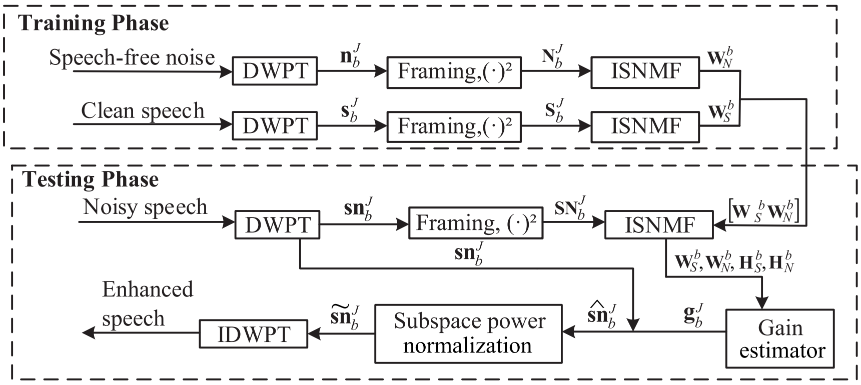

Fig. 4. Block diagram of the proposed speech enhancement system.

and then decomposes recursively on both to produce a set of subband signals $\mathbf{s}_b^J$, where $b = 0, 1, 2, ..., 2^J - 1$, b indicates the subband index, and J denotes the level of DWPT. The IDWPT is to reconstruct the decomposed results to form the time signal $\mathbf{f}$. The reconstructed signal is the sum of the set of subband signals $\mathbf{s}_b^J$.

The detailed SE framework is shown in Fig. 4, and it consists of two phases: the training and enhancement ones. In both of them, the matrix of subband b from the DWPT is further squared to obtain a nonnegative matrix $\mathbf{S}_b^J$.

In the training stage, the nonnegative matrix $\mathbf{S}_b^J$ from the noise-free speech is used to create a speech basis matrix $\mathbf{W}_S^b$ by NMF. Likewise, a speech-free noise is used to create a noise basis matrix $\mathbf{W}_N^b$ by NMF. Finally, we get a double-wide matrix which is horizontally concatenated through these two basis matrices, $\mathbf{W}_V^b = \left[\mathbf{W}_S^b \ \mathbf{W}_N^b \right]$. For the enhancement stage, the $\mathbf{SN}_b^J$ of the subband signal $\mathbf{sn}_b^J$ for the input noise-corrupted speech is analysed via NMF, by keeping the basis matrix $\mathbf{W}_V^b$. So only the weight matrix needs to be iteratively updated during the analysis. Finally we have

$$\mathbf{SN}_b^J = \mathbf{X}_b^s + \mathbf{X}_b^n \approx \left[\mathbf{W}_b^S \ \mathbf{W}_b^N \right] \begin{bmatrix} \mathbf{H}_b^S \\ \mathbf{H}_b^N \end{bmatrix}, \tag{7}$$

where $\mathbf{H}_b^S$, $\mathbf{H}_b^N$ denote the components in the weight matrix $\mathbf{SN}_b^J$ for the speech and noise, respectively.

The gain estimation of the subband signal $\mathbf{sn}_b^J$ for the noise-corrupted speech is calculated, and the enhanced subband signal is expressed as (Wang *et al.*, 2016):

$$\mathbf{g}_b^J = .\sqrt{(\mathbf{W}_S^b \mathbf{H}_S^b./(\mathbf{W}_S^b \mathbf{H}_S^b + \mathbf{W}_N^b \mathbf{H}_N^b)}, \tag{8}$$

$$\widehat{\mathbf{sn}}_b^J = \mathbf{sn}_b^J . \times \mathbf{g}_b^J, \tag{9}$$

where the symbols "$.\sqrt{}$", "./" and ".×" denote the element-wise square root, division, and multiplication operations, respectively. The $\mathbf{g}_b^J$ is gain estimation for subband $\mathbf{sn}_b^J$, the $\widehat{\mathbf{sn}}_b^J$ is the enhanced subband signal.

In order to compensate the noise effect and obtain the power normalised subband signal, a power normalisation scheme is applied to the estimated subband signal as:

$$\widetilde{\mathbf{sn}}_b^J = \frac{\sigma_{b,c}}{\sigma_b} \widehat{\mathbf{sn}}_b^J, \tag{10}$$

where $\sigma_{b,c}$ represents the clean speech root mean square value which is calculated at the training phase, and σ_b denotes the root mean square value of the enhanced subband signal. Finally, we can obtain the enhanced speech signal by using the IDWPT on the $\widetilde{\mathbf{sn}}_b^J$.

3. Experimental results

In this section, we first introduced the design of the dataset and experimental parameters. Next, introducing the utilised evaluation criteria, i.e.: Short Time Objective Intelligibility, Perceptual Evaluation of Speech Quality and Segmental SNR. Then, the overall performance of the proposed method was compared with conventional methods using the above evaluation criteria. Finally, we analysed the overall performance of these methods in the seen noise and unseen noise cases.

3.1. Dataset

For the experiment, the UW/NU corpus was utilised to train and test the proposed SE system (Panfili *et al.*, 2017). The corpus contains 72 lists, each consisting of ten phonetically balanced sentences. The corpus audio files are in WAV format, sampled at 44.1 kHz with 16-bit quantisation, high-pass filtered from 60 to 22,000 Hz, and smoothed at 100 Hz, the 100 Hz is a parameter setting of the high-pass filtered, the length of these utterances is around 2 to 3 s. We used 50 randomly chosen utterances from this database as our training utterances. And the test sets consisted of other 30 utterances, every utterance is downsampled at 8 kHz. Five representative noises were selected from the NOISEX dataset for training and testing purposes, including factory, babble, white, pink, HF channel (Varga, Steeneken, 1993). Furthermore, three noises, "factory," "babble," "white," were used to create the noise basis matrix of NMF at the training phase. Since the noises are around four minutes long, we randomly cut the first two minutes of each noise when constructing the training sets. Then we added those random cuts to the training utterances at −5, 0, 5, 10 dB. The next two minutes of noise were randomly cut for the test set, and the test sets were obtained by mixing test utterances with those random cuts at −5, 0, 5, 10 dB. Eventually, we got a set of training sets and a set of test sets (30 utterances × 5 noises × 4 SNR).

The number of frames of speech basis matrix was set to 40, and because there are three kinds of noise, the parameter values of noise basis matrix was set to 120. The db10 wavelet, which has a good regularity and whose fitted signal is relatively smooth, was selected, and the level of the wavelet was set to 3, which will be analysed in the following sections. The STFT-NMF were computed using the frame size and shift size with the value of 256 and 80 samples, respectively.

3.2. Evaluation criteria

The performance of our proposed SE method was evaluated by Short Time Objective Intelligibility (STOI) (Taal *et al.*, 2011), Perceptual Evaluation of Speech Quality (PESQ) (Rix *et al.*, 2001), and Segmental SNR (segSNR) (Hansen, Pellom, 1998), respectively. These evaluation criteria are widely used to assess the enhancement speech signals since STOI is highly correlated with the intelligibility score of human speech, PESQ is closely related to voice quality, and segSNR can show the quantity of noise reduction in the enhanced speech. STOI returns score in the range of 0 to 1, and the value of PESQ is between 0 and 4.5, where higher STOI and PESQ values imply better intelligibility and quality of the speech. The segSNR is limited within the [−10, 35] dB, higher values imply less noise in the speech.

3.3. Results

The level of DWPT may have an effect on SE. In order to determine how the level of DWPT affects the result, we used different resolution levels p ($p = 2, 3, 4, 5$) to enhance speech. The results are shown in Fig. 5, we found that when the p is greater than 4 the enhancement effect decreases significantly. This may happen because with the further decomposition of DWPT, the frequency band of the filter will become more and more narrow, thereby reducing the speech information contained in the frequency band. Therefore, the NMF cannot enhance the speech of each band any better, and this reduces the overall enhancement effect. And if the p is smaller 3, the DWPT will not be able to provide more details for the signal, this will also reduce the enhancement effect, especially in the case of a low SNR. Compared with p equal to 3, when p equal to 4, although the quality of speech is improved, the noise in the speech is not reduced and the amount of calculation is doubled. Therefore, we finally chose $p = 3$ to enhance speech.

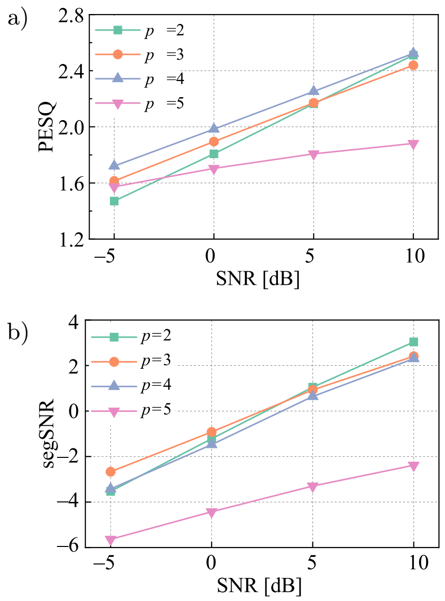

Fig. 5. Comparative performance evaluation of the level of the wavelet.

The experiments showed that the generalised KL cost function defined as

$$d_{KL}(x,y) = x\log\frac{x}{y} - x + y \qquad (11)$$

works best in speech separation tasks (Li *et al.*, 2017; Sun, Fevotte, 2014).

So, we combined a KL divergence and DWPT denoted by "DWPT-KLNMF" as a method to compare with the proposed one. We denoted the traditional NMF method uses STFT as "STFT-NMF". Therefore, there were 5 methods to be compared: "STFT-NMF", "STFT-ISNMF", "DWPT-NMF", "DWPT-KLNMF", and our proposed method.

The overall performance of these methods under various signal-to-noise ratio (SNR) conditions are represented by the histograms shown in Figs 6–8. They show that the proposed method outperforms the other four methods. For the speech quality, 4.2% and 2.9% average improvement is found compared to "STFT-NMF" and "DWPT-NMF", respectively. For the speech intelligibility, the proposed method improves on average by 10.6% and 2.8% compared to "STFT-NMF" and "DWPT-NMF". And for segSNR, the improvements on average is by 1.46 dB and 0.25 dB compared to "STFT-NMF" and "DWPT-NMF".

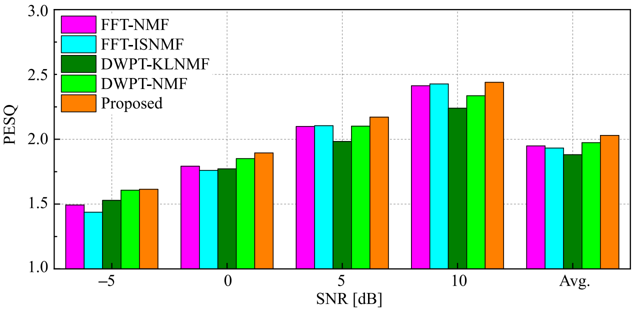

Fig. 6. Comparative performance evaluation of the different speech enhancement methods using perceptual evaluation of speech quality (PESQ) under various SNR conditions.

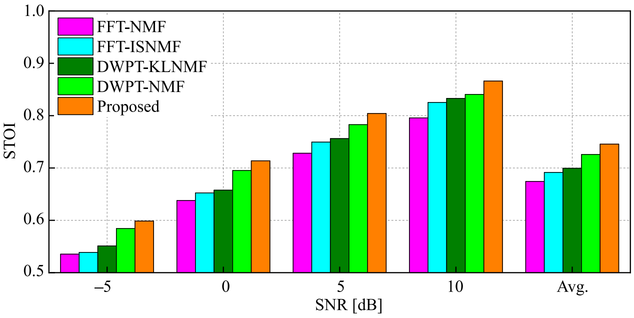

Fig. 7. Comparative performance evaluation of the different speech enhancement methods using short time objective intelligibility (STOI) under various SNR conditions.

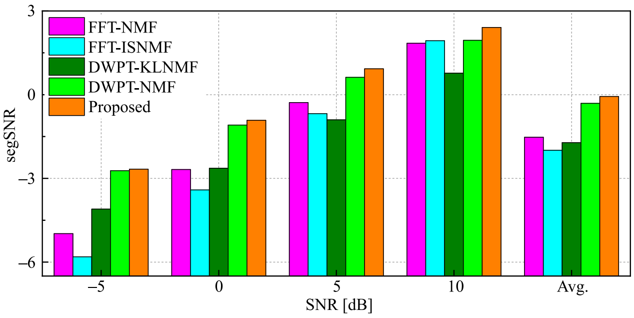

Fig. 8. Comparative performance evaluation of the different speech enhancement methods using segmental SNR (segSNR) under various SNR conditions.

By comparing Figs 6, 7, and 8, we found that the proposed method improves all evaluation criteria under various SNR, particularly the speech intelligibility and segSNR. However, for the speech quality, the proposed method is not particularly good, especially in a high SNR condition. The main reason for this result is that the PESQ rather concerns non-speech segments, which weakens the effect of IS divergences and DWPT on speech signal. But STOI pays attention to the speech segments and less to the non-speech segments, and the segSNR is focused on all frames of speech. Therefore, these two evaluation indicators are better than PESQ.

We can also observe that the effects of noise under different SNRs on SE are different. Compared with the evaluation criteria in high SNR conditions, the proposed method is better than "STFT-NMF" in the case of a low SNR. This could be because as the SNR increases, the distortion of STFT becomes smaller, making DWPT no longer advantageous. The opposite performance trend can be observed from the "DWPT-NMF", the proposed method is better than "DWPT-NMF" in the case of a high SNR. This is because the noisy speech spectrogram under the severe noise condition has a larger dynamic range than the spectrogram with little noise, and the IS cost function is less sensitive to large dynamic ranges than the other distance. In other words, a low SNR will weaken the effect of the IS cost function.

At the same time, we analysed the overall performance of these methods in the seen noise case and in the unseen noise case. The seen noise case refers to the noise used in the training phase, the seen noise includes factory noise, babble noise, and white noise. The unseen noise case refers to the noise not used during the training phase, the unseen noise includes pink noise and HF channel noise. Although our proposed method improves most of evaluation criteria, studying the impact of unseen noise on SE will help us enhance the performance of SE methods.

As shown in Fig. 9, for the speech intelligibility and the segSNR, the proposed method outperforms the other three methods in the seen noise and unseen noise cases. Compared to traditional methods, the proposed method performs better in the unseen noise case. It is proved that the proposed method can improve the speech quality and intelligibility in complex noise environments and outperforms the traditional methods.

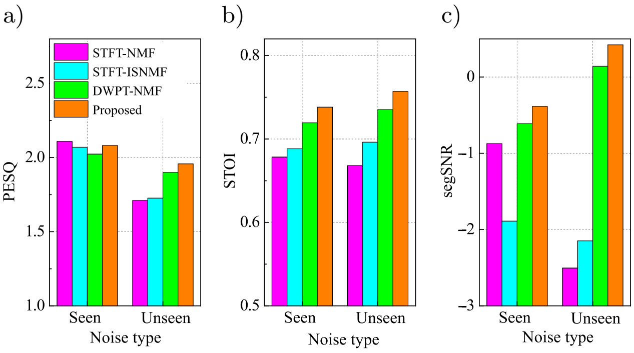

Fig. 9. Comparative performance evaluation of the different speech enhancement methods using (a) PESQ, (b) STOI, (c) segSNR, under seen and unseen noise conditions.

4. Conclusion

In this paper, we propose a novel SE method which combines the DWPT and the ISNMF. First, the noise speech was split into a series of subband signals using the DWPT. Then the ISNMF was applied to enhance the contaminated speech. The performance of the proposed method was compared with four traditional methods. The results show that the proposed method outperformed the conventional methods in speech quality and intelligibility. Besides, it is also demonstrated that the proposed method performs well not only in the seen noise condition, but also in the case of unseen noise. In the future, we intend to study the phase of the speech, which also plays an important role in SE, as the NMF does not consider it.

Acknowledgments

This work is supported by National Natural Science Foundation of China (51775547), the Fundamental Research Funds for the Central Universities (2015XKMS021), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

1. Bavkar S., Sahare S. (2013), PCA based single channel speech enhancement method for highly noisy environment, *Proceedings of International Conference on Advances in Computing*, pp. 1103–1107, Mysore, doi: 10.1109/ICACCI.2013.6637331.

2. Boll S. (1979), Suppression of acoustic noise in speech using spectral subtraction, *IEEE Transactions on Acoustics Speech & Signal Processing*, **27**(2): 113–120, doi: 10.1109/TASSP.1979.1163209.

3. Bouzid A., Ellouze N. (2016), Speech enhancement based on wavelet packet of an improved principal component analysis, *Computer Speech & Language*, **35**: 58–72, doi: 10.1016/j.csl.2015.06.001.

4. Chien J.T., Yang P.K. (2015), Bayesian factorization and learning for monaural source separation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**(1): 185–195, doi: 10.1109/TASLP.2015.2502141.

5. Coifman R.R., Wickerhauser M.V. (1992), Entropy-based algorithms for best basis selection, *IEEE Transactions on Information Theory*, **38**(2): 713–718, doi: 10.1109/18.119732.

6. Févotte C., Le Roux J., Hershey J.R. (2013), Non-negative dynamical system with application to speech and audio, *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3158–3162, Vancouver, doi: 10.1109/ICASSP.2013.6638240.

7. Gokhale M., Khanduja D.K. (2010), Time domain signal analysis using wavelet packet decomposition approach, *International Journal of Communications, Network and System Sciences*, **3**(3): 321–329, doi: 10.4236/ijcns.2010.33041.

8. Grancharov V., Samuelsson J., Kleijn B. (2006), On causal algorithms for speech enhancement, *IEEE Transactions on Speech & Audio Processing*, **14**(3): 764–773, doi: 10.1109/TSA.2005.857802.

9. Hansen J.H., Pellom B.L. (1998), An effective quality evaluation protocol for speech enhancement algorithms, *Proceedings of Fifth International Conference on Spoken Language Processing*, pp. 0917–0921, Sydney.

10. Islam M.S., Al Mahmud T.H., Khan W.U., Ye Z. (2019), Supervised single channel speech enhancement based on dual-tree complex wavelet transforms and nonnegative matrix factorization using the joint learning process and subband smooth ratio mask, *Electronics*, **8**(3): 353–371, doi: 10.3390/electronics8030353.

11. Krawczyk-Becker M., Gerkmann T. (2016), An evaluation of the perceptual quality of phase-aware single-channel speech enhancement, *Journal of the Acoustical Society of America*, **140**(4): EL364–EL369, doi: 10.1121/1.4965288.

12. Lai Y.-H., Chen F., Wang S.-S., Lu X., Tsao Y., Lee C.-H. (2016), A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation, *IEEE Transactions on Biomedical Engineering*, **64**(7): 1568–1578, doi: 10.1109/TBME.2016.2613960.

13. Lee D.D., Seung H.S. (1999), Learning the parts of objects by non-negative matrix factorization, *Nature*, **401**(6755): 788–791, doi: 10.1038/44565.

14. Lee S., Han D.K., Ko H. (2017), Single-channel speech enhancement method using reconstructive NMF with spectrotemporal speech presence probabilities, *Applied Acoustics*, **117**: 257–262, doi: 10.1016/j.apacoust.2016.04.024.

15. Li J., Sakamoto S., Hongo S., Akagi M., Suzuki Y.I. (2011), Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication, *Speech Communication*, **53**(5): 677–689, doi: 10.1016/j.specom.2010.04.009.

16. Li Y., Zhang X., Sun M. (2017), Robust Nonnegative matrix factorization with β-divergence for speech separation, *ETRI Journal*, **39**(1): 21–29, doi: 10.4218/etrij.17.0115.0122.

17. Luts H. *et al.* (2010), Multicenter evaluation of signal enhancement algorithms for hearing aids, *Journal of the Acoustical Society of America*, **127**(3): 1491–1505, doi: 10.1121/1.3299168.

18. Magron P., Virtane B. (2018), Expectation-maximization algorithms for Itakura-Saito nonnegative matrix factorization, *Proceedings of 2018 Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 856–860, Graz, doi: 10.21437/Interspeech.2018-1840.

19. Mavaddaty S., Ahadi S.M., Seyedin S. (2017), Speech enhancement using sparse dictionary learning in wavelet packet transform domain, *Computer Speech & Language*, **44**: 22–47, doi: 10.1016/j.csl.2017.01.009.

20. Mohammadiha N., Smaragdis P., Leijon A. (2013), Supervised and unsupervised speech enhancement using nonnegative matrix factorization, *IEEE Transactions on Audio, Speech, and Language Processing*, **21**(10): 2140–2151, doi: 10.1109/TASL.2013.2270369.

21. Mowlaee P., Saeidi R. (2014), Time-frequency constraints for phase estimation in single-channel speech enhancement, *Proceedings of 2014 14th International Workshop on Acoustic Signal Enhancement*, pp. 337–341, Juan-les-Pins, doi: 10.1109/IWAENC.2014.6954314.

22. Nakano M., Kameoka H., Le Roux J., Kitano Y., Ono N., Sagayama S. (2010), Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence, *Proceedings of 2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 283–288, Kittila, doi: 10.1109/MLSP.2010.5589233.

23. Nie S., Shan L., Wenju L., Xueliang Z., Jianhua T. (2018), Deep learning based speech separation via NMF-style reconstructions, *IEEE/ACM Transactions on Audio Speech & Language Processing*, **26**(11): 2043–2055, doi: 10.1109/TASLP.2018.2851151.

24. Panfili L. M., Haywood J., McCloy D.R., Souza P.E., Wright R.A. (2017), *The UW/NU Corpus, Version 2.0*, https://depts.washington.edu/phonlab/projects/uw-nu.php.

25. Rix A.W., Beerends J.G., Hollier M.P., Hekstra A.P. (2001), Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, *Proceedings of 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP)*, pp. 749–752, Salt Lake City, doi: 10.1109/ICASSP.2001.941023.

26. Saleem N., Khattak M.I.I., Ali M.Y., Shafi M. (2019), Deep neural network for supervised single-channel speech enhancement, *Archives of Acoustics*, **44**(1): 3–12, doi: 10.24425/aoa.2019.126347.

27. Saleem N., Khattak M.I., Shafi M. (2018), Unsupervised speech enhancement in low SNR environments via sparseness and temporal gradient regularization, *Applied Acoustics*, **141**: 333–347, doi: 10.1016/j.apacoust.2018.07.027.

28. Scalart P., Filho J.V. (1996), Speech enhancement based on a priori signal to noise estimation, *Proceedings of 1996 IEEE International Conference on Acoustics*, pp. 629–632, Atlanta, doi: 10.1109/ICASSP.1996.543199.

29. Sun D.L., Fevotte C. (2014), Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence, *Proceedings of 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP)*, pp. 6201–6205, Florence, doi: 10.1109/ICASSP.2014.6854796.

30. Sun P., Qin J. (2016), Wavelet packet transform based speech enhancement via two-dimensional SPP estimator with generalized gamma priors, *Archives of Acoustics*, **41**(3): 579–590, doi: 10.1515/aoa-2016-0056.

31. Taal C.H., Hendriks R.C., Heusdens R., Jensen J. (2011), An algorithm for intelligibility prediction of time–frequency weighted noisy speech, *IEEE Transactions on Audio, Speech, and Language Processing*, **19**(7): 2125–2136, doi: 10.1109/TASL.2011.2114881.

32. Varga A., Steeneken H.J. (1993), 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*, **12**(3): 247–251, doi: 10.1016/0167-6393(93)90095-3.

33. Varshney Y.V., Abbasi Z.A., Abidi M.R., Farooq O. (2017), Frequency selection based separation of speech signals with reduced computational time using sparse NMF, *Archives of Acoustics*, **42**(2): 287–295, doi: 10.1515/aoa-2017-0031.

34. Veisi H., Sameti H., Aroudi A. (2015), Hidden Markov model-based speech enhancement using multivariate Laplace and Gaussian distributions, *IET Signal Processing*, **9**(2): 177–185, doi: 10.1049/iet-spr.2014.0032.

35. Wang D., Jiang M., Niu F., Cao Y., Zhou C. (2018a), Speech Enhancement Control Design Algorithm for Dual-Microphone Systems Using β-NMF in a Complex Environment, *Complexity*, **2018**, Article ID 6153451, doi: 10.1155/2018/6153451.

36. Wang D., Chen J. (2018), Supervised speech separation based on deep learning: An overview, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **26**(10): 1702–1726, doi: 10.1109/TASLP.2018.2842159.

37. Wang D., Hansen J.H.L. (2018), Speech enhancement for cochlear implant recipients, *Journal of the Acoustical Society of America*, **143**(4): 2244–2254, doi: 10.1121/1.5031112.

38. Wang M., Zhang E., Tang Z. (2018b), Speech enhancement based on NMF under electric vehicle noise condition, *IEEE Access*, **6**: 9147–9159, doi: 10.1109/ACCESS.2018.2797165.

39. Wang S.S., Chern A., Tsao Y., Hung J.W., Lai Y.H., Su B. (2016), Wavelet speech enhancement based on nonnegative matrix factorization, *IEEE Signal Processing Letters*, **23**(8): 1101–1105, doi: 10.1109/LSP.2016.2571727.

40. Wang S.S. *et al.* (2015), Improving denoising autoencoder based speech enhancement with the speech parameter generation algorithm, *Proceedings of 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 365–369, Hong Kong, doi: 10.1109/APSIPA.2015.7415295.