WAVELET SPEECH ENHANCEMENT BASED ON NONNEGATIVE MATRIX FACTORIZATION

Syu-Siang Wang[1], *Alan Chern*[2], *Yu Tsao*[2], *Jeih-weih Hung*[4], *Xugang Lu*[3], *Ying-Hui Lai*[2], *Borching Su*[1]

[1]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[3]National Institute of Information and Communications Technology, Japan
[4]Dept of Electrical Engineering, National Chi Nan University, Taiwan

ABSTRACT

For most of the state-of-the-art speech enhancement (SE) techniques, a spectrogram is usually preferred than the respective time-domain raw data since it reveals more compact presentation together with conspicuous temporal information over a long time span. However, two problems can potentially cause distortions in the conventional nonnegative matrix factorization (NMF) based SE algorithms. One is related to the overlap-and-add (OLA) operation used in the short-time Fourier transform (STFT) based signal reconstruction, and the other is concerned with directly using the phase of the noisy speech as the phase of the enhanced speech in signal reconstruction. These two problems can cause information loss or discontinuity when comparing the original clean signal with the reconstructed signal. To solve these two problems, we propose a novel SE method that adopts the algorithms of discrete wavelet packet transform (DWPT) and NMF. In brief, the DWPT is first applied to split a time-domain speech signal into a series of subband signals without introducing any distortion. Then we exploit NMF to highlight the speech component for each subband. Finally, the enhanced subband signals are joined together via the inverse DWPT to reconstruct a noise-reduced signal in time domain. We evaluate the proposed DWPT-NMF based speech enhancement method on the Mandarin hearing in noise test (MHINT) task. Experimental results show that this new method behaves very well in prompting speech quality and intelligibility and it outperforms the conventional STFT-NMF based method.

Index Terms— short-time Fourier transform, discrete wavelet packet transform, NMF, speech enhancement

1. INTRODUCTION

Speech enhancement (SE) techniques, which estimate clean speech components in noise-corrupted utterances are employed to increase quality and intelligibility of speech signals. The speech interaction applications in noise are benefited from the applications of these techniques [1]. Two stages are included in most conventional spectrum-wise SE framework [2] explicitly or implicitly, viz. noise tracking and signal gain estimation (or filter estimation). In the noise tracking stage, the power of background noise from the noise-corrupted spectrogram is estimated. Several well-known noise tracking schemes have been proposed, for example, minimum statistics (MS) tracking [3], and improved minima controlled recursive averaging (IMCRA) [4]. In the signal gain estimation stage, the estimated noise power information is utilized to determine the gain factor on the noise-corrupted spectrogram to predict the clean speech components. Based on this idea, many famous algorithms have been proposed, such as spectral subtraction (SS) [5], Wiener filtering [6], minimum mean-square error log-spectral amplitude estimation (LSA) [7], maximum likelihood spectral amplitude estimation (MLSA) [8], and maximum a posteriori spectral amplitude estimation (MAPA) [9]. Recently, there are a lot of advance progress in exploring new techniques for SE, for example, deep neural network [10, 11], deep denoising auto-encoder [12–14], sparse coding [15], and nonnegative matrix factorization (NMF) [16, 17]. Most of these techniques consist of off-line and on-line phases in clean speech estimation. The off-line phase prepares a model to characterize the correlation between clean speech and noise; this model is then adopted in the on-line phase to recover the clean signals given the noisy inputs. In this study, we focus on the NMF-based SE method. The NMF algorithm approximates a data matrix $\mathbf{V}$ by a product of another two matrices, the basis matrix $\mathbf{W}$ and the encoding matrix $\mathbf{H}$, viz. $\mathbf{V} \approx \mathbf{WH}$, where all of the three matrices $\mathbf{V}$, $\mathbf{W}$ and $\mathbf{H}$ contain nonnegative entries only [18]. In NMF-based SE methods the data matrix $\mathbf{V}$ to be processed is usually the magnitude spectrogram of a speech utterance. Given a pre-trained and fixed basis spectral matrix $\mathbf{W}$ that consists of speech and noise components, the magnitude spectrogram $\mathbf{V}$ is factorized via NMF. The approximated result $\mathbf{WH}$ for $\mathbf{V}$ is further split into speech and noise components, both of which are used in the signal gain estimation for the SE method. Variations of NMF-based SE methods have been proposed, and several well-known algorithms have been developed, for example, segmental NMF (SNMF) [19], Bayesian NMF (BNMF) [20], and non-negative dynamical system (NDS) [21].

There are two potential limitations in the conventional NMF-based SE system. First, the spectrogram being analyzed in the NMF methods is mostly estimated by the well-known short-time Fourier transform (STFT). Despite the success of SE, the STFT operation brings moderate distortion to the time-domain signal primarily due to the obligatory segmentation and windowing processes. This distortion can be clearly observed when comparing the original clean signal with the reconstructed signal, which is from the inverse STFT with OLA. Second, most NMF-based SE methods restore clean speech spectral magnitudes given the noisy utterances, and the phase of the noisy speech is often adopted to reconstruct the enhanced speech signal [22, 23]. Clearly, the noisy phase may cause distortion on the enhanced speech signal. To overcome the above two limitations, this study proposes an NMF-based SE method in the wavelet domain. In the proposed method, a discrete wavelet packet transform (DWPT) [24] is first employed to decompose the time-domain signal into various subband signals. Squaring and framing processes are then applied to each subband signal. Next, an NMF-wise noise estimation scheme is used to estimate noise with respect to each subband for the subsequent signal gain estimation. Finally, all of the subband signals are updated via the individual gain and then passed through the inverse DWPT to produce the enhanced

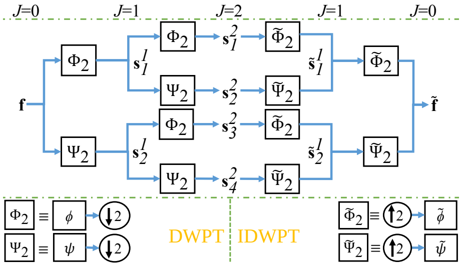


Fig. 1. An example of level-2 ($J = 2$) DWPT and IDWPT.

time-domain signal. We evaluate this DWPT-NMF based SE method on the MHINT task, and the experimental results confirm that this novel method outperforms the conventional STFT-based SE method and significantly promotes the speech quality and intelligibility under noise-corrupted situations.

2. DISCRETE WAVELET PACKET TRANSFORM

The DWPT and inverse DWPT (IDWPT) are performed by a set of well-defined low-pass and high-pass filters together with a down/up-sampling process, and they serve as a perfect (distortionless) analysis/synthesis for an arbitrary signal $\mathbf{f}$. Figure 1 depicts the flowchart of the concatenation of the DWPT and IDWPT with level 2. For the left side of Fig. 1, a (full-band) time signal $\mathbf{f}$ is decomposed into two subband signals carrying information of the low and high-frequency components. The length of each subband signal is half of that of the original full-band signal due to the factor-2 down-sampling operation. The decomposition operation is then applied again to each of the two subband signals, and four subband signals are generated accordingly. The DWPT process can be formulated in Eq. (1):

$$\mathbf{s}_b^J = DWPT_b^J\{\mathbf{f}\}, \quad b = 1, 2, 3, \cdots, 2^J, \tag{1}$$

where $\mathbf{s}_b^J$ denotes any subband signal produced by DWPT, J denotes the level of DWPT, and b refers to the subband index.

On the other hand, the IDWPT integrates the subband signals and reconstructs a full-band time signal $\tilde{\mathbf{f}}$. As shown in the right side of Fig. 1, a factor-2 up-sampling operation with the subsequent low-pass/high-pass filters are applied to the four level-2 subband signals to construct two level-1 subband signals, which in turn undergo the same up-sampling and filtering process to generate a full-band time signal $\tilde{\mathbf{f}}$. Thus the IDWPT can be formulated in Eq. (2):

$$\tilde{\mathbf{f}} = IDWPT^J\{\mathbf{s}\}, \tag{2}$$

where $\mathbf{s}$ denotes the set of all level-J subband signals $\{\mathbf{s}_b^J\}_{b=1}^{2^J}$. Provided a well-defined filter set (as symbolized by ϕ, $\tilde{\phi}$, ψ and $\tilde{\psi}$), the input signal will be identical to the reconstructed signal, namely

$$\tilde{\mathbf{f}} = IDWPT^J\{\mathbf{s}\} = \mathbf{f}. \tag{3}$$

More details of DWPT/IDWPT can be found in [24] and [25].

3. STFT-NMF SPEECH ENHANCEMENT SYSTEM

The conventional SE framework based on STFT-wise spectrogram modification [26] is illustrated in Fig. 2. The input real-valued time signal $\mathbf{f}$ is first converted to its complex-valued spectrogram $\mathbf{F}$ via STFT. Then the SE system compensates the magnitude part $|\mathbf{F}|$ of the spectrogram while keeping its phase part $\angle\mathbf{F}$ unaltered. Finally, the new spectrogram $\tilde{\mathbf{F}} = |\tilde{\mathbf{F}}| \exp(j\angle\mathbf{F})$ with an updated magnitude $|\tilde{\mathbf{F}}|$ and the original phase $\angle\mathbf{F}$ is converted back to the time domain via inverse STFT (ISTFT) to reconstruct the enhanced time signal $\tilde{\mathbf{f}}$.

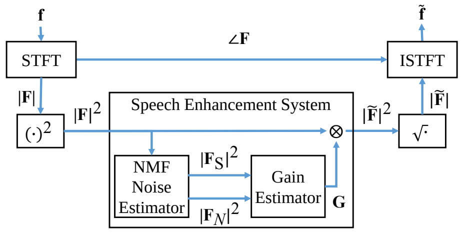


Fig. 2. The structure of a conventional STFT-based SE system.

From Fig. 2, the noise estimation block is performed by the NMF technique.There are off-line and on-line phases for NMF noise estimation. In the off-line phase, a noise-free speech spectral basis $\mathbf{W}_S$ obtained from the power spectrogram of clean speech utterances in the training set is performed by NMF. In addition, the pure noise spectral basis matrix $\mathbf{W}_N$ is obtained in the same manner as that of obtaining $\mathbf{W}_S$ using the speech-free noise in the training set. The concatenated basis, $\mathbf{W}_{SN} = [\mathbf{W}_S, \mathbf{W}_N]$ is further applied to the on-line phase to form the activation matrix, $\mathbf{H}_{SN} = [\mathbf{H}_S^\top, \mathbf{H}_N^\top]^\top$ for an input testing spectrogram $|\tilde{\mathbf{F}}|^2 = \mathbf{W}_{SN}\mathbf{H}_{SN}$. Finally, the gain that applies to the noise-corrupted spectrogram is represented as follows,

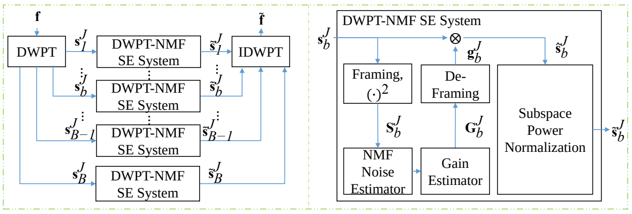
$$\mathbf{G} = \sqrt{(\mathbf{W}_S\mathbf{H}_S)./(\mathbf{W}_S\mathbf{H}_S + \mathbf{W}_N\mathbf{H}_N)}, \tag{4}$$

where "$\sqrt{\cdot}$" denotes an element-wise square root operation. It is noteworthy that when the power spectrogram is used for analysis, the gain function $\mathbf{G}$ shown in Eq. (4) behaves as a Wiener filter. This NMF-based SE method that applies to the STFT-derived spectrogram is denoted by "STFT-NMF" in later discussions for simplicity.

4. DWPT-NMF SPEECH ENHANCEMENT SYSTEM

In this study, we propose a novel SE method that adopts NMF-wise compensation directly on the time signals, while these time signals are in fact the DWPT (filtered and down-sampled) subband outputs for the original time signal. The block diagram of the overall framework of this SE method is depicted in Fig. 3(a). For simplicity, we use "DWPT-NMF" to stand for the proposed method.

As shown in Fig. 3(a), the DWPT is first applied to a noise-corrupted signal $\mathbf{f}$ to produce a set of subband signals $\{\mathbf{s}_b^J\}$. Then all of the subband signals are individually enhanced via NMF-wise compensation. Finally, these updated subband signals are joined together via the IDWPT and the enhanced signal $\tilde{\mathbf{f}}$ is reconstructed.



(a) The procedure of DWPT speech enhancement framework.

(b) DWPT-NMF-based speech enhancement system

Fig. 3. The structure for proposed DWPT-NMF SE system.

The SE procedures for each subband signal as depicted in Fig. 3(b) are clarified in the following. The subband signal $\mathbf{s}_b^J$ is first segmented into overlapped frames without further windowing (equivalent to using a rectangular window), and these frames are then arranged in sequence to be the columns of a matrix. Each element of the matrix is further squared in order to produce a nonnegative matrix, denoted by $\mathbf{S}_b^J$, for the subsequent NMF processing.

4.1. NMF-based noise estimation and gain estimation

Similar with the NMF noise estimation procedures stated in Section 3, two phases are adopted here. At the off-line phase, the nonnegative matrices $\mathbf{S}_b^J$ for the clean utterances in the training set with respect to a specific subband b are concatenated and then analyzed by NMF to create a speech basis matrix $\mathbf{W}_S^b$. Likewise, the noise basis matrix $\mathbf{W}_N^b$ associated with the subband b is created thereby using the speech-free noise in the training set. As for the on-line phase, the matrix $\mathbf{S}_b^J$ of the DWPT subband signal $\mathbf{s}_b^J$ for the input noise-corrupted utterance is NMF-encoded with the fixed basis matrix $\mathbf{W}^b = [\mathbf{W}_S^b \;\; \mathbf{W}_N^b]$ such that $\mathbf{S}_b^J \approx \mathbf{W}^b\mathbf{H}^b = \mathbf{W}_S^b\mathbf{H}_S^b + \mathbf{W}_N^b\mathbf{H}_N^b$, where $\mathbf{H}^b$ denotes the NMF encoding matrix, and $\mathbf{H}_S^b$ and $\mathbf{H}_N^b$ are respectively the speech and noise partitions of $\mathbf{H}^b$. Furthermore, the gain estimation as for subband b is achieved by

$$\mathbf{G}_b^J = \sqrt{(\mathbf{W}_S^b\mathbf{H}_S^b)./(\mathbf{W}_S^b\mathbf{H}_S^b + \mathbf{W}_N^b\mathbf{H}_N^b)}. \qquad (5)$$

Finally, the overlap-add process as the de-framing scheme is applied to $\mathbf{G}_b^J$ in Eq. (5) to obtain a gain sequence $\mathbf{g}_b^J$ that has the same size as the original subband signal $\mathbf{s}_b^J$. Therefore, $\mathbf{s}_b^J$ is modulated with $\mathbf{g}_b^J$ to produce a new subband signal as

$$\hat{\mathbf{s}}_b^J = \mathbf{s}_b^J . \times \mathbf{g}_b^J . \qquad (6)$$

Notably, since de-framing is processed on the gain estimate $\mathbf{G}_b^J$, the possible distortion caused by STFT does not occur here.

4.2. Subband power normalization

In order to compensate for the noise effect, a power normalization procedure is applied to the enhanced subband signal $\hat{\mathbf{s}}_b^J$ in Eq. (6). At the off-line phase, we concatenate the DWPT subband-b signals associated with all the clean speech utterances in the training set as a clean samples, and then we calculate the root mean square (rms) value for these samples, denoted by $\sigma_{b,c}$. At the on-line phase, the rms value of $\hat{\mathbf{s}}_b^J$ is also computed and denoted by $\hat{\sigma}_b$, and then we obtain the power normalized subband signal by $\tilde{\mathbf{s}}_b^J = \frac{\sigma_{b,c}}{\hat{\sigma}_b}\hat{\mathbf{s}}_b^J$. As a result, the power of the enhanced subband signal $\tilde{\mathbf{s}}_b^J$ is always equal to $\sigma_{b,c}^2$, regardless of different noise-corrupted signals to be enhanced.

5. EXPERIMENTS

5.1. Experimental setup

Evaluation experiments were conducted on the MHINT database [27], which contained 240 clean utterances recorded in a soundbooth with the background noise level below 45 dB sound pressure level and pronounced by a Mandarin male speaker at a sampling rate of 16 kHz. In this study, these utterances were down-sampled to be 8-kHz data. The averaged length of each utterance was around 3 seconds. Among these 240 utterances, 10 utterances were selected as the training data and another 50 utterances were the testing data for the SE task. In addition, eight types of noise: subway, exhibition, car, street, restaurant, babble, airport, and train-station, drawn from Aurora-2 database [28] were artificially added to the clean testing utterances to generate the noise-corrupted counterparts at six signal-to-noise ratios (SNRs) ranging from 20 dB to -5 dB with a 5-dB interval. Furthermore, four types of stationary noise, namely 'subway', 'exhibition', 'car', and 'street', drawn from Aurora-2 database were selected and used together to create the noise basis matrices of NMF at the off-line phase for noise estimation. Finally, the number of columns of each speech basis matrix ($\mathbf{W}_S$ and $\mathbf{W}_S^b$) was set to 40, while that of each noise basis matrix ($\mathbf{W}_N$ and $\mathbf{W}_N^b$) was set to 160.

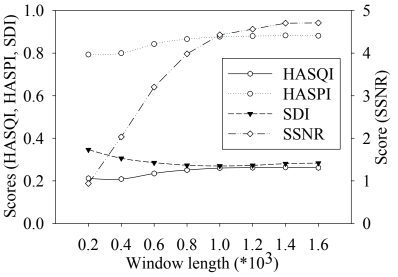


Fig. 4. The averaged scores of HASQI, HASPI, SDI and SSNR for DWPT-NMF with different framing window lengths

The applied frame size and frame shift for STFT-NMF were 256 samples and 80 samples, respectively. For the proposed DWPT-NMF, a 20-sample frame shift was used, while the frame size was varied. In addition, the level of DWPT/IDWPT was set to 3.

5.2. Evaluation methods

The SE methods were evaluated by the quality test in terms of the hearing aids speech quality index (HASQI) [29], and the perceptual test in terms of the hearing aids speech perception index (HASPI) [30]. Notably, HASQI and HASPI, resepctively, are developed to evaluate sound quality and perception for both hearing impaired patients and normal hearing people. It has been confirmed that these evaluations provide considerably high correlation scores with human quality assessment and perception. The HASQI and HASPI scores are both ranged from 0 to 1. Higher scores of HASQI and HASPI correspond to better sound quality and intelligibility, respectively.

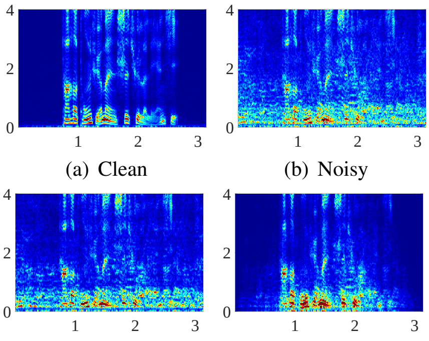
5.3. Performance evaluation

We first investigated the effects of STFT/ISTFT and DWPT/IDWPT on SE processes in Table 1, which lists the results of STFT/ISTFT(N), STFT/ISTFT(S), and DWPT/IDWPT on HASQI, HASPI and mean-squared error (MSE). In the experiments, we prepared paired utterances: a clean utterance and its noisy version. Here the noise utterance was artificially generated by contaminating the clean utterance with the restaurant noise at 0 dB SNR. We tested the transformation performance of STFT/ISTFT. Consider an SE system with STFT/ISTFT: we first assumed that the SE system can perfectly restore the clean magnitude given the noisy magnitude, and considered two conditions for the phase part: (a) using the noisy phase as the restored phase (denoted as STFT/ISTFT(N)); (b) using the clean phase as the restored phase (denoted as STFT/ISTFT(S)); clearly, the condition (b) represented a perfect restoration in the frequency domain. In addition to STFT/ISTFT, we tested the transformation performance of DWPT/IDWPT. For a fair comparison, we assumed that the SE system can perfectly restore clean speech input in each subband given the noisy input for the DWPT/IDWPT results.

From Table 1, the same HASPI score for all three systems suggesting that the phase part has negligible effect on the intelligibility. However, STFT/ISTFT(N) has lower HASQI score than

Table 1. Transformed efficiency of STFT and DWPT on HASQI, HASPI, and MSE.

Indexes	HASQI	HASPI	MSE
STFT/ISTFT(N)	0.623	1.000	0.634
STFT/ISTFT(S)	0.991	1.000	0.523
DWPT/IDWPT	1.000	1.000	0.000



(a) Clean

(b) Noisy

(c) STFT-NMF

(d) DWPT-NMF

Fig. 5. The spectrograms of (a) Clean, (b) Noisy, (c) STFT-NMF and (d) DWPT-NMF. In the figures the horizontal axis is for time in second and the vertical axis for frequency in kHz.

STFT/ISTFT(S), suggesting that the inaccurate phase information may distort the restored signals. Next, when compared to DWPT/IDWPT, STFT/ISTFT(S) gives a higher MSE value and a lower HASQI score. Please note that STFT/ISTFT(S) has perfect magnitude and phase information; the result confirms that STFT/ISTFT could generate distortions on the enhanced signals. Finally, these results demonstrate that DWPT/IDWPT outperforms both STFT/ISTFT(N) and STFT/ISTFT(S), suggesting that DWPT/IDWPT can be used as a plug-in unit in many advanced SE algorithms to further improve the performance.

Table 2. The HASQI results at six SNR conditions.

SNR (dB)	**20**	**15**	**10**	**5**	**0**	-5
Baseline	**0.452**	0.359	0.261	0.163	0.084	0.036
STFT-NMF	0.418	0.344	0.261	0.175	0.097	0.043
DWPT-NMF	0.447	**0.394**	**0.328**	**0.258**	**0.178**	**0.105**

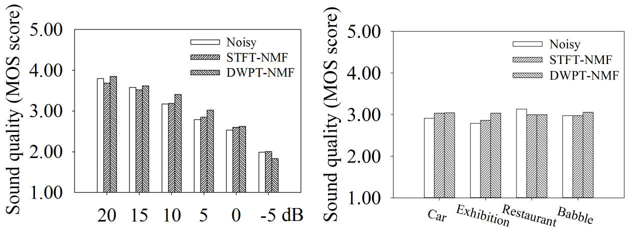
Second, the effect of different assignments of the frame size in the proposed DWPT-NMF was investigated. The corresponding evaluation was performed on the noise-corrupted data with four noise types and six SNR levels, and the results in terms of HASQI, HASPI, speech distortion index (SDI) and segmental SNR (SSNR) were shown in Fig. 4. From this figure, we observed that increasing the length of the framing window from 200 to 1000 always brings better results in all of the evaluation metrics. However, most of the metric scores reached a peak by further enlarging the frame size around 1000. One possible reason is the curse of dimensionality issue that large input feature dimensions cause imperfect basis matrix ($\mathbf{W}_S$ and $\mathbf{W}_N$) estimation. As a result, a frame with 1000 samples was applied to DWPT-NMF for the following experiments.

Third, the quality of the restored spectrogram for an utterance contaminated by babble noise at an SNR of 0 dB was examined in Fig. 5. Comparing these spectrograms, we can see that the conventional STFT-NMF exhibits higher speech distortion as well as more noise residues in the restored spectrogram than DWPT-NMF. Therefore, DWPT-NMF was shown to be superior to STFT-NMF.

Finally, we compared STFT-NMF and DWPT-NMF in terms of the HASQI and HASPI metric scores associated with the enhanced signals. Tables 2 and 3 listed the corresponding results at six SNR cases while averaged over four noise types. From these tables, we found that DWPT-NMF gives rise to higher HASQI and HASPI scores than STFT-NMF and the unprocessed baseline in almost all cases revealing that DWPT-NMF was quite effective in improving both the quality and intelligibility of speech signals. To further confirm the significance of improvements, the one-way analysis of variance (ANOVA) and Tukey post-hoc comparisons were used to analyze the HASQI and HASPI metric scores of STFT-NMF and DWPT-NMF with baseline over all noise conditions (eight noise types with six SNR conditions). The ANOVA results confirmed that both HASQI and HASPI scores differed significantly with p-values less than 0.001 while comparing DWPT-NMF with STFT-NMF; the Tukey post-hoc comparisons further verified the significant differences for HASQI scores: (baseline, STFT-NMF), and (STFT-NMF, DWPT-NMF), and for HASPI scores: (baseline, STFT-NMF), and (baseline, DWPT-NMF). In addition to objective evaluations, we also conducted subjective listening tests for the baseline, STFT-NMF, and the proposed DWPT-NMF. The subjective listening test had a single-blind design that seven un-trained but experienced normal hearing subjects did not know which speech enhancement method was used. The TDH-50P earphone [31] commonly used in audiometer for hearing test provided the original noisy and processed speech sounds to the human subjects, and was calibrated to the 65 dB SPL before the listening test based on ANSI S3.7 standard [32]. Six SNRs with four noise types (car, exhibition, restaurant, and babble) were used to form the test set. All measurement procedures were performed in a quiet environment where the background noise level was below 50 dB SPL. Fig. 6 (a) shows the MOS scores at six SNR conditions while averaged four noise types. From the figure, we first observe that the MOS scores for three processed techniques are consistently decreased along with SNRs. In addition, the proposed DWPT-NMF technique outperforms baseline and STFT-NMF in most SNRs except that at -5 dB. The averaged MOS scores in four noise types are listed in the Fig. 6 (b). The quality of DWPT-NMF reconstructed signals has the highest score over most noise types except that in restaurant. These results demonstrate that the proposed DWPT-NMF SE system can recover high-quality clean utterances from noisy input in most of noise environments.

Table 3. The HASPI results at six SNR conditions.

SNR (dB)	**20**	**15**	**10**	**5**	**0**	-5
Baseline	0.998	0.994	0.976	0.881	0.541	0.164
STFT-NMF	0.998	0.994	0.977	0.893	0.604	0.212
DWPT-NMF	**0.999**	**0.998**	**0.993**	**0.977**	**0.900**	**0.618**



(a) On six SNR conditions.

(b) On four noise types.

Fig. 6. MOS scores for Noisy, STFT-NMF, and DWPT-NMF.

6. CONCLUSION

This study proposed a novel SE framework based on DWPT and NMF. It is shown that DWPT serves as a better choice than the conventional STFT in the preparation of the data for the subsequent NMF-wise enhancement scheme. The proposed DWPT-NMF SE framework provides the noise-corrupted speech with a significant improvement in both quality and intelligibility. Please note that the main focus of the present study is to validate that the DWPT can be a suitable alternative processing for STFT for NMF-based SE. In the future, we will explore the combination of DWPT with advanced NMF methods (such as SNMF, BNMF, and NDS) to see if further improvement can be achieved.

7. REFERENCES

[1] W. Hartmann, A. Narayanan, E. Fosler-Lussier, and D. Wang, "A direct masking approach to robust ASR," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 1993–2005, 2013.

[2] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2005.

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

[4] I. Cohen, "Speech enhancement using a noncausal a priori snr estimator," *IEEE Signal Processing Letters*, vol. 11, no. 9, pp. 725–728, 2004.

[5] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[6] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *ICASSP*, pp. 629–632, 1996.

[7] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. ICASSP*, pp. 789–792, 1999.

[8] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[9] T. Lotter and P. Vary, "Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model," *EURASIP journal on applied signal processing*, vol. 2005, pp. 1110–1126, 2005.

[10] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*, pp. 7092–7096, 2013.

[11] 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.

[12] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Ensemble modeling of denoising autoencoder for speech spectrum restoration," in *Proc. INTERSPEECH*, vol. 14, pp. 885–889, 2014.

[13] S.-S. Wang, H.-T. Hwang, Y.-H. Lai, Y. Tsao, X. Lu, H.-M. Wang, and B. Su, "Improving denoising auto-encoder based speech enhancement with the speech parameter generation algorithm," in *Proc. APSIPA*, pp. 365–369, 2015.

[14] M. Sun, X. Zhang, T. F. Zheng, *et al.*, "Unseen noise estimation using separable deep auto encoder for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 93–104, 2016.

[15] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech enhancement using generative dictionary learning," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1698–1712, 2012.

[16] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[17] 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.

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

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

[20] J.-T. Chien and P.-K. Yang, "Bayesian factorization and learning for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 185–195, 2016.

[21] C. Févotte, J. Le Roux, and J. R. Hershey, "Non-negative dynamical system with application to speech and audio," in *Proc. ICASSP*, pp. 3158–3162, 2013.

[22] P. Mowlaee, R. Saeidi, and Y. Stylanou, "Interspeech 2014 special session: Phase importance in speech processing applications," in *Proc. INTERSPEECH*, 2014.

[23] P. Mowlaee and J. Kulmer, "Phase estimation in single-channel speech enhancement: Limits-potential," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 8, pp. 1283–1294, 2015.

[24] M. Gokhale, D. K. Khanduja, *et al.*, "Time domain signal analysis using wavelet packet decomposition approach," *Int'l J. of Communications, Network and System Sciences*, vol. 3, no. 03, p. 321, 2010.

[25] D. D. Ariananda, M. K. Lakshmanan, and H. Nikookar, "An investigation of wavelet packet transform for spectrum estimation," *arXiv preprint arXiv:1304.3795*, 2013.

[26] N. Mohammadiha, J. Taghia, and A. Leijon, "Single channel speech enhancement using bayesian nmf with recursive temporal updates of prior distributions," in *Proc. ICASSP*, pp. 4561–4564, IEEE, 2012.

[27] L. L. Wong, S. D. Soli, S. Liu, N. Han, and M.-W. Huang, "Development of the mandarin hearing in noise test (MHINT)," *Ear and hearing*, vol. 28, no. 2, pp. 70S–74S, 2007.

[28] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[29] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (HASQI)," *Journal of the Audio Engineering Society*, vol. 58, no. 5, pp. 363–381, 2010.

[30] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, 2014.

[31] M. Valente, L. G. Potts, and L. M. Valente, "Differences and intersubject variability of loudness discomfort levels measured in sound pressure level and hearing level for TDH-50P and ER-3A earphones," *Journal of the American Academy of Audiology*, vol. 8, no. 1, 1997.

[32] ANSI (1995). S3.7-1995 (R 2008), "ANSI S3.7-American national standard method for coupler calibration of earphones," American National Standards Institute, New York.