SPEECH ENHANCEMENT USING SEGMENTAL NONNEGATIVE MATRIX FACTORIZATION

Hao-Teng Fan[1], Jeih-weih Hung [1], Xugang Lu[2], Syu-Siang Wang[3], and Yu Tsao[3]

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

ABSTRACT

The conventional NMF-based speech enhancement algorithm analyzes the magnitude spectrograms of both clean speech and noise in the training data via NMF and estimates a set of spectral basis vectors. These basis vectors are used to span a space to approximate the magnitude spectrogram of the noise-corrupted testing utterances. Finally, the components associated with the clean-speech spectral basis vectors are used to construct the updated magnitude spectrogram, producing an enhanced speech utterance. Considering that the rich spectral-temporal structure may be explored in local frequency and time-varying spectral patches, this study proposes a segmental NMF (SNMF) speech enhancement scheme to improve the conventional frame-wise NMF-based method. Two algorithms are derived to decompose the original nonnegative matrix associated with the magnitude spectrogram; the first algorithm is used in the spectral domain and the second algorithm is used in the temporal domain. When using the decomposition processes, noisy speech signals can be modeled more precisely, and spectrograms regarding the speech part can be constituted more favorably compared with using the conventional NMF-based method. Objective evaluations using perceptual evaluation of speech quality (PESQ) indicate that the proposed SNMF strategy increases the sound quality in noise conditions and outperforms the well-known MMSE log-spectral amplitude (LSA) estimation.

Index Terms—nonnegative matrix factorization, NMF, speech enhancement, sub-band processing, patch processing

1. INTRODUCTION

The primary objective of single channel speech enhancement is to alleviate the effect of noise in speech signals and to improve speech quality. In contemporary communication applications, speech enhancement plays an important role as a pre-processor, suppressing speech distortion caused by noise. Generally, speech enhancement methods can be categorized into two broad classes: unsupervised and supervised. Unsupervised methods include a wide range of approaches such as spectral subtraction (SS) [1]-[3], Wiener [4, 5] and Kalman filtering [6], short-time spectral amplitude (STSA) estimation [7], estimations based on super-Gaussian prior distributions for DFT coefficients of speech [8, 9], and schemes based on periodic models of speech signals [10]. In these unsupervised methods, statistical models are used for both speech and noise, and the model parameters regarding clean speech are estimated via noisy observations without any prior information concerning the noise type or speaker identity. However, a challenge of most unsupervised speech enhancement methods is estimating the power spectral density (PSD) of noise [11, 12], which is particularly difficult when the interfering noise is non-stationary. By contrast, supervised speech enhancement methods use distinct models for clean speech and noise signals, and the parameters of each model are estimated using the respective samples. An interaction model is then defined by combining speech and noise models, and finally the noise reduction task is conducted. Some examples of the supervised methods include codebook-based approaches [13] and hidden Markov model (HMM) based methods [14]. One advantage of these methods is that they do not need to estimate the power spectral density (PSD) of noise using a separate algorithm. The supervised approaches have been shown to produce enhanced speech signals with higher quality than those produced using the unsupervised methods. This is expected because more prior information is incorporated into the algorithm when using supervised methods than it is when using unsupervised methods, and the considered models are trained for each specific type of signal. The required prior information regarding noise type (and speaker identity, in some cases) can be provided by the user or a separate acoustic environment classification algorithm [15], or obtained using a built-in classification scheme [13].

A successful supervised speech enhancement algorithm is based on the nonnegative matrix factorization (NMF) technique. With NMF, the basis spectra for clean speech and noise are first estimated using the corresponding training samples. Next, both speech and noise spectral bases are jointly used to approximately span the magnitude spectrogram of the noise-corrupted utterance. Finally, the portion spanned by the clean speech spectra bases is extracted and used to produce the enhanced testing utterance. Because of its effective performance levels, NMF-based speech enhancement has been extensively investigated [16]-[18].

However, there is still room for improvement in conventional NMF-based speech enhancement: First, since NMF is directly applied to frame-wise full-band spectra, the learned basis vectors may omit discriminative information embedded in different frequency components, and thus decomposing the spectrum to low and high frequency portions in order to focus on local frequency structure appears to catch the difference between speech and noise more accurately. Second, the frame-wise processing of NMF may lack an invariant structure for describing the time-varying characteristics of spectra, and thus using a temporal window to embrace the neighboring spectra for the subsequent NMF processing could help to catch the invariant or stable structure. In light of these observations, this study focuses on exploring basis functions that encode discriminative and invariant structure of speech and noise.

In this study, a segmental NMF (SNMF) speech enhancement scheme was proposed for improving the conventional NMF-based method. We presented two instantiations of SNMF based on the unique structures of speech signals: 1) distinct characteristics are presented in high and low frequency bands, and 2) temporal cues are vital. First, sub-band processing was performed to divide spectrograms into sub-band blocks; each sub-band spectrogram was

then individually enhanced via the standard NMF process. Second, patch processing was conducted to incorporate the temporal information regarding the speech spectra. In the patch processing, each spectrum frame was augmented into a spectrum patch with neighboring spectrum frames, and then the standard NMF process was applied to the spectrum patches to perform speech enhancement. According to the experimental results, SNMF with these two instantiations provides a more favorable level of performance than conventional NMF-based method in the perceptual estimation of speech quality (PESQ) evaluation [26] across various signal-to-noise ratio (SNR) conditions.

2. SPEECH ENHANCEMENT USING THE NMF TECHNIQUE

This section provides a review of the NMF technique and the related procedure specifically for speech enhancement.

2.1 The NMF technique

NMF is a technique that involves projecting the columns of a nonnegative matrix onto a space spanned by a set of basis vectors. NMF has been widely used as a source separation technique applied to monaural mixtures [19, 20]. Recently, NMF has also been used to estimate clean speech from a noisy observation [21]–[23]. When applied to speech source separation, a sufficient separation can be expected only when speaker-dependent bases are learned. By contrast, regarding noise reduction, even if a general speaker-independent basis matrix of speech is learned, a satisfactory level of enhancement can be achieved [24]. Nevertheless, in some cases where the basis matrices of speech and noise are similar (e.g., speech degraded with multi-talker babble noise), additional constraints are typically imposed to the conventional NMF in order to improve noise reduction [21].

Given a nonnegative data matrix $\mathbf{V} \in \mathbf{R}^{N\times M}$, NMF is used to calculate two nonnegative matrices $\mathbf{W} \in \mathbf{R}^{N\times r}$ and $\mathbf{H} \in \mathbf{R}^{r\times M}$, such that

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

One of the objective functions to be minimized in NMF to obtain the nonnegative matrices **W** and **H** in Eq. (1) is:

$$\mathrm{J}(\mathbf{W},\mathbf{H}) = \|\mathbf{V}-\mathbf{WH}\|^2 = \textstyle\sum_{i,j}(\mathbf{V}_{ij}-(\mathbf{WH})_{ij})^2. \tag{2}$$

W and **H** are often called the basis matrix and encoding matrix, respectively, because each column of data matrix **V** in Eq. (1) approximately lies in the space spanned by the columns of **W** with the (encoded) coordinate coefficients stored in the column of **H**. The column vectors of **W** are also known as the "building blocks" for data matrix **V** because these vectors serve as the basis for the data vector space (provided they are linearly independent). The number of basis vectors, *r*, is often chosen to be lower than *N* and *M*, which are the size of each sample (data vector) and the total number of samples, respectively. In general, the two matrices **W** and **H** are obtained by iteratively minimizing the objective function defined in NMF as in Eq. (1). Additional details regarding the computation of **W** and **H** are described in [16]–[18].

2.2 NMF-based speech enhancement

NMF-based speech enhancement consists of a training stage and an enhancement stage. In the training phase, it is assumed that a clean speech magnitude spectrogram, $\mathbf{V}_s^{Tn} \in \mathbf{R}^{N_f\times M_s}$, is available, where N_f and M_s denote the numbers of frequency bins and speech frames, respectively. Notably, $\mathbf{V}_s^{Tn}$ can be formed by (horizontally) concatenating the magnitude spectrograms of clean training utterances, Each column of $\mathbf{V}_s^{Tn}$, denoted by X_t^{Tn}, is an $N_f \times 1$ vector representing the spectral vector of a specific time frame in the ordered (magnitude) spectrogram series.

By applying NMF, $\mathbf{V}_s^{Tn}$ can be approximated as

$$\mathbf{V}_s^{Tn} \approx \mathbf{W}_s^{Tn}\mathbf{H}_s^{Tn}, \tag{3}$$

where $\mathbf{W}_s^{Tn} \in \mathbf{R}^{N_f\times r}$ and $\mathbf{H}_s^{Tn} \in \mathbf{R}^{r\times M_s}$, in which r is the number of basis vectors as the columns of $\mathbf{W}_s^{Tn}$ chosen to represent each source spectral vector X_t^{Tn}. Each column of $\mathbf{W}_s^{Tn}$ is one of the spectral "building blocks".

Likewise, a speech-free noise magnitude spectrogram, $\mathbf{V}_n^{Tn} \in \mathbf{R}^{N_f\times M_n}$, is derived from a long noise segment, where N_f is the number of frequency bins and M_n is the number of noise frames. According to NMF, we have

$$\mathbf{V}_n^{Tn} \approx \mathbf{W}_n^{Tn}\mathbf{H}_n^{Tn}, \tag{4}$$

where $\mathbf{W}_n^{Tn} \in \mathbf{R}^{N_f\times r}$ and $\mathbf{H}_n^{Tn} \in \mathbf{R}^{r\times M_n}$.

In particular, during the training phase the standard NMF process is performed to iteratively update the four matrices, $\mathbf{W}_s^{Tn}$, $\mathbf{V}_s^{Tn}$, $\mathbf{W}_n^{Tn}$ and $\mathbf{V}_n^{Tn}$ by minimizing $\|\mathbf{V}_s^{Tn}-\mathbf{W}_s^{Tn}\mathbf{H}_s^{Tn}\|^2$ and $\|\mathbf{V}_n^{Tn}-\mathbf{W}_n^{Tn}\mathbf{H}_n^{Tn}\|^2$, respectively, as in Eq. (2).

In the enhancement phase, the two basis spectral matrices obtained in the training phase, $\mathbf{W}_s^{Tn}$ and $\mathbf{W}_n^{Tn}$, are assumed to continue providing a suitable basis functions for describing speech and noise in the noise-corrupted testing utterances. The two matrices are horizontally concatenated to form a double-wide matrix, $\mathbf{W}_c^{Tn} = [\mathbf{W}_s^{Tn}\ \mathbf{W}_n^{Tn}]$, and thus $\mathbf{W}_c^{Tn} \in \mathbf{R}^{N_f\times 2r}$, which acts as the basis matrix in the NMF process as described in Eq. (1) for approximating a data matrix. This data matrix corresponds to the magnitude spectrogram of a testing utterance that is mixed with clean speech and noise and is denoted by $\mathbf{V}_{mix}^{Tt}$, and can be approximated via NMF:

$$\mathbf{V}_{mix}^{Tt} \approx \mathbf{W}_c^{Tn}\mathbf{H}_c^{Tn} = [\mathbf{W}_s^{Tn}\ \mathbf{W}_n^{Tn}]\begin{bmatrix}\mathbf{H}_s^{Tt}\\ \mathbf{H}_n^{Tt}\end{bmatrix}. \tag{5}$$

Different from the training phase, here the basis matrix $\mathbf{W}_c^{Tn} = [\mathbf{W}_s^{Tn}\ \mathbf{W}_n^{Tn}]$ remains unchanged while the encoding matrix $\mathbf{H}_c^{Tn} = [\mathbf{H}_s^{Tt\prime}\ \mathbf{H}_n^{Tt\prime}]'$ is iteratively updated in order to achieve a better approximation. Finally, the component associated with the clean-speech basis spectra constitutes the enhanced magnitude spectrogram for the testing utterances:

$$\mathbf{V}_s^{Tt} \approx \mathbf{W}_s^{Tn}\mathbf{H}_s^{Tt}. \tag{6}$$

Finally, the enhanced magnitude spectrogram $\mathbf{V}_s^{Tt}$ in Eq. (6) together with the original phase spectrogram is converted to the time domain and the enhanced testing utterance is obtained accordingly.

3. PROPOSED SEGMENTAL NMF SPEECH ENHANCMENT SCHEME

Speech signals exhibit a unique spectral-temporal structure. In the spectral domain, speech signals exhibit distinct characteristics in the high and low frequency bands. In particular, major speech components are primarily located at lower frequencies. In addition, speech is a time-varying signal, and its temporal information plays a crucial role in identifying the corresponding characteristics. This section develop two speech enhancement approaches based on the concept of NMF, which decompose the non-negative matrix associated with the magnitude spectrogram of speech, namely sub-band and patch processing approaches. Figure 1 shows the applications of the two approaches to the spectrogram. Details are presented in the following discussion.

3.1 Spectral-domain SNMF

In the procedure shown in Eqs. (3) and (4), $\mathbf{V}_s^{Tn}$ and $\mathbf{V}_n^{Tn}$ are processed in a full-band manner. Here, sub-band processing was used

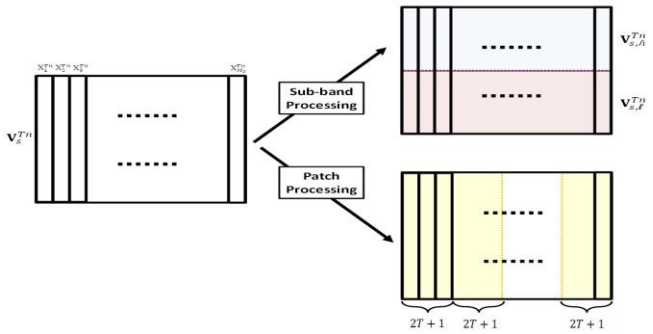

Fig 1. The proposed structural NMF with the sub-band and patch processing.

to segment the full-band speech and noise magnitude spectrograms into sub-bands, and this approach is called spectral-domain SNMF (denoted by SNMF (S) as a short-hand notation hereafter). First, the speech magnitude spectrogram, $\mathbf{V}_s^{Tn}$, that corresponds to clean training utterances as noted in subsection 2.2, is separated into $\mathbf{V}_{s,\ell}^{Tn}$ and $\mathbf{V}_{s,\hbar}^{Tn}$ (viz. $\mathbf{V}_s^{Tn} = [\mathbf{V}_{s,\hbar}^{Tn\prime}\ \mathbf{V}_{s,\ell}^{Tn\prime}]'$), which are factorized via NMF:

$$\mathbf{V}_{s,\ell}^{Tn} \approx \mathbf{W}_{s,\ell}^{Tn}\mathbf{H}_{s,\ell}^{Tn}, \tag{7}$$

$$\mathbf{V}_{s,\hbar}^{Tn} \approx \mathbf{W}_{s,\hbar}^{Tn}\mathbf{H}_{s,\hbar}^{Tn}, \tag{8}$$

where $\mathbf{V}_{s,\ell}^{Tn} \in \mathbf{R}^{N_{f,\ell}\times M_s}$, $\mathbf{W}_{s,\ell}^{Tn} \in \mathbf{R}^{N_{f,\ell}\times r}$, $\mathbf{H}_{s,\ell}^{Tn} \in \mathbf{R}^{r\times M_s}$, $\mathbf{V}_{s,\hbar}^{Tn} \in \mathbf{R}^{N_{f,\hbar}\times M_s}$, $\mathbf{W}_{s,\hbar}^{Tn} \in \mathbf{R}^{N_{f,\hbar}\times r}$, and $\mathbf{H}_{s,\hbar}^{Tn} \in \mathbf{R}^{r\times M_s}$. The speech-free noise magnitude spectrograms are prepared in the same sub-band manner as the speech part. Accordingly, we obtain the sub-band basis matrices, $\mathbf{W}_{n,\ell}^{Tn}$ and $\mathbf{W}_{n,\hbar}^{Tn}$, for the noise part.

In the enhancement phase, the magnitude spectrogram of each testing utterance is divided into high and low frequency bands, $\mathbf{V}_{mix,\ell}^{Tt}$ and $\mathbf{V}_{mix,\hbar}^{Tt}$. NMF is then applied to $\mathbf{V}_{mix,\ell}^{Tt}$ and $\mathbf{V}_{mix,\hbar}^{Tt}$ individually with the fixed spectral basis matrices $\mathbf{W}_{s,\ell}^{Tn}$ and $\mathbf{W}_{s,\hbar}^{Tn}$ prepared in the training phase. Consequently, we have

$$\mathbf{V}_{mix,\ell}^{Tt} \approx \begin{bmatrix}\mathbf{W}_{s,\ell}^{Tn} & \mathbf{W}_{n,\ell}^{Tn}\end{bmatrix}\begin{bmatrix}\mathbf{H}_{s,\ell}^{Tt}\\ \mathbf{H}_{n,\ell}^{Tt}\end{bmatrix}, \tag{9}$$

$$\mathbf{V}_{mix,\hbar}^{Tt} \approx \begin{bmatrix}\mathbf{W}_{s,\hbar}^{Tn} & \mathbf{W}_{n,\hbar}^{Tn}\end{bmatrix}\begin{bmatrix}\mathbf{H}_{s,\hbar}^{Tt}\\ \mathbf{H}_{n,\hbar}^{Tt}\end{bmatrix}. \tag{10}$$

Finally, with the estimated encoding matrices $\mathbf{H}_{s,\ell}^{Tt}$ and $\mathbf{H}_{s,\hbar}^{Tt}$ associated with the speech part for the high and low sub-bands, respectively, we can obtain

$$\mathbf{V}_{s,\ell}^{Tt} \approx \mathbf{W}_{s,\ell}^{Tn}\mathbf{H}_{s,\ell}^{Tt}. \tag{11}$$

$$\mathbf{V}_{s,\hbar}^{Tt} \approx \mathbf{W}_{s,\hbar}^{Tn}\mathbf{H}_{s,\hbar}^{Tt}. \tag{12}$$

Finally, we concatenate the two matrices in Eqs. (11) and (12) to obtain the updated full-band spectrogram as $\mathbf{V}_s^{Tt} = [\mathbf{V}_{s,\hbar}^{Tt\prime}\mathbf{V}_{s,\ell}^{Tt\prime}]'$.

3.2 Temporal-domain SNMF

The goal of the patch processing is to capture the temporal information of speech signals. Thus the clean speech magnitude spectrogram, $\mathbf{V}_s^{Tn}$, established by the entire set of speech signals is first segmented into groups to be processed further. Speaking in detail, a sliding window is applied to $\mathbf{V}_s^{Tn}$ to capture its temporal information, forming a patch Y_t^{Tn} at each time frame instant t by *vertically* concatenating $2T+1$ neighboring frames of $\mathbf{V}_s^{Tn}$. That is, $\mathrm{Y}_t^{Tn} = [\mathrm{X}_{t-T}^{Tn}; .. \mathrm{X}_t^{Tn} \ldots; \mathrm{X}_{t+T}^{Tn}]$, where X_t^{Tn} is the t^{th} column of $\mathbf{V}_s^{Tn}$. Specifically, we set $\mathrm{X}_{t'}^{Tn} = \mathrm{X}_1^{Tn}$ for $t' \le 0$ and $\mathrm{X}_{t'}^{Tn} = \mathrm{X}_{M_s}^{Tn}$ for $t' \ge M_s$. An extended speech spectrogram which contains the patches at different time instants is thus created: $\mathbf{V}_{s,P}^{Tn} = [\mathrm{Y}_1^{Tn}, .. \mathrm{Y}_t^{Tn} \ldots, \mathrm{Y}_{M_s}^{Tn}]$, where $\mathbf{V}_{s,P}^{Tn} \in \mathbf{R}^{(N_f\times(2T+1))\times M_s}$. As for the noise part, an extended noise spectrogram, $\mathbf{V}_{n,P}^{Tn} \in \mathbf{R}^{(N_f\times T)\times M_n}$, which consists of the spectral patches from the original noise spectrogram $\mathbf{V}_n^{Tn}$ is prepared similarly to $\mathbf{V}_{s,P}^{Tn}$. Analogous to the procedures mentioned in sections 2.2 and 3.1, NMF is subsequently performed on $\mathbf{V}_{s,P}^{Tn}$ and $\mathbf{V}_{s,P}^{Tn}$ in the training phase to obtain

$$\mathbf{V}_{s,P}^{Tn} \approx \mathbf{W}_{s,P}^{Tn}\mathbf{H}_{s,P}^{Tn}, \tag{13}$$

$$\mathbf{V}_{n,P}^{Tn} \approx \mathbf{W}_{n,P}^{Tn}\mathbf{H}_{n,P}^{Tn}. \tag{14}$$

In the enhancement phase, the extended spectrogram for each noise-corrupted utterance, denoted by $\mathbf{V}_{mix,P}^{Tt}$, is approximated via NMF with the fixed basis matrix $[\mathbf{W}_{s,P}^{Tn}\ \mathbf{W}_{n,P}^{Tn}]$:

$$\mathbf{V}_{mix,P}^{Tt} \approx \begin{bmatrix}\mathbf{W}_{s,P}^{Tn} & \mathbf{W}_{n,P}^{Tn}\end{bmatrix}\begin{bmatrix}\mathbf{H}_{s,P}^{Tt}\\ \mathbf{H}_{n,P}^{Tt}\end{bmatrix}, \tag{15}$$

The speech part of the right-hand side in Eq. (15) is then extracted:

$$\mathbf{V}_{s,P}^{Tt} \approx \mathbf{W}_{s,P}^{Tn}\mathbf{H}_{s,P}^{Tt}. \tag{16}$$

Denoting $\mathbf{V}_{s,P}^{Tt} = [\mathrm{Y}_1^{Tt}, .. \mathrm{Y}_t^{Tt} \ldots, \mathrm{Y}_{M_s}^{Tt}]$, where $V_t^{Tt} \in \mathbf{R}^{(N_f\times(2T+1))\times 1}$ is the updated spectral patch at the frame time instant t, we construct the enhanced magnitude spectrum for the t^{th} frame by averaging the $(2T+1)$ sub-vectors in V_t^{Tt}:

$$U_t^{Tt} = \frac{1}{2T+1}\left(\left[V_t^{Tt}\right]_1^{N_f} + \left[V_t^{Tt}\right]_{N_f+1}^{2N_f} + \cdots + \left[V_t^{Tt}\right]_{2TN_f+1}^{(2T+1)N_f}\right), \tag{17}$$

where $[V_t^{Tt}]_a^b$ is the sub-vector containing the a^{th} to b^{th} entries of V_t^{Tt}. Therefore, $\mathbf{U}_s^{Tt} = [U_1^{Tt}, U_2^{Tt}, .., U_{M_z}^{Tt}]$ is the finally enhanced magnitude spectrogram. This NMF-based method with patch processing is termed temporal-domain SNMF and denoted by SNMF (T) as a compact notation in the discussions hereafter.

4. EXPERIMENTS

This section describes the experimental setups used to evaluate the proposed approaches. In addition to the conventional NMF and new presented SNMF schemes, a well-known speech enhancement method, MMSE log-spectral amplitude (LSA) estimation, was implemented for comparison. The conventional NMF enhancement algorithm is denoted as NMF for simplicity. Besides SNMF (S) and SNMF (T) used in isolation, we further developed an integrated scheme where the enhanced spectrogram is the average of the output spectrograms of the SNMF (S) and SNMF (T) processes, and this scheme is denoted by SNMF (ST) in the following experimental results and discussions.

4.1. Experimental setup

This section presents the database, configurations of the speech enhancement systems, and the evaluation metric.

4.1.1 Speech data preparation

The experiments used the utterances included in the Aurora-2 database [25], which contained connected English digit utterances generated by both female and male speakers at a sampling rate of 8 kHz. Parts of these utterances were contaminated by various types of noise at different SNRs. In the experiments, 39 noise-free clean utterances produced by a target female speaker were used for generating the NMF basis matrix $\mathbf{W}_s^{Tn}$ in Eq. (3). A segment of speech-free airport noise was used to prepare $\mathbf{W}_n^{Tn}$ in Eq. (4). 50 airport-noise corrupted utterances belonging to the same speaker were used to form the testing set. The SNR levels of the noise-corrupted utterances were varied from 0 dB to 20 dB.

4.1.2 Speech enhancement setup

Some information about the setup of NMF-based speech enhancement used in this study is as follows:

1) Each utterance was split into overlapped frames. The frame duration and frame shift were set to 20 ms and 10 ms, respectively. A Hamming window was then applied to each frame signal.

2) The number of frequency bins, N_f, for the short-time Fourier transform (STFT) was set to 257.

3) The ranks (the numbers of columns) of both the NMF basis matrix $\mathbf{W}_s^{Tn}$ in Eq. (3) and $\mathbf{W}_n^{Tn}$ in Eq. (4) were assigned to 20. The maximum number of iterations in the NMF process was 100.

4) For sub-band processing, the high and low frequency bands are divided at 2,000 Hz.

5) For patch processing, the number of frames contained in a spectral patch, $2T + 1$, is set to 3.

4.1.3 Objective evaluation metric

Perceptual estimation of speech quality (PESQ) [26]-[28] was used as the evaluation metric. PESQ indicates the quality difference between the enhanced and clean speech signals, and it is analogous to the mean opinion score, which is a subjective evaluation index. The PESQ score ranges from 0.5 to 4.5, and a high score indicates that the enhanced utterance is close to the clean utterance.

4.2. Experimental results

This section presents the experimental results and discussions. The spectrogram analyses among the speech enhancement algorithms are presented, and the PESQ results for each algorithm are reported.

4.2.1 Spectrogram analysis

The spectrograms processed using LSA, NMF and SNMF (ST) methods are shown and compared in Fig. 2. Because the LSA method compensates for noise components in a frame-wise manner (which is different from the NMF-related technique), it was expected to provide complementary effect to the NMF-based method. Therefore, the LSA and SNMF (ST) were integrated, and the resultant scheme is denoted as LSA+SNMF (ST). In Fig. 2, the content of the utterance is "FAK_3615A.08", which was acquired from the Aurora-2 database and produced by the target speaker.

The six panels in Fig. 2 indicate that Ⓒ LSA reduced noise markedly, whereas the NMF-based approaches, including Ⓓ NMF and Ⓕ SNMF (ST), maintained a favorable speech signal structure. Ⓔ SNMF (ST) exhibited greater capability to reduce noise components than Ⓓ NMF did. Finally, Ⓕ LSA+SNMF (ST) gave an improvement to Ⓔ SNMF (ST).

4.2.2 PESQ results

Table I shows the PESQ results of the various instantiations of SNMF scheme including SNMF (S), SNMF (T) and SNMF (ST), respectively. The PESQ results of noisy signals and the enhanced signals via the conventional NMF noted in subsection 2.2 are also listed and denoted as Noisy and NMF, respectively, in this table.

Table 1 indicates that both SNMF (S) and SNMF (T) outperform NMF in most SNR cases, confirming the effectiveness of incorporating either of the spectral and temporal information into the conventional NMF-based speech enhancement algorithm. Next, SNMF (ST) provides the highest PESQ results among the four approaches, indicating that the integration of spectral and temporal information by pairing SNMF (S) with SNMF (T) enables SNMF to achieve better performance than each component approach.

Next, we compare LSA and its combination with NMF and SNMF. In Table II, we list the results of LSA (Ⓒ in Fig. 2), LSA+NMF, LSA+SNMF (S), and LSA+SNMF (ST) (Ⓕ in Fig. 2). Comparing Tables I and II shows that LSA+NMF outperforms LSA in most conditions, while LSA+SNMF (S) gives even better performance than LSA+NMF. The results indicate that the unsupervised LSA algorithm can be suitably integrated with supervised NMF- and SNMF-based methods to reduce the noise effect further. Moreover, it is noted that LSA+SNMF (ST) performs the best among the four approaches in Table II, confirming that the frame-wise LSA is well additive to the jointly sub-band (spectral) and patch (temporal) NMF scheme to provide superior noise reduction.

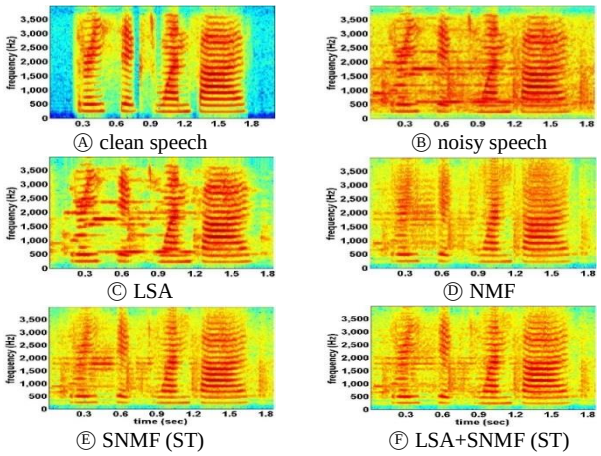

Fig 2. Spectrograms of: Ⓐ clean speech, Ⓑ noisy speech, the enhanced speech via ⒸMMSE-LSA, Ⓓ NMF, Ⓔ SNMF (ST), and Ⓕ LSA+SNMF (ST).

Table I. PESQ for various enhancement methods in different SNRs

SNR	Noisy	NMF	SNMF (S)	SNMF (T)	SNMF (ST)
0 dB	1.30	1.67	**1.70**	1.67	1.70
5 dB	1.77	2.07	2.10	2.10	**2.12**
10 dB	2.06	2.33	2.38	2.34	**2.39**
15 dB	2.39	2.53	2.60	2.56	**2.63**
20 dB	2.78	2.79	2.89	2.79	**2.91**

Table II. PESQ for various combinative methods in different SNRs

SNR	LSA	LSA+NMF	LSA+SNMF (S)	LSA+SNMF (ST)
0 dB	1.68	1.96	1.94	**1.98**
5 dB	2.23	2.36	2.42	**2.43**
10 dB	2.48	2.58	2.62	**2.65**
15 dB	2.80	2.81	2.88	**2.89**
20 dB	3.04	2.91	3.04	**3.05**

5. CONCLUSION

This study proposed a segmental NMF-based speech enhancement scheme to improve the conventional frame-wise NMF-based method. The spectral and temporal structures of speech signals were considered to derive sub-band and patch processing approaches, enabling the nonnegative magnitude spectrogram matrix to be decomposed into sub-matrices. The enhanced spectrograms derived by these sub-matrices characterized speech signals more precisely than that obtained via the conventional NMF. The experimental results demonstrated both sub-band and patch processing approaches (SNMF (S) and SNMF (T)) outperform the NMF-based method as they operate in isolation, and these two approaches are well additive to each other, making the resulting method SNMF (ST) provide further improvement. Moreover, integrating the well-known LSA with any of the SNMF instantiations consistently produce better noise reduction than the individual component method. In the future, we will conduct additional sets of objective evaluations and subjective listening tests to further examine the presented SNMF.

6. REFERENCE

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2), pp. 113–120, 1979.

[2] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 208-211, 1979.

[3] S. Kamath, P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. IV-4164, 2002.

[4] C. Plapous, C. Marro, P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*. 14(6), pp. 2098–2108, 2006.

[5] P. Scalart, J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 629-632, 1996.

[6] V. Grancharov and J. S. B. Kleijn, "On causal algorithms for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3), pp. 764–773, 2006.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Audio, Speech, and Language Processing*, 32(6), pp. 1109–1121, 1984.

[8] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions on Audio, Speech, and Language Processing*, 13(5), pp. 845–856, 2005.

[9] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Transactions on Audio, Speech, and Language Processing*, 15(6), pp. 1741–1752, 2007.

[10] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Transactions on Audio, Speech, and Language Processing*, 20(7), pp. 1948–1963, 2012.

[11] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, 9(5), pp. 504–512, 2001.

[12] I. Cohen, "Noise spectrum estimation in adverse environments : Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, 11(5), pp. 466–475, 2003.

[13] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1), pp. 163–176, 2006.

[14] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3), pp. 882–892, 2007.

[15] K. El-Maleh, A. Samouelian, and P. Kabal, "Frame level noise classification in mobile environments," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 237–240, 1999.

[16] N. Mohammadiha, J. Taghia, and A. Leijon, "Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.4561-4564, 2012.

[17] N. Mohammadiha, T. Gerkmann and A. Leijon, "A new approach for speech enhancement based on a constrained nonnegative matrix factorization," in *Proceedings of the IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 1-5, 2011.

[18] K. W. Wilson, B. Raj, P. Smaragdisy and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4029-4032, 2008.

[19] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1), pp. 1–12, 2007.

[20] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3), pp. 1066-1074, 2007.

[21] N. Mohammadiha and A. Leijon, "Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement," *IEEE Transactions Audio, on Speech, and Language Processing*, 21(5), pp. 998-1011, 2013.

[22] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized nonnegative matrix factorization with temporal dependencies for speech denoising," in *Proceedings of International Conference Spoken Language Processing*, pp. 411–414, 2008.

[23] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 17–20, 2011.

[24] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization," in *Proceedings of IEEE Workshop Applications of Signal Processing to Audio Acoustics*, pp. 45–48, 2011.

[25] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of the 2000 Automatic Speech Recognition: Challenges for the new Millenium*, pp. 181-188, 2000.

[26] ITU-T Recommendation P.862, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, 2001.

[27] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 749-752, 2001.

[28] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1), pp. 229-238, 2008.