Speech Enhancement via Ensemble Modeling NMF Adaptation

Jeremy Chiaming Yang[1], Syu-Siang Wang[1], Yu Tsao[1], and Jeih-weih Hung[2]

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

***Abstract*--Nonnegative matrix factorization (NMF)-based speech enhancement algorithm has been proven to provide satisfactory performance when the prior information about speaker and noise types are given. In most real-world scenarios, however, such prior information may not always be accessible. Therefore, an adaptation technique is favorable to adapt the NMF matrices to match the testing condition for a better enhancement performance. In this study, we proposed a novel ensemble modeling (ESM) algorithm for NMF adaptation. The algorithm effectively uses the local information of the entire training data to facilitate an effective and efficient online NMF adaptation. The adapted NMF matrices were used to perform speech enhancement. Experimental results on the perceptual evaluation of speech quality (PESQ) confirmed the effectiveness of the ESM algorithm in various signal-to-noise ratio conditions.**

I. INTRODUCTION

Speech enhancement is a key component in many speech applications to handle the noise issues in real-world conditions. Numerous speech enhancement approaches have been proposed. Among them, nonnegative matrix factorization (NMF) is a successful one that has been extensively studied [1, 2]. Previous studies have confirmed that the NMF-based speech enhancement can achieve satisfactory performance when the clean speech data (used to train the clean spectral dictionary matrix) and testing utterances are from the same speaker. However, in most real-world scenarios, they often come from different speakers. The mismatch can notably degrade the enhancement performance.

In this study, we propose a novel ensemble modeling (ESM) algorithm for NMF adaptation to handle the abovementioned mismatch issue. The proposed ESM algorithm can be divided into two stages, offline and online stages. In the offline stage, ESM prepares multiple NMF basis matrices, each matrix characterizing local information of the entire set of training data. In the online stage, ESM estimates a new NMF basis matrix that matches the testing condition with a mapping function. The parameters in the mapping function are estimated by a set of adaptation data, which has similar acoustic characteristics to the testing condition. The resulting adapted NMF basis matrix is then used to perform NMF speech enhancement upon the testing utterances. Experimental results confirmed that this ESM-wise NMF adaptation achieves better performance than conventional NMF-based method in the speech quality (PESQ) evaluation [3] across various signal-to-noise ratio (SNR) conditions within a limited number of adaptation speech samples. For clarity, the symbols used in this paper are summarized in Table 1.

Symbol	Description
$\mathbf{V}_s^{Tn}$	Clean-speech magnitude spectrogram
$\mathbf{V}_n^{Tn}$	Speech-free noise magnitude spectrogram
$\mathbf{V}_{mix}^{Tt}$	Testing noisy magnitude spectrogram
$\mathbf{V}_s^{Tt}$	Enhanced speech magnitude spectrogram
$\mathbf{W}_s^{Tn}$	Spectral dictionary matrix of clean speech
$\mathbf{W}_n^{Tn}$	Spectral dictionary matrix of noise
$\mathbf{W}_{s,Adp}^{Tn}$	Adapted spectral dictionary matrix of clean speech
$\mathbf{W}_{s,C_i}^{Tn}$	Spectral dictionary matrix for the C_i-th cluster
$\mathbf{H}_s^{Tn}$	Coefficient matrix of clean speech
$\mathbf{H}_n^{Tn}$	Coefficient matrix of noise
$\mathbf{H}_s^{Tt}$	Coefficient matrix of enhanced speech

Table 1. Summary of the symbols in this paper.

II. NMF-BASED SPEECH ENHANCEMENT

NMF-based speech enhancement comprises two phases, training and enhancement [2]. In the training phase, $\mathbf{V}_s^{Tn}$ can be approximated as:

$$\mathbf{V}_s^{Tn} \approx \mathbf{W}_s^{Tn}\mathbf{H}_s^{Tn} \quad (1)$$

In the same way, $\mathbf{V}_n^{Tn}$ can be also decomposed as:

$$\mathbf{V}_n^{Tn} \approx \mathbf{W}_n^{Tn}\mathbf{H}_n^{Tn} \quad (2)$$

The four matrices, $\mathbf{W}_s^{Tn}$, $\mathbf{V}_s^{Tn}$, $\mathbf{W}_n^{Tn}$ and $\mathbf{V}_n^{Tn}$ are updated by minimizing $\|\mathbf{V}_s^{Tn} - \mathbf{W}_s^{Tn}\mathbf{H}_s^{Tn}\|^2$ and $\|\mathbf{V}_n^{Tn} - \mathbf{W}_n^{Tn}\mathbf{H}_n^{Tn}\|^2$. In the enhancement phase, $\mathbf{V}_{mix}^{Tt}$ is approximated by Eq. (3):

$$\mathbf{V}_{mix}^{Tt} \approx \mathbf{W}_c^{Tn}\mathbf{H}_c^{Tn} = [\mathbf{W}_s^{Tn}\ \mathbf{W}_n^{Tn}]\begin{bmatrix}\mathbf{H}_s^{Tt}\\ \mathbf{H}_n^{Tt}\end{bmatrix}. \quad (3)$$

Finally, the matrix of clean speech predicts the enhanced speech magnitude spectrogram for the testing utterances:

$$\mathbf{V}_s^{Tt} \approx \mathbf{W}_s^{Tn}\mathbf{H}_s^{Tt}. \quad (4)$$

Then the enhanced magnitude spectrogram $\mathbf{V}_s^{Tt}$ in Eq. (4) with the original phase spectrogram is transformed into the enhanced testing utterance. More details about the NMF-based speech enhancement can be found in [1, 2].

III. ENSEMBLE MODELING NMF ADAPTATION

In this study, we proposed an ensemble modeling (ESM) algorithm to adapt the speech bases in $\mathbf{W}_s^{Tn}$ (in Eq. (1)) to $\mathbf{W}_{s,Adp}^{Tn}$, which is then used to replace $\mathbf{W}_s^{Tn}$ (in Eq. (3)) for speech enhancement. The system architecture of the ESM algorithm is presented in Fig.1. The overall ESM algorithm is divided into two stages, offline and online. From Fig.1, the offline stage divides the entire set of training data into several clusters (C_N clusters in Fig. 1), with each cluster including training data with similar acoustic characteristics. Next, each cluster of training data is used to compute a speech basis matrix. Thus we have C_N matrices: $\mathbf{W}_{s,C_1}^{Tn}$, $\mathbf{W}_{s,C_2}^{Tn}$,…, $\mathbf{W}_{s,C_N}^{Tn}$.

When directly estimating these C_N matrices, the order of basis vector among the matrices may not align. Thus, we compute the basis matrices via the following mapping function [4]:

$$\mathbf{W}_{s,C_i}^{Tn} = \mathbf{A}_i \mathbf{W}_{s,SI}^{Tn} + \mathbf{B}_i \quad (5)$$

, where $\mathbf{W}_{s,SI}^{Tn}$ is the basis matrix computed by the entire set of training data. $\{\mathbf{A}_i, \mathbf{B}_i\}$ denotes the parameters of the mapping function that are estimated by the C_i-th cluster of training data. The online stage computes another basis matrix $\mathbf{W}_{s,Adp}^{Tn}$ by

$$\mathbf{W}_{s,Adp}^{Tn} = \mathbf{G}(\mathbf{W}_{s,C_1}^{Tn}, \mathbf{W}_{s,C_2}^{Tn}, \ldots, \mathbf{W}_{s,C_N}^{Tn}) \quad (6)$$

, where $\mathbf{G}(.)$ is a mapping function. The parameters in $\mathbf{G}(.)$ are computed based on the adaptation data. There are various selections for the mapping function $\mathbf{G}(.)$ [5]. In this study, we test the linear combination with bias (LCB) mapping function, and thus Eq. (6) becomes:

$$\mathbf{W}_{s,Adp}^{Tn} = \boldsymbol{a_1}\mathbf{W}_{s,C_1}^{Tn} + \boldsymbol{a_2}\mathbf{W}_{s,C_2}^{Tn} + \ldots + \boldsymbol{a_N}\mathbf{W}_{s,C_N}^{Tn} + \mathbf{B} \quad (7)$$

, where $\{\boldsymbol{a_1}, \boldsymbol{a_2}$ ..., $\boldsymbol{a_N}\}$ and $\mathbf{B}$ denote the coefficients and the bias matrix, respectively.

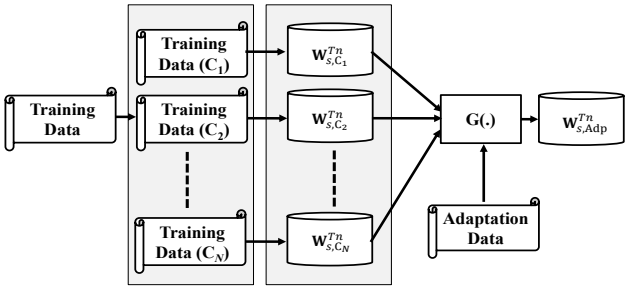

Fig 1. The proposed ESM algorithm for NMF adaptation.

IV. EXPERIMENTS

A. Setting

This section elaborates the data preparation, configurations of the speech enhancement system, and the evaluation metric.

1) Speech data preparation

The utterances were extracted from the Aurora-2 database [6]. In the experiments, 390 noise-free clean utterances produced by 10 speakers were used to generate the spectral dictionary matrix $\mathbf{W}_s^{Tn}$ in Eq. (1). Speech-free noises were used to produce $\mathbf{W}_n^{Tn}$ in Eq. (2). We mixed 10 utterances belonging to the same female speaker with babble noise to form the evaluation set. The SNR levels of the noise-mixed utterances differed from 0 dB to 20 dB.

2) Speech enhancement setup

The experimental setup is presented as follows:

① Every utterance was divided into overlapped frames, with 20 ms and 10 ms, respectively, for frame duration and shift.

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

③ The ranks of both the NMF basis matrices $\mathbf{W}_s^{Tn}$ and $\mathbf{W}_n^{Tn}$ were assigned to 20. The maximum number of iterations in the NMF updating process was 100.

④ The ESM algorithm is only applied to speech NMF basis matrices, and the noise NMF basis matrix is not adapted.

3) Objective evaluation metric

PESQ [3] was used as the evaluation metric. The PESQ score ranges from 0.5 to 4.5, and a high score reveals that the enhanced utterance is close to the clean utterance.

B. Experimental results

Table I shows the PESQ results of Noisy (where no speech enhancement is performed) and speaker-independent (SI) NMF (conventional NMF-based speech enhancement). From Table I, we can note that NMF can improve the PESQ scores moderately, despite the fact that the used basis matrix is not matched to the testing utterance.

Table I. PESQ scores for Noisy and NMF in different SNRs

SNR	0 dB	5 dB	10 dB	15 dB	20 dB
Noisy	1.3450	1.6559	1.9589	2.3103	2.6239
NMF	1.7043	2.0488	2.3937	2.6839	2.8742

Table II shows the PESQ scores of speech enhancement using the ESM NMF adaptation with different number of adaptation data frames. Notably, 10 frames amount to a 0.3-second speech data. By comparing Tables I and II, we observe that ESM NMF adaptation can effectively improve the PESQ scores over the conventional NMF counterpart, even when very limited adaptation data is used (10 frames). We also noticed that the performance saturated quickly (for example at 0dB SNR, using 30 adaptation frames can already achieves the optimal PESQ scores). The results confirm that the ESM NMF adaptation can provide rapid adaptation capability.

Table II. PESQ scores for of ESM in different SNRs

SNR/ #adapt frame	0 dB	5 dB	10 dB	15 dB	20 dB
10	1.8802	2.161	2.4816	2.7529	2.9573
20	1.8996	2.1773	2.5031	2.7876	3.0012
30	1.9345	2.2093	2.5306	2.8112	3.0129
40	1.9345	2.2093	2.5306	2.8112	3.0129
50	1.9345	2.2093	2.5306	2.8112	3.0129

V. CONCLUSION

This study proposed an ESM algorithm for NMF adaptation. By using ESM to adapt speech basis matrix, the performance of NMF-based speech enhancement has been notably improved. In the future, we will investigate different criteria for offline preparation and online mapping function estimation for the ESM algorithm.

REFERENCE

[1] K. W. Wilson, B. Raj, P. Smaragdisy, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. ICASSP 2008*, pp. 4029-4032, 2008.

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

[3] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on ASLP*, vol. 16, pp. 229-238, 2008.

[4] E. M. Grais and H. Erdogan, "Adaptation of speaker-specific bases in non-negative matrix factorization for single channel speech-music separation," in *Proc. INTERSPEECH 2011, pp. 569-572, 2011.*

[5] Y. Tsao, P. Lin, T.-y. Hu, and X. Lu, "Ensemble environment modeling using affine transform group," *Speech Communication*, vol. 68, pp. 55-68, 2016.

[6] D. Pearce and H.-G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *in Proc. ITRW ASR*, 2000.