Classifying NMF Components Based on Vector Similarity for Speech and Music Separation

Nengheng Zheng*, Yi Cai*, Xia Li* and Tan Lee†
*Shenzhen Key Lab of Telecommunication and Information Processing, College of Information Engineering, Shenzhen University, Shenzhen, China. E-mail: nhzheng@szu.edu.cn
†Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, China
E-mail: tanlee@ee.cuhk.edu.hk

***Abstract*—This paper presents a nonnegative matrix factorization (NMF) components classification algorithm for single-channel speech and music separation. Music only and music-speech mixture segments are firstly classified from the audio stream via audio segmentation technique. Then NMF is applied for signal decomposition. The basis matrix of the NMF output of music only segments provides the prior knowledge of music component in the mixture signal. NMF components, i.e. basis and gain vectors of the mixture signal are classified into speech and music based on the vector similarity between each basis vector and the priori music basis matrix. A set of SNR-dependent thresholding coefficients are empirically determined for the classification. The separated speech and music signals are reconstructed from the respectively classified NMF components. Experimental results show the effectiveness of the proposed method for speech and music separation, and its superior performance over the traditional NMF-based separation methods.**

I. INTRODUCTION

Speech separation refers to the problem of separating speech signal and interfering signal from their mixture. It is a fundamental problem of signal processing that has been studied for many years. Blind source separation (BSS) approaches such as independent component analysis (ICA) have demonstrated good success when multiple input mixtures are available [1]. However, in many practical applications, only a single-channel signal recording is available. In this case, source separation is nearly an irresolvable problem without priori knowledge of the component sources, and conventional BSS methods are generally not applicable.

Many attempts to single-channel speech separation have been reported, with different kinds of priori knowledge being exploited. In computational auditory scene analysis (CASA) approaches [2], the input signal is segregated into coherent time-frequency regions, which are then classified as individual sources based on some primitive acoustic cues such as F0s, onset/offset timing, temporal continuity, etc. High-level knowledge about the speech was found useful to source separation. Sameti [3] and Xiao [4] exploited the priors embedded in the acoustic models of speech and noise for speech enhancement. Ji [5] proposed a corpus-based approach, in which the target speech was reconstructed from clean speech segments selected from a large speech corpus. By matching long speech segments, the temporal dynamics and speaker characteristics of the target speech were obtained.

As a recently proposed signal matrix decomposition technique, nonnegative matrix factorization (NMF) [6] has been widely implemented for analyzing and classifying various kinds of signals, including image, audio and biomedical signals, etc. In particular, several NMF-based speech separation techniques have been proposed in the past ten years, e.g. [7-12]. In these works, NMF was adopted for signal decomposition and various priors of the sources were deployed to classify the NMF output into speech and interfering source. For example, Grais [9] presented an NMF-based speech and music separation method. The basis vectors of each source were generated from the training data and served as the representative bases for the respective sources in the mixture. Only gain matrices were updated during NMF of the mixture and each source was reconstructed as a linear combination of their pre-trained basis vectors. The major problem is that the pre-trained basis vector can not represent the processing sources well due to the mismatches between the processing signals and the training data, which degraded the separation accuracy drastically. Zheng [10] proposed an algorithm that utilizes linguistic knowledge to separate two mixed speech sources with equal intensity. Linguistic contents and acoustic models were used to predict the initial basis and gain matrices, and syllable level NMF was implemented with several factorization constraints. However, the linguistic knowledge is usually unavailable in practical application. In [12], basis matrix of music was trained from music only segments among the audio stream, and NMF was implemented for speech and music separation with the pre-trained music basis and several factorization constraints. However, the random initialization of speech matrix may lead to undesired results and the constraints imposed to direct the factorization may not always reflect the spectral and temporal properties of the mixing sources. More explicit priori

This research is jointly supported by the National Science Foundation of China (Ref: 60901061), Guangdong Science and Technology Plan (Ref: 2011B010200045), the General Research Funds (Ref: CUHK 414108) from the Hong Kong Research Grant Council, and a project grant from the Shun Hing Institute of Advanced Engineering, CUHK.

knowledge should be incorporated to supervise the components classification.

This paper deals with single-channel speech and music separation task. An NMF-based separation algorithm is proposed. Priors of the music are online-learned. A vector similarity measure, together with a set of SNR-dependent classification thresholds, is implemented to classify the basis and gain vectors of the NMF output of the mixture signal. A series of separation experiments are carried out to demonstrate the effectiveness of the proposed method.

II. NMF FOR SIGNAL ANALYSIS AND BLIND SOURCE SEPARATION

A. NMF for signal representation

NMF is a factorization method that approximates an observation matrix $\mathbf{Y}$ by the product of a basis matrix $\mathbf{B}$ and a gain matrix $\mathbf{G}$, i.e.

$$\mathbf{Y} = \mathbf{BG} + \mathbf{V}, \tag{1}$$

where $\mathbf{Y} \in R^{N\times T}$, $\mathbf{B} \in R^{N\times J}$ and $\mathbf{G} \in R^{J\times T}$. Approximant of $\mathbf{Y}$ is given by $\hat{\mathbf{Y}} = \mathbf{BG}$, and $\mathbf{V}$ is the approximation error. Elements in $\mathbf{Y}$, $\mathbf{B}$ and $\mathbf{G}$ are all nonnegative [6]. Imposing appropriate factorization constraints, e.g. the least Euclidian distance, least divergence and sparse distribution, NMF decomposes a signal into basis elements that represent locally key parts of the signal. For example, with the Euclidian distance as the factorization cost, i.e.

$$\zeta_{Euc} = \| \mathbf{Y} - \mathbf{BG} \|^2. \tag{2}$$

The factorization can be achieved by minimizing ζ_{Euc} via a gradient descent iterating process. In each iteration, the new basis and gain matrices, $\hat{\mathbf{B}}$ and $\hat{\mathbf{G}}$, are element-wise updated as

$$\begin{aligned} \hat{\mathbf{B}}:\ [\mathbf{B}]_{n,j} &\leftarrow [\mathbf{B}]_{n,j} - \beta^B_{n,j} \frac{\partial J_{Euc}}{\partial [\mathbf{B}]_{n,j}} \quad n = 1,2,\cdots,N \\ & \qquad\qquad\qquad\quad j = 1,2,\cdots,J \\ \hat{\mathbf{G}}:\ [\mathbf{G}]_{j,t} &\leftarrow [\mathbf{G}]_{j,t} - \beta^G_{j,t} \frac{\partial J_{Euc}}{\partial [\mathbf{G}]_{j,t}} \quad t = 1,2,\cdots,T \end{aligned}, \tag{3}$$

where $\beta^B_{n,j} = \frac{[\mathbf{B}]_{n,j}}{[\mathbf{BGG}^{\mathrm{T}}]_{n,j}}$, $\beta^G_{j,t} = \frac{[\mathbf{G}]_{j,t}}{[\mathbf{B}^{\mathrm{T}}\mathbf{BG}]_{j,t}}$ (4)

are the descending rate for $\mathbf{B}$ and $\mathbf{G}$, respectively.

B. NMF-based blind source separation

Given $\mathbf{Y}$ being a mixture of speech and music, (1) can be rewritten as

$$\mathbf{Y} = \mathbf{B}_s\mathbf{G}_s + \mathbf{B}_m\mathbf{G}_m + \mathbf{V} = [\mathbf{B}_s\ \mathbf{B}_m]\begin{bmatrix}\mathbf{G}_s \\ \mathbf{G}_m\end{bmatrix} + \mathbf{V}, \tag{5}$$

where subscripts *s* and *m* denote speech and music components, respectively. For blind separation task without any priori knowledge of the two sources, basis and gain matrices are usually initialized with random elements. According to the time-frequency properties of the two acoustical signals, a number of factorization constraints have been developed for the separation task. Some of them are listed following.

- The sparseness constraint [13]. For example, the constraint for music can be defined as

$$\zeta_{Spa}(\mathbf{G}_m) = \sum_{j=1}^{J}\sum_{t=1}^{T} |[\mathbf{G}_m]_{j,t} / \sigma_j| \tag{6}$$

where $\sigma_j = \sqrt{\frac{1}{T}\sum_{t=1}^{T}[\mathbf{G}_m]^2_{j,t}}$. (7)

For speech and music spectra, the major energy is occupied by the harmonic frequency components. The sparseness constraint is very successful in NMF processing of the two signals.

- The temporal continuity constraint [7]. For example,

$$\zeta_{Tem}(\mathbf{G}_m) = \sum_{j=1}^{J}\frac{1}{\sigma_j^2}\sum_{t=1}^{T}([\mathbf{G}_m]_{j,t} - [\mathbf{G}_m]_{j,t-1})^2 \tag{8}$$

denotes the temporal continuity constraint for music. The constraint for speech can be defined similarly. The temporal continuous properties of the frequency bins can be easily found from the spectrograms of speech and music.

- The orthogonality constraint [11]. Assuming speech and music being two independent random processes, we can define the orthogonality constraint as

$$\zeta_{Oth} = \sum_{n}\sum_{t}[\mathbf{B}_s\mathbf{G}_s]_{n,t} \cdot [\mathbf{B}_m\mathbf{G}_m]_{n,t}. \tag{9}$$

In speech and music separation tasks, usually, better performances can be achieved using a combination of the abovementioned constraints.

III. NMF COMPONENTS CLASSIFICATION FOR SPEECH-MUSIC SEPARATION

The blind separation based on NMF usually does not result in good separated sources due to the two problems:

- In typical NMF, $\mathbf{B}$ and $\mathbf{G}$ are initialized randomly. With randomized initialization, NMF may not converge to an optimal factorization and may lead to undesired results.
- The constraints imposed to direct the factorization may not accurately reflect the spectral and temporal properties of the mixing sources. More explicit prior knowledge should be incorporated to supervise the components classification.

In this study, we present an NMF components classification method to separate speech and music sources from their mixture. Priors of the processing music sources, i.e. the musical basis matrix are online obtained and vector similarity between each mixture component and the priori matrix is computed to supervise the classification. Figure 1 shows the block diagram of the proposed separation system. Details of the key processing blocks are described following.

A. Audio segmentation

It is reasonable to assume that the audio stream contains music-only segments (e.g., before and after a sentence, or where long speech pauses happen). The audio stream is

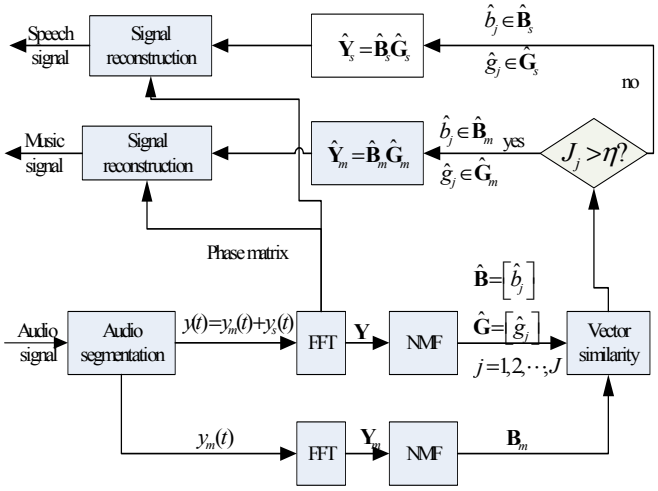

Fig. 1 Block diagram for the proposed speech and music separation system

classified into music-only segments, noted as $y_m(t)$, and music-speech mixture segments, noted as $y(t)$, via a state-of-the-art audio segmentation method.

B. Learning the musical priors via NMF

For music only segments, an N by T nonnegative matrix $\mathbf{Y}_m$ is constructed from the magnitude of the short-time Fourier spectra of $y_m(t)$, where N is the number of frequency bins and T is the number of short-time frames. NMF processing as given in (3) and (4) is implemented to generate the basis and gain matrices of music, i.e., $\mathbf{B}_m$ and $\mathbf{G}_m$. Instead of using only ζ_{Euc}, the optimizing cost function adopted here is a weighted combination of three constraints, i.e.

$$\zeta=\zeta_{Euc}+w_t\zeta_{Tem}(\mathbf{G}_m)+w_s\zeta_{Spa}(\mathbf{G}_m), \qquad (10)$$

where w_t and w_s are weighting parameters of the respective constraints. The matrix $\mathbf{B}_m$ provides some prior knowledge of the musical components in the mixture assuming that music segments from the same song have a certain degree of consistency in musical properties.

C. Speech-music classification of NMF components

For each segment of mixture signal $y(t)$, a matrix $\mathbf{Y}$ is constructed similarly to $\mathbf{Y}_m$. NMF is applied to $\mathbf{Y}$ to generated basis and gain matrices

$$\hat{\mathbf{B}}=[\hat{\boldsymbol{b}}_1\cdots\hat{\boldsymbol{b}}_j\cdots\hat{\boldsymbol{b}}_J];\quad \hat{\mathbf{G}}=\begin{bmatrix}\hat{\boldsymbol{g}}_1\\ \vdots\\ \hat{\boldsymbol{g}}_j\\ \vdots\\ \hat{\boldsymbol{g}}_J\end{bmatrix}, \qquad (11)$$

with the factorization constraint

$$\zeta=\zeta_{Euc}+w_t\zeta_{Tem}(\hat{\mathbf{G}})+w_s\zeta_{Spa}(\hat{\mathbf{G}}). \qquad (12)$$

In (11), each pair of column vector $\hat{\boldsymbol{b}}_j$ and row vector $\hat{\boldsymbol{g}}_j$ could belong to either music or speech. For each $\hat{\boldsymbol{b}}_j$, its similarity to $\mathbf{B}_m$ is computed as

$$\gamma_j=\cos\angle(\hat{\boldsymbol{b}}_j,\mathbf{B}_m)=\sum_{i=1}^{I}\frac{[\hat{\boldsymbol{b}}_j\bullet\boldsymbol{b}_i^m]}{\|\hat{\boldsymbol{b}}_j\|\cdot\|\boldsymbol{b}_i^m\|}, \qquad (13)$$

in which $\boldsymbol{b}_i^m$ are the column vectors in $\mathbf{B}_m$. Now the speech and music components classification is given by

$$\begin{cases}\hat{\boldsymbol{b}}_j\in\hat{\mathbf{B}}_m,\hat{\boldsymbol{g}}_j\in\hat{\mathbf{G}}_m, & if\ \gamma_j>\eta\\ \hat{\boldsymbol{b}}_j\in\hat{\mathbf{B}}_s,\hat{\boldsymbol{g}}_j\in\hat{\mathbf{G}}_s, & if\ \gamma_j\le\eta\end{cases}, \qquad (14)$$

where η is the classifying threshold.

Note that the number of basis vectors for representing the signal, i.e. the column dimension size I and J for $\mathbf{B}_m$ and $\hat{\mathbf{B}}$, respectively, is an important parameter in NMF. As abovementioned, NMF decomposes a signal into basis vectors representing locally key parts of the whole signal. A small size of bases will blur the distinct key parts; on the other hand, a large size of bases may result in over fitting to the noise components. Therefore, selecting appropriate base sizes for different signals should be considered. In this study, analytical experiments have been carried to select the parameters. For NMF of music only signal, it is found that the number of basis vectors does not affect the classification performance significantly. As we can see in (13), the possibility of each $\hat{\boldsymbol{b}}_j$ belonging to music, γ_j, is measure as the sum of distances from $\hat{\boldsymbol{b}}_j$ to all $\boldsymbol{b}_i^m$ among $\mathbf{B}_m$. Therefore, changing I may change the values of γ_j, but will not change the rank of all γ_j. Specifically, I equals to 8 for all experiments in this study. For NMF of the mixture signal, experiments show that large number of J is preferred and J =128 is selected as the tradeoff between the accuracy and the complexity.

D. Signal reconstruction

With the classified basis and gain vectors, we can reconstruct the separated music and speech matrices as

$$\begin{aligned}&\hat{\mathbf{B}}_m=[\hat{\boldsymbol{b}}_j\,|\gamma_j>\eta];\quad \hat{\mathbf{G}}_m=[\hat{\boldsymbol{g}}_j\,|\gamma_j>\eta];\quad \hat{\mathbf{Y}}_m=\hat{\mathbf{B}}_m\hat{\mathbf{G}}_m\\ &\hat{\mathbf{B}}_s=[\hat{\boldsymbol{b}}_j\,|\gamma_j\le\eta];\quad \hat{\mathbf{G}}_s=[\hat{\boldsymbol{g}}_j\,|\gamma_j\le\eta];\quad \hat{\mathbf{Y}}_s=\hat{\mathbf{B}}_s\hat{\mathbf{G}}_s\end{aligned}. \qquad (15)$$

Finally, time domain speech and music signals can be respectively reconstructed from $\hat{\mathbf{Y}}_s$ and $\hat{\mathbf{Y}}_m$, with the phase matrix derived from the mixture signal.

IV. Determine the Classifying Threshold

The coefficient η is a key factor to the separation system. However, there is no theoretically tractable η for the task. We propose an empirical determined η via a series of analytical experiments.

It is found that for each separation trial, the similarity measure γ_j is likely to be Gaussian distributed. The optimal η can be approximated as

$$\eta=\overline{\gamma}+\alpha\overline{\gamma}, \qquad (16)$$

where $\overline{\gamma}$ denotes the mean of γ_j, and α is signal-to-noise rate (SNR) dependent and denotes the deviation of η from $\overline{\gamma}$.

To derive the SNR-dependent α, the following analytical experiments have been conducted:

- For every possible combination of speech and music signals $y_m(t)$ and $y_s(t)$, a mixture $y(t)$ with a specific SNR is generated. Basis matrix for music only signal $\mathbf{B}_m$, basis and gain matrices for mixture signal $\hat{\mathbf{B}}$, $\hat{\mathbf{G}}$, are computed. For each $\hat{\boldsymbol{b}}_j \in \hat{\mathbf{B}}$, its similarity to $\mathbf{B}_m$ is also computed as (13).
- Resort $\hat{\boldsymbol{b}}_j$ and $\hat{\boldsymbol{g}}_j$ in ascending γ_j. After resorting, $\hat{\boldsymbol{b}}_1$ will most unlikely belong to music, and most likely belong to speech; on the other hand, $\hat{\boldsymbol{b}}_J$ will most likely belong to music.
- Increase j from 1 to J. For each j, compute
$$\hat{\mathbf{Y}}_s = \hat{\mathbf{B}}(:,1:j)\hat{\mathbf{G}}(1:j,:), \quad (17)$$
and reconstruct $\hat{y}_s(t)$ from $\hat{\mathbf{Y}}_s$. Then the quality of $\hat{y}_s(t)$ in reference with the original speech $y_s(t)$, noted as $Q(j)$, is computed.
- The optimal threshold for the current separation trial is given as
$$\eta = \gamma\left(\arg\max_j Q(j)\right). \quad (18)$$
The deviation parameter is given by
$$\alpha = \frac{\eta - \overline{\gamma}}{\overline{\gamma}}. \quad (19)$$
- The above steps are repeated for different SNR levels and for each pair of $y_m(t)$ and $y_s(t)$. A statistically "optimal" SNR-dependent α can be computed as the average of all α 's for the specific SNR .

Figure 2 exemplifies, with a particular separation trial, the variation of reconstructed speech quality $Q(j)$ as γ increases. Here, we adopted two speech quality measurements, i.e. perceptual evaluation of speech quality (PESQ) [14], and SNR. As illustrated in the figure, at the beginning, both PESQ and SNR increase as γ increases. This is because that with increasing γ, more NMF components, which is very likely to be speech, are used to reconstruct the speech signal, such that the speech quality keeps an increasing trend. However, after $\gamma > \eta$, components likely to be music are misused for speech reconstruction. Consequently, the speech quality degrades. Note that the best PESQ and SNR achieve at different γ in this example. This is a common phenomenon in speech quality measurement. As higher SNR achievement usually brings extra speech distortion in speech separation, in this study, the optimal threshold η is defined as γ where highest PESQ achieves.

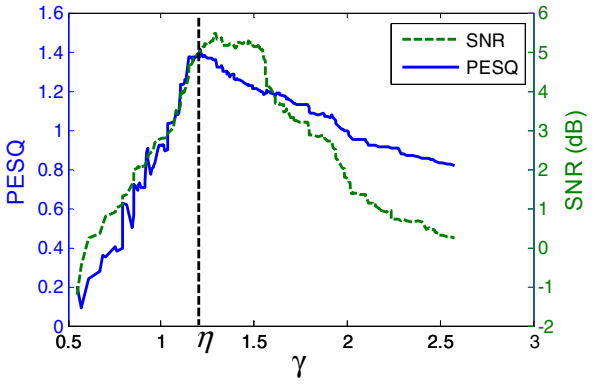

Fig. 2 Quality of reconstruct speech (PESQ & SNR) as γ increases.

Table I gives the SNR-dependent values of α. Each value is statistically concluded from experiments with 30 different pairs of $y_m(t)$ and $y_s(t)$.

TABLE I
THE EMPIRICALLY DETERMINED SNR-DEPENDENT α

SNR (in dB)	-5	0	5	10	15
α	-0.25	-0.12	0.07	0.42	0.79

These SNR-dependent α values will be used to derive η as in (16) for the separation tests in the following experiments.

V. EXPERIMENTS

A. Experimental setup

A series of separation experiments were conducted to evaluate the proposed system performance. Ten speech files randomly selected from the TIMIT corpus are used in the experiment. The duration of each speech file is about 3 seconds. For music, we selected 3 kinds, namely, piano, pop and rock, 10 segments for each kind. The duration of each music segment is about 10 seconds. Therefore, there are totally 30 pairs of $y_m(t)$ and $y_s(t)$ for the experiment. All speech and music files used in the test experiments are different from those used for training the SNR-dependent α as described in Section IV. Each pair of $y_m(t)$ and $y_s(t)$ are added to generate the mixture $y(t)$ with varying energy of $y_m(t)$ according to the SNRs. The duration of $y(t)$ is about 10 seconds, within which only about 3 seconds segment containing the added speech and music, other containing music only. The SNR is defined as
$$SNR = 10 * \log_{10} \frac{E\{y_s(t)\}}{E\{y_m(t)\}}, \quad (20)$$
where $E\{\}$ denotes the time-duration normalized signal energy.

B. Baseline system for comparison

For comparison, a semi-blind NMF with four factorization constraints as proposed in [12] is adopted as the baseline system. In this system, basis matrix generated from the music

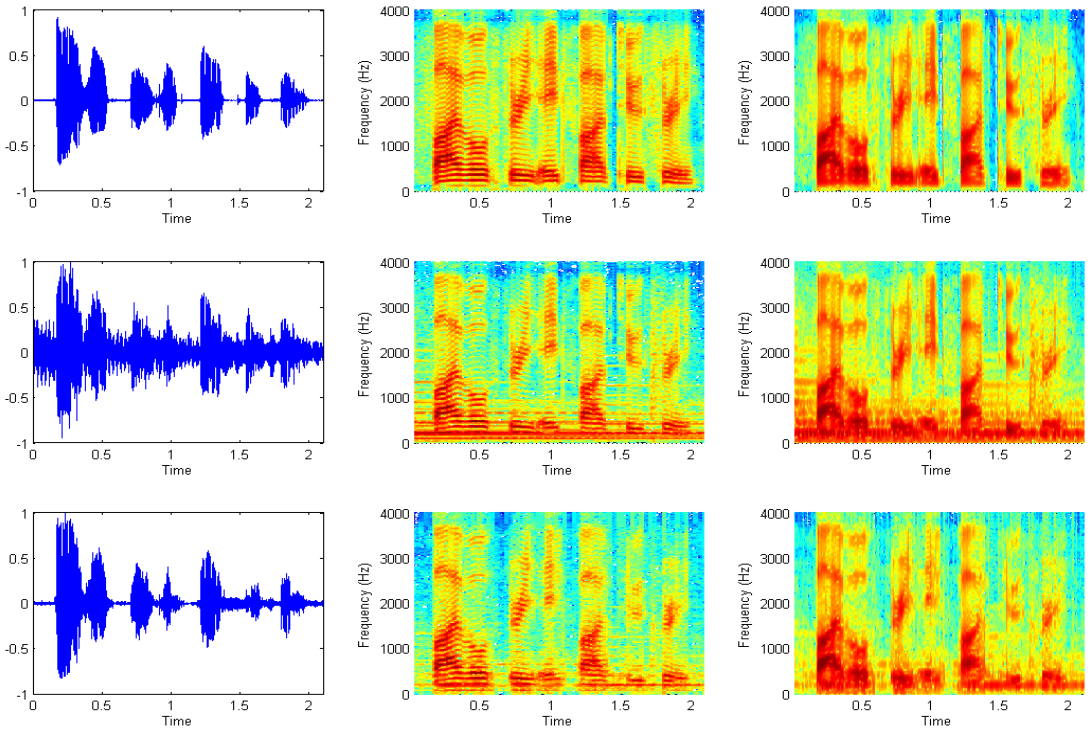

Fig. 3 Waveforms, narrow- and wide-band spectrograms of clean speech, 0 dB mixture, and reconstructed speech.

TABLE II
SNR (IN DB) OF MIXTURE AND RECONSTRUCTED SPEECH BY TWO SEPARATION SYSTEMS: BASELINE/**PROPOSED**

Mixture	Separated speech		
	Piano	Pop	Rock
-5	1.90 / **6.55**	1.48 / **4.27**	1.42 / **4.49**
0	5.15 / **9.01**	3.63 / **6.90**	3.23 / **6.70**
5	7.32 / **11.23**	6.86 / **9.23**	7.05 / **10.15**
10	10.31/**13.85**	9.62 /**12.03**	9.87 / **13.71**
15	12.08/**16.55**	12.16/**16.01**	13.20/**17.64**

only segments was adopted as the initial music basis matrix. The difference of the baseline method from the proposed one is that the speech-music separation is achieved as in (5) with the following factorization constraints

$$\zeta = \zeta_{Euc} + w_o \zeta_{Oth} + w_t \zeta_{Tem}(\mathbf{G}_m) + w_s \zeta_{Spa}(\mathbf{G}_m), \qquad (21)$$

where w_o, w_t and w_s are pre-trained weighting parameters of the respective constraints. No further classification of NMF components is implemented.

C. Experimental results

Figure 3 shows an example of the separation results. In this example, waveforms and spectrograms (narrow- and wide-band) are given for clean speech, 0 dB mixture and reconstructed speech. We can see from the waveforms that most of the music components have been removed and some distortion happens on the speech. The spectrograms tell that harmonic and formant structures of the original speech have been well preserved in the reconstructed speech.

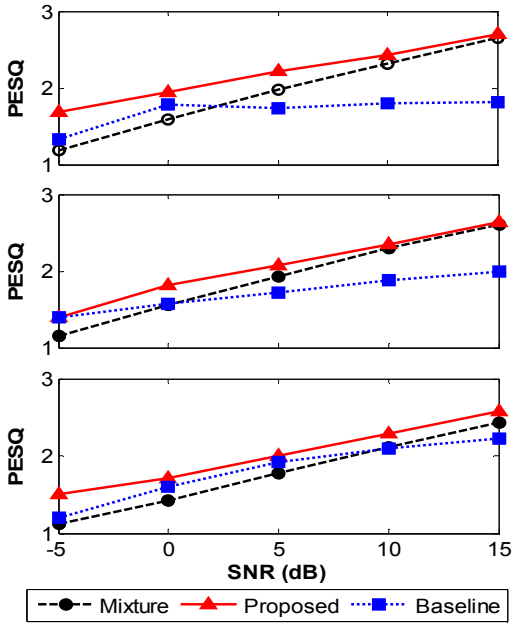

Fig. 4 PESQ of mixture and reconstructed speeches with two separation methods. The interfering music is piano, pop and rock for the subfigures (top to bottom), respectively.

Table 2 and Figure 4 compare the separation performances of the proposed system and the baseline system. Mixture signals with SNR (speech to music) of -5, 0, 5, 10, 15 dB have been tested. For each SNR condition, three different kinds of interfering music, e.g. piano, pop and rock have been used. We can see from both SNR and PESQ performances that the baseline system works only for mixtures with SNR lower than 5 dB; for mixture with SNR greater than 5 dB, its performance is insignificant or even degrades. This is consistent to previous studies on NMF-based speech and music separation, e.g. [7, 8, 11, 12], where only performance for mixture of 0 dB and lower were reported.

For the proposed system, it is clear that both SNR and PESQ have been improved and, especially, significant improvement has been achieved at low SNR cases. For high SNR cases, the proposed system also achieves better performance.

VI. CONCLUSIONS

A NMF-based speech and music separation system was presented. Music only segments were firstly detected via audio segmentation. The NMF output, e.g. the basis matrix of the music only segments provides the priori knowledge for music components in the mixture. With the help of such priors, speech/music classification of the basis and gain vectors of the mixture was obtained by comparing the vector similarity between the basis vector and the priori music bases to a SNR-dependent thresholding coefficient. Experimental results showed the effectiveness of the proposed separation method for mixture with a wide range of SNRs, e.g. from -5 dB to 15 dB. The superiority of the proposed system over the baseline system was also demonstrated.

REFERENCES

[1] J. T. Chien, and B. C. Chen, "A new independent component analysis for speech recognition and separation," *IEEE Trans. Audio, Speech, and Language Processing*, 14(4), pp. 1245-1254, 2006.

[2] G. Hu, and D. L. Wang, "An auditory scene analysis approach to monaural speech segregation," In Hansler E. and Schmidt G. (ed.), *Topics in Acoustic Echo and Noise Control*, Springer, Heidelberg, pp. 485-515. 2006.

[3] H. Sameti, and L. Deng, "Nonstationary-state hidden Markov model representation of speech signals for speech enhancement," *Signal Processing*, 82, pp.205-227, 2002.

[4] X. Xiao, P. Lee, and R. M. Nickel, "Inventory based speech enhancement for speaker dedicated speech communication systems", *Proc. ICASSP* 2009, pp.3877-3880.

[5] M. Ji, R. Srinivasan, and D. Crookes, "A corpus-based approach to speech enhancement from nonstationary noise," *Proc. Interspeech* 2010, Makuhari, Chiba, Japan, pp. 1097-1100.

[6] D. Lee, and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Inforamtion Processing Systems*, 13, pp. 556-562, 2001.

[7] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, and Language Processing*, 5(3), pp. 1066–1074, 2007.

[8] T. Virtanen, and A. T. Cemgil, "Mixtures of gamma priors for non-negative matrix factorization based speech separation," *Proc. of the 8th International Conference on Independent Component Analysis and Blind Signal Separation*, 2009.

[9] E. M. Grais, and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," *Proc. 7th International Conference on Digital Signal Processing*, 2011, pp. 1-6.

[10] Nengheng Zheng, Tan Lee, and Chun-Man Mak, "Model-based non-negative Matrix factorization for single-channel speech separation," *Proc. IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC)*, pp. 385-388, Xi'an, China, September 2011.

[11] S.-Y. Jeong, K. Kim, J.-H. Jeong, and K.-C. Oh, "Semi-blind disjoint non-negative matrix factorization for extracting target source from single channel noisy mixture," *Proc. IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, pp. 18-21, 2009.

[12] Y. Cai, N. H. Zheng, and X. Li, "Semi-blind Speech and Music Signal Separation based on Non-negative Matrix Factorization, " *Proc. 6th National conference on Harmonic Human-Machine Environment: Multitmedia Technology*, 2010 (in Chinese).

[13] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, 5, pp. 1457–1469, 2004

[14] 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," *Proc. ICASSP*, pp.749-752, 2001.