Biomedical Signal Processing and Control 36 (2017) 168–175

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Non-negative matrix factorization for speech/music separation using source dependent decomposition rank, temporal continuity term and filtering

S. Abdali, B. NaserSharif*

Department of Computer Engineering, K.N. Toosi University of Technology, Tehran, Iran

ARTICLE INFO

Article history:
Received 24 November 2016
Received in revised form 11 March 2017
Accepted 26 March 2017

Keywords:
Non-negative matrix factorization (NMF)
Cost function
Regularization term
Filter
Decomposition rank

ABSTRACT

Non-negative matrix factorization (NMF) is a recently well-known method for separating speech from music signal as a single channel source separation problem. In this approach, spectrogram of each source signal is factorized as a multiplication of two matrices known as basis and weight matrices. To obtain a good estimation of signal spectrogram, weight and basis matrices are updated based on a cost function, iteratively. In standard NMF, each frame of signal is considered as an independent observation and this assumption is a drawback for NMF. For overcoming this weakness, a regularization term is added to the cost function to consider spectral temporal continuity. Furthermore, in the standard NMF, the same decomposition rank is usually used for different sources. In this paper, in accompany with using a regularization term, we propose to apply a filter to the signals estimated by NMF. The filter is constructed by signals which are estimated using a regularized NMF method. Moreover, we propose to use different decomposition ranks for speech and music signals as different sources. Experimental results on one hour of speech and music signals show that the proposed method increases signal to inference ratio (SIR) values for speech and music signals in comparison to conventional NMF methods.

1. Introduction

Speech/Music separation is one of the single channel source separation methods that can be used as a pre-processing step for different applications such as in-car voice command systems or searching keywords in spoken news archives. It can be also used for improving the quality of hearing aid's output (as a medical equipment) when background music is considered as the background noise.

Many approaches have been proposed for single channel source separation problem. Most of these approaches need training data for each source signal. The training data can be modeled by probabilistic models such as Gaussian Mixture Model (GMM) or Hidden Markov Model (HMM).

These models are used in the separation stage under this assumption that appeared sources in the mixed signal have the same energy level as the training data. Resolving this limitation requires complicated computations [1–5].

Another technique for training data modeling is to train non-negative basis sets for each source. This method is more flexible where it does not need any assumption about the energy differences between the signals in the training and separation stages. The main drawback of this approach is that any non-negative linear combination of the trained vectors is not necessarily a valid estimation of the source signals. This may decrease separation quality [9]. Non-negative matrix factorization (NMF) is an example of such methods. Usually, NMF method is used for decomposing a mixed signal spectrogram. In standard NMF, each source signal is defined as the multiplication of two matrices with non-negative elements known as weight and basis matrices. NMF has two stages: training and test (decomposition) stages. In the training stage, training data are used for training basis vectors of each source's basis matrix. In this stage, a cost function is used to estimate the matrices. Then, in the decomposition stage, estimated basis vectors are used for frame-by-frame separation of mixed signal without considering any smooth transition between frames or other information that may exist in the subsequent frames. Until now, different cost functions such as Euclidean distance, Kullback–Leibler (KL) and Itakora–Saito (IS) divergences have been used [6–10]. A problem with Standard NMF is that each element of the basis matrix is considered as an independent observation. Many approaches have

* Corresponding author.
E-mail address: bnasersharif@kntu.ac.ir (B. NaserSharif).

http://dx.doi.org/10.1016/j.bspc.2017.03.010
1746-8094/

been proposed to solve this problem. One of these approaches is to add a regularization term to the cost function of NMF, which codes prior knowledge and shows temporal continuity of spectrum [11,12]. In fact, this term is used to penalize big differences of adjacent frames in the weight matrix [11,13]. To determine this term, different methods have been proposed based on used NMF cost function and its mathematical properties. For example, in [14,15], a term is considered to compensate sparseness of weight matrix besides temporal continuity term. In another approach, the temporal continuity is considered using a regularization term based on HMM and smoothness for rows of weight matrix [9]. In [16,17], this regularization term is obtained using statistical properties of spectrum's temporal continuity. Moreover, it has been proposed to apply post processing methods such as Wiener mask for enhancing source separation [18].

In this paper, for improving separation quality, besides using regularization term, we propose to apply a filter on separated output signals. In addition, we propose to use different decomposition ranks for music and speech signals.

The remainder of this article is organized as follows: Section 2 introduces the NMF method mathematically, then in Section 3, KL divergence, which broadly is used as the NMF cost function for speech processing, has been described. In Section 4, our proposed method and the motivations have been propounded and finally in Sections 5 and 6 the experimental results and conclusion are presented, respectively.

2. Non-negative matrix factorization

2.1. Basic definition

Mathematically, the NMF is formulated as follows. Let $V \in R_+^{M\times N}$ be a non-negative matrix, mean all the coefficients of which are positive or null, of size $M \times N$ (in music applications, V will very often the amplitude spectrogram). Non-negative matrix factorization approximates V by $\tilde{V}$ as follows:

$$\tilde{V} = BW \tag{1}$$

where $B \in R_+^{M\times K}$ and $W \in R_+^{K\times N}$, and where K is the factorization rank, generally chosen such that $K(M+N) \leq MN$. The matrix B is called basis or the codebook. The matrix W is called the weight or activation matrix. Each column vector in matrix V is estimated by a weighted linear combination of basis vectors which are the same B columns. Weights for basis vectors appear in corresponding columns in matrix W. To approximate data in V as a non-negative linear combination of its component vectors, the non-negative basis vectors in matrix B are optimized. The following figure shows an example of how to decompose a piece of music [9] (Fig. 1):

The matrices B and W are estimated by solving following optimization problem [18]:

$$\min C(V||BW) \quad \text{where } B, W \geq 0 \tag{2}$$

where C is a cost function which estimates distance between V and BW. Different cost functions lead to different kinds of NMF. One of the well-known cost functions is KL divergence.

2.2. Training of speech and music

With two sets of training data for speech and music signals, the Fast Fourier Transform (FFT) is computed for each signal to obtain magnitude spectrogram of speech and music signals. Then, NMF is used for decomposing speech and music spectrograms into base and weight matrices. In other word, the aim of using NMF is to model the training data as a set of basis vectors to represent the spectral characteristics for each source signal [18].

$$S_{\text{train}} \approx B_{\text{speech}} W_{\text{speech}} \tag{3}$$

$$M_{\text{train}} \approx B_{\text{music}} W_{\text{music}} \tag{4}$$

Based on the used cost function, B and W are updated iteratively. B_{speech} and B_{music} have normalized columns, and after each iteration their columns should be normalized again. The initial values for B and W are random positive values [18].

2.3. Decomposition of the mixed signal

In the decomposition stage, NMF should be used again to decompose the signal X spectrogram. In this step, basis matrix is obtained from the training phase matrices as follows [18]:

$$X \approx \left[B_{\text{speech}} B_{\text{music}}\right] W \tag{5}$$

$$\tilde{S} \approx B_{\text{speech}} W_{\text{s}} \tag{6}$$

$$\tilde{M} \approx B_{\text{music}} W_{\text{m}} \tag{7}$$

where W_s and W_m are sub matrices in matrix W which correspond to the speech and music components respectively. $\tilde{M}$ and $\tilde{S}$ matrices contain estimations for the spectral magnitude of the music and speech signals [18].

3. NMF cost functions and regularization term

3.1. KL based cost function

Cost function based on KL divergence is defined as [7,8]:

$$D_{KL}(V||BW) = \sum_{i,j}\left(V_{ij}\log\frac{V_{i,j}}{(BW)_{i,j}} - V_{i,j} + (BW)_{i,j}\right) \tag{8}$$

To minimize the KL based cost function, B and W matrices can be computed through the following iterative updates:

$$B \longleftarrow B \otimes \frac{(V/BW)W^{\mathrm{T}}}{1W^{\mathrm{T}}} \tag{9}$$

$$W \longleftarrow W \otimes \frac{B^{\mathrm{T}}(V/BW)}{B^{\mathrm{T}}1} \tag{10}$$

where 1 is a matrix of ones with the same size of V, all multiplication and divisions are element wise and $\otimes$ indicates element wise multiplication [6,7].

As mentioned before, standard NMF does not consider any probabilistic assumption and so basis vector elements are considered independent. This tends to poor basis and weight vectors estimation. Some approaches use prior knowledge of basis vectors in the standard NMF training step [14].

A well-known method for considering the prior knowledge is based on adding a regularization term to the NMF cost function. This regularization term considers temporal continuity of components to penalize large changes between the weights of adjacent frames. This term can be computed as the sum of the squared differences between the weights as follows [14]:

$$C_{\text{TC}} = \sum_{j=1}^{J}\frac{1}{\sigma_j^2}\sum_{t=2}^{T}\left(w_{t,j} - w_{t-1,j}\right)^2 \tag{11}$$

$$\sigma_j = \sqrt{\frac{1}{T}\sum_{t=1}^{T} w_{t,j}^2} \tag{12}$$

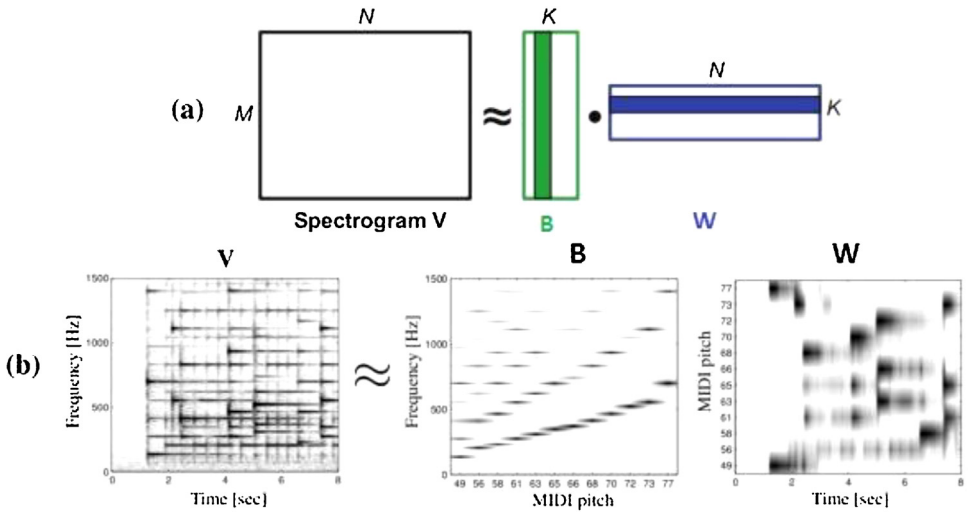

Fig. 1. Non-negative matrix factorization.

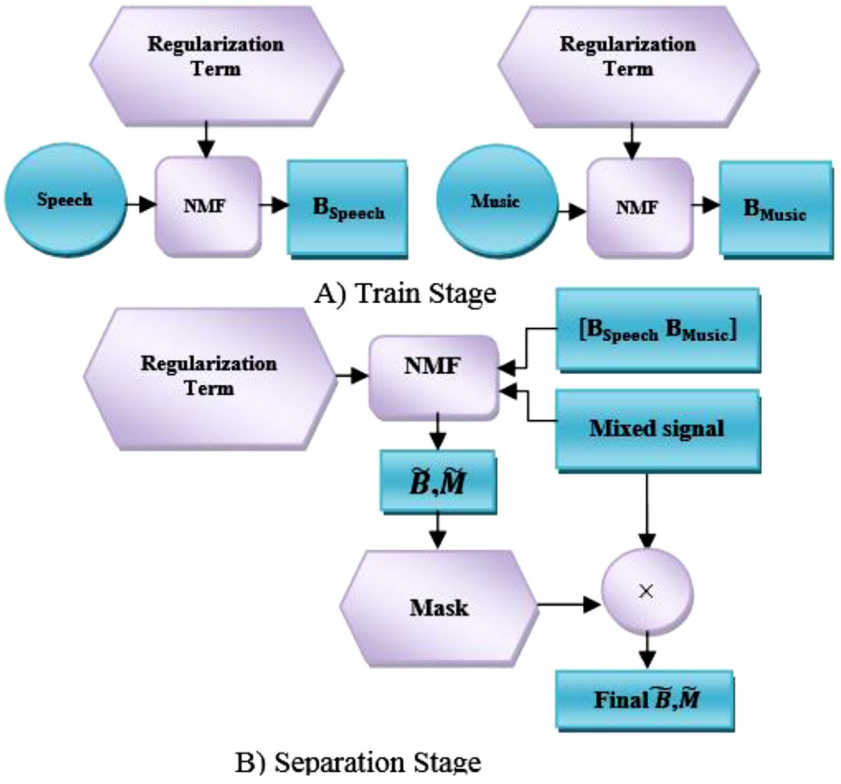

Fig. 2. Proposed method.

σ_j is an estimation of weights standard deviation used to prevent the numerical scale of the weights from affecting the penalty cost. Finally, the cost function can be rewritten as:

$$C_{KLTC} = D_{KL}(V||BW) + C_{TC}. \quad (13)$$

4. Proposed method

As mentioned before, a regularization term is added to NMF cost function to consider spectral temporal continuity in the basis matrix. In this work, for improving the quality of separated signals, we propose to apply a filter to separated signals obtained from NMF with a regularization term. Fig. 2 shows the process in detail. Moreover, by considering special properties of music and speech signals, we propose to use different decomposition ranks for each signal. In Section 4.1, we will discuss this method and motivations.

4.1. Using different decomposition ranks for different sources

Our observations show that generally, music spectra have more variation than speech spectra. Fig. 3 shows variances of music and speech power spectra in a 30 ms frame. Vowels and consonants are basic speech sounds, which have different spectral variances.

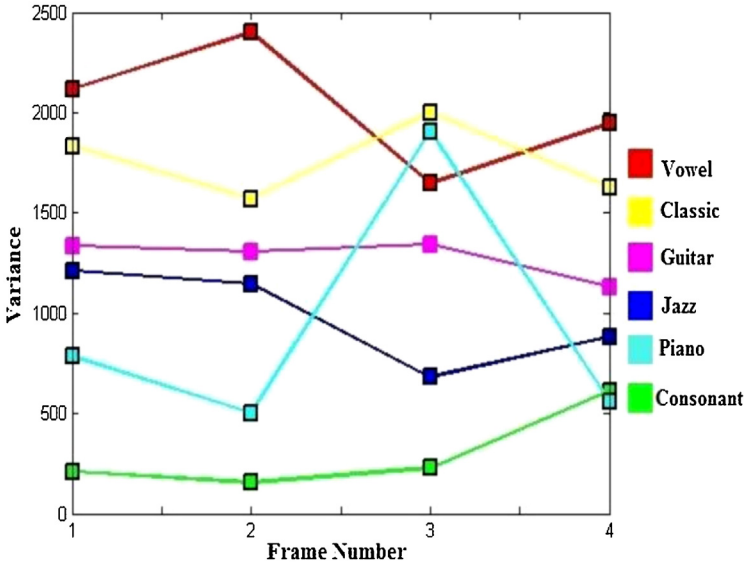

Fig. 3. Variance of music and speech power spectra in 30 ms frames for consonant "*f*" and vowel "a" and 4 genres of music.

Based on this difference, we should use different ranks for different parts of speech signal and should have different basis vectors for vowel and consonants, but this has its own difficulties. Therefore, in this paper, we focus only on music and because spectral variance of music is usually higher than that of consonant, it is reasonable to have smaller decomposition ranks for music signals. Therefore, we propose to use smaller rank for music signal in comparison to speech rank. However, according to the relation (5), the basis matrices of speech and music signals should have the same size. So, we propose to repeat the basis matrix of music signal to reach the size of speech basis matrix. Then, to reconstruct the music signal, we separate the non-repetitive part of music basis matrix and reconstruct the music signal using this non-repetitive part. This procedure is shown in Fig. 4 for sample music rank 256 and speech rank 1024.

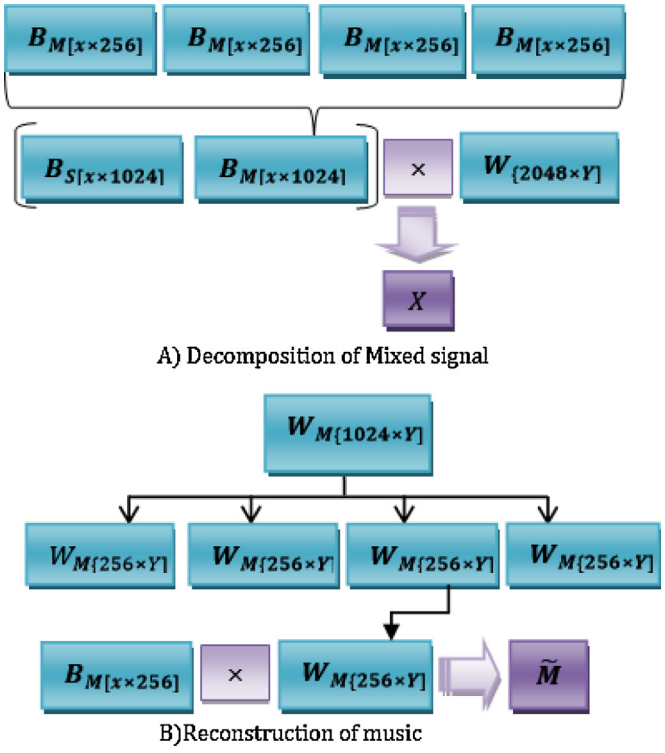

Fig. 4. Using different ranks for different sources.

4.2. Filter

In NMF based speech/music signal separation, $\tilde{S}$ and $\tilde{M}$ shown in relations (6) and (7) are usually used as final estimations of source signals spectrograms directly, although estimated spectrograms $\tilde{S}$ and $\tilde{M}$ are not the exact summation of spectrogram X and the decomposition error is not equal to zero. If we assume that the existing noise in the mixed signal is small, we can write:

$$X \approx \tilde{S} + \tilde{M} \tag{18}$$

To make the decomposition error zero, we use the initial estimated $\tilde{S}$ and $\tilde{M}$ to construct a mask as follows:

$$H = \frac{\tilde{S}^p}{\tilde{S}^p + \tilde{M}^p} \tag{19}$$

where $p > 0$ is a constant parameter and division is element wise. It can be seen that $H \in (0, 1)$ and using different values for p produce different masks. These masks scale each frequency component in the mixed signal with a rate that describes how much each signal participates in the mixed signal [18]:

$$\tilde{S} = H \otimes X \tag{20}$$

$$\tilde{M} = (1 - H) \otimes X \tag{21}$$

$\tilde{S}$ and $\tilde{M}$ are final estimations of speech and music signals, respectively. In relations (20) and (21), 1 is a matrix which all its elements are one and $\otimes$ indicates element wise multiplication.

Using this mask, we can reduce estimation error to zero and guarantee that summation of source signals is equal to the mixed signal.

If $p = 2$ in Eq. (19), we obtain Wiener mask:

$$H_{\text{wiener}} = \frac{\tilde{S}^2}{\tilde{S}^2 + \tilde{M}^2} \tag{22}$$

In this work, we have tried different masks to evaluate the effect of parameter p on the quality of separation.

5. Experimental results

In our evaluation experiments, sampling rate of speech and music signals is 16 kHz and we used 30 ms frames with 50% overlap. For speech training data, 1081 utterances of both male and female speakers have been chosen fr TIMIT database, which this set, totally consists of 30 min of speech. For music training set, 4 jingles of pop, jazz, piano and classical music have been chosen from GTZAN database [21]. Theses 4 jingles have been repeated equally to reach the length of speech set. So, totally, 1 h of speech and music signals are used for the training set. For speech test set, 100 utterances with length 1.5 s have been chosen from TIMIT database. Moreover, for music test set, the same 4 jingles of GTZAN are used.

To discuss the effects of choosing different ranks for different sources, first, we chose rank 1024 for both speech and music signals. Then, for comparing equal rank approach with different rank approach, in first experiment, we fixed speech rank equal to 1024 and then change the rank of music signal. In second experiment, we tried speech rank 512 for different music ranks. These results are reported in Tables 1 and 2.

Meanwhile, as it has been shown in the works of Emad et al. [20] that the best results are obtained for $p = 2$, we have tried different values greater than 2 for this parameter to evaluate its effect on the quality of separation.

To compare and evaluate the methods, we have used 3 standard measurements: Signal to Inference ratio (SIR), Signal to Artifact ratio (SAR) and Signal to Distortion ratio (SDR) [19]. These numerical performance criteria are calculated using energy ratios expressed in decibels (dB). In [19], source to distortion ratio is defined as:

$$\mathrm{SDR} = 10\log_{10}\frac{{s_{\mathrm{target}}}^2}{{e_{\mathrm{interf}} + e_{\mathrm{music}} + e_{\mathrm{artif}}}^2} \tag{23}$$

It measures the amount of distortion introduced in the output signal and is defined as the ratio between the energy of the clean signal and that of the distortion.

The source to interferences ratio is defined as the ratio between the power of the target signal and that of the interference signal and is used to measure the amount of undesired interference remaining in the separated signal:

$$\mathrm{SIR} = 10\log_{10}\frac{{s_{\mathrm{target}}}^2}{{e_{\mathrm{interf}}}^2} \tag{24}$$

The source to artifact ratio measures the quality in terms of absence of artificial noise:

$$\mathrm{SAR} = 10\log_{10}\frac{{s_{\mathrm{target}} + e_{\mathrm{interf}} + e_{\mathrm{Music}}}^2}{{e_{\mathrm{artif}}}^2} \tag{25}$$

These four measures are inspired by the usual definition of the SMR, with a few modifications.

$$\mathrm{SMR} = 10\log_{10}\frac{{s_{\mathrm{target}}}^2}{{e_{\mathrm{Music}}}^2} \tag{26}$$

In these measures s_{target} denotes a version of the true desired source modified by distortions and e_{interf}, e_{Music} and e_{artif} are the interferences, music and artifacts error terms, respectively [19].

We have reported our results for different input signal to music ratio (SMR) values: 5, 10, 15 and 20 db. In all tables, TC, R_s and R_M stands for temporal continuity, speech rank and music rank, respectively.

The experimental results of KL based methods have been shown in Tables 1 and 2. Table 1 includes the results of speech rank 512 and different music ranks including 512, 256, 128 and 64. According to these results, SIR values (the best measurement for separation ratio) of NMF-KL + Wiener and NMF-KL + TC are higher than standard KL. When we use both TC and Wiener simultaneously, SIR values increase by almost 1 db. Other results are related to using different ranks. As we expect, by decreasing rank of music, SIR values of both speech and music signals are increased considerably. When we choose music rank 64, we have the highest improvement for music SIR values. As can be seen in Table 2, which includes the results of speech rank 1024, we obtain improvements in SIR values just like the results of rank 512. The only difference is that choosing speech rank 1024 tends to higher speech SIRs in comparison to Table 1. In addition, SAR and SDR values of proposed method are better than those of standard KL in both tables. This shows that our proposed method does not increase the distortion and artificial noise of separated speech signal.

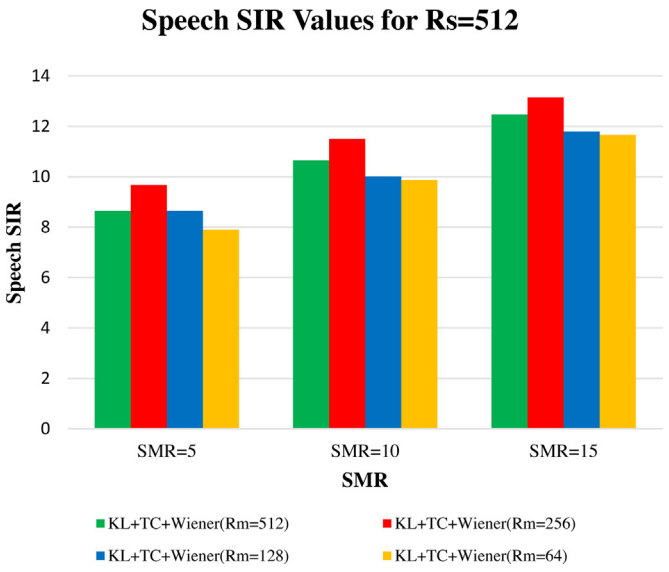

Fig. 5. Speech SIR values vs. SMR values for Rs = 512.

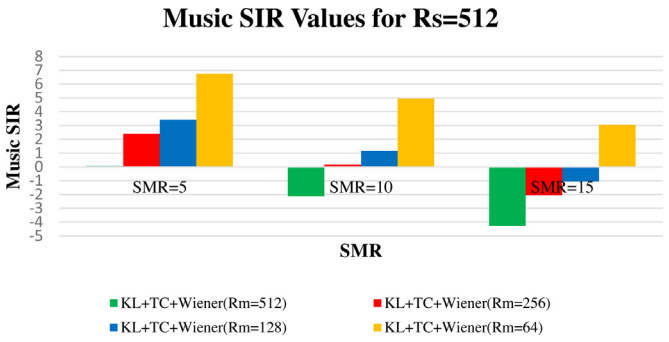

Fig. 6. Music SIR values vs. SMR values for Rs = 512.

Table 3 represents the effect of choosing different values for parameter p in relation (19). Based on the result shown in this table, by increasing SIR values of both speech and music signals increase. On the other hand, SDR and SAR values decrease. Therefore, we can conclude that for higher p values, we have better separation ratio, but more distortion and artificial noise as well. So, choosing a value for parameter p in relation (19) is a trade-off between SIR values and SDR/SAR values.

For a better representation of results, changes in SIR values for different SMR values have been shown in Figs. 5–10.

Table 1
Experimental results of KL based methods for speech rank 512.Bold values shows the best obtained results.

	Input SMR (db)	SDR speech (db)	SAR speech (db)	SIR speech (db)	SIR music (db)
NMF-KL	5	−2.6725	−1.3485	7.1152	−0.5862
$R_s = 512$	10	−2.0894	−1.2267	9.2788	−2.8329
$R_M = 512$	15	−1.7127	−1.1424	11.2726	−5.0034
NMF-KL + Wiener	5	−2.8214	−1.8341	8.4644	−0.2075
$R_s = 512$	10	−2.3386	−1.6736	10.4449	−2.4635
$R_M = 512$	15	−2.0454	−1.5017	12.2173	−4.6115
NMF-KL + TC	5	−2.6794	−1.4235	7.2810	−0.2091
$R_s = 512$	10	−2.1277	−1.3191	9.4926	−2.4461
$R_M = 512$	15	−1.8130	−1.2826	11.4845	−4.5662
NMF-KL + TC + Wiener	5	−2.8334	−1.9055	8.6478	0.0684
$R_s = 512$	10	−2.3446	−1.7263	10.6522	−2.1300
$R_M = 512$	15	−2.0509	−1.6289	12.4682	−4.2604
NMF-KL + TC + Wiener	5	−2.1318	−1.3597	**9.6783**	**2.3993**
$R_s = 512$	10	−1.7167	−1.1849	**11.5060**	**0.1695**
$R_M = 256$	15	−1.4509	−1.0737	**13.1409**	**−2.0538**
NMF-KL + TC + Wiener	5	−2.1730	−1.0583	**8.6478**	**3.3993**
$R_s = 512$	10	−1.5753	−0.8190	**10.0048**	**1.1743**
$R_M = 128$	15	−1.1803	−0.6555	**11.7892**	**−1.0390**
NMF-KL + TC + Wiener	5	−1.9178	−0.7509	**7.9014**	**6.7364**
$R_s = 512$	10	−1.3199	−0.5241	**9.8676**	**4.9414**
$R_M = 64$	15	−0.9586	−0.4073	**11.6620**	**3.0608**

Table 2
Experimental results of KL based methods for speech rank 1024.

	Input SMR (db)	SDR speech (db)	SAR speech (db)	SIR speech (db)	SIR music (db)
NMF-KL	5	−2.4434	−1.3795	8.1877	0.7183
$R_s = 1024$	10	−1.9345	−1.2455	10.3295	−1.4532
$R_M = 1024$	15	−1.6211	−1.1504	12.1709	−3.6005
NMF-KL + Wiener	5	−2.5637	−1.8320	10.0166	1.0237
$R_s = 1024$	10	−2.1377	−1.6382	11.8525	−1.1654
$R_M = 1024$	15	−1.8289	−1.4684	13.4154	−3.2703
NMF-KL + TC	5	−2.4573	−1.4570	8.5224	0.8526
$R_s = 1024$	10	−1.9775	−1.3256	10.6213	−1.2654
$R_M = 1024$	15	−1.7000	−1.2605	12.4645	−3.3732
NMF-KL + TC + Wiener	5	−2.5744	−1.8772	10.3534	1.2581
$R_s = 1024$	10	−2.1562	−1.6781	12.1448	−0.8996
$R_M = 1024$	15	−1.8957	−1.5532	13.6712	−3.0049
NMF-KL + TC + Wiener	5	−2.0311	−1.3301	10.2034	2.3306
$R_s = 1024$	10	−1.6623	−1.1737	11.9560	0.0780
$R_M = 256$	15	−1.4284	1.0722	13.4537	−2.0963
NMF-KL + TC + Wiener	5	−2.2453	−1.1764	8.2189	2.2836
$R_s = 1024$	10	−1.6324	−0.9072	10.1893	0.2287
$R_M = 128$	15	−1.2495	−0.7419	11.9553	−1.8601
NMF-KL + TC + Wiener	5	−2.0570	−0.8744	7.7960	6.1027
$R_s = 1024$	10	−1.4374	−0.6337	9.7841	4.4212
$R_M = 64$	15	−1.0421	−0.4844	11.5790	2.5008

Table 3
Experimental results for choosing different p values for speech rank 1024.

	Input SMR (db)	SDR speech (db)	SAR speech (db)	SIR speech (db)	SIR music (db)
NMF-KL + TC + Wiener	5	−2.5744	−1.8772	10.3534	1.2581
	10	−2.1562	−1.6781	12.1448	−0.8996
	15	−1.8957	−1.5532	13.6712	−3.0049
NMF-KL + TC + Filter ($p = 3$)	5	−2.6542	−1.9931	10.6800	1.3512
	10	−2.2999	−1.8617	12.6057	−0.683
	15	−2.0453	−1.7194	13.9333	−2.8359
NMF-KL + TC + Filter ($p = 4$)	5	−2.8055	−2.1967	11.1714	1.4594
	10	−2.3762	−1.9489	12.7714	−0.6873
	15	−2.1346	−1.8144	14.0030	−2.8934

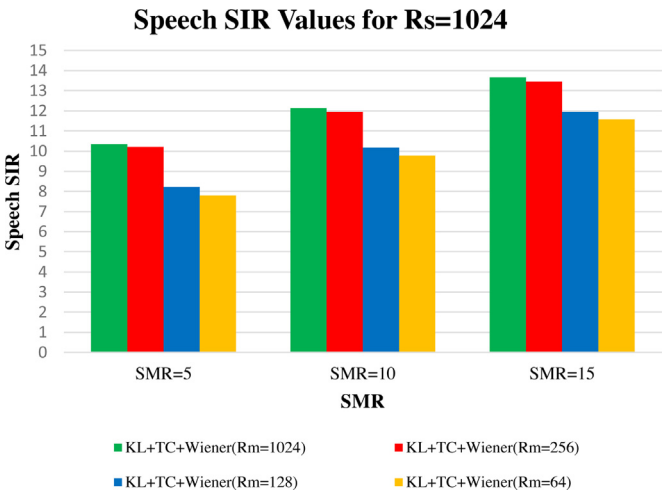

Fig. 7. Speech SIR values vs. SMR values for Rs =1024.

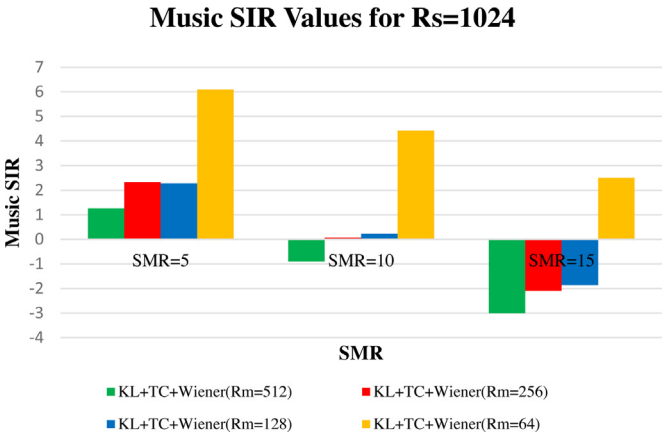

Fig. 8. Music SIR values vs. SMR values for Rs = 1024.

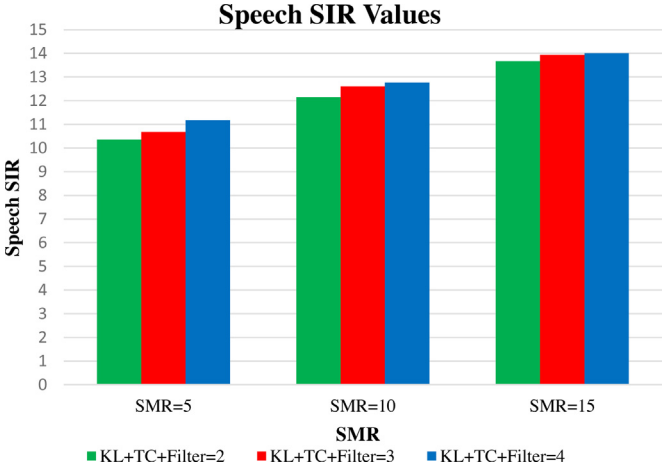

Fig. 9. Speech SIR values vs. SMR values for different *p*.

6. Conclusion

In this paper, for separating speech from music signal, we proposed a novel approach to improve performance of non-negative matrix factorization algorithm. Meanwhile, we proposed to use different decomposition ranks for speech and music signals as different sources. We used smaller ranks for music signal than speech signal rank. This is due to the fact that there is more variation in music spectrum in comparison to speech spectrum. Thus, we need smaller basis vectors to describe behavior of music spectrum. Furthermore, besides using a regularization term, we apply a filter to signals estimated by NMF. In this way, we can obtain better estimations for music and speech signals because of constructing a better filter. Our experiments on one hour of speech and music signals show that the improved NMF method has higher signal to inference ratio (SIR) values for both speech and music signals in comparison to the conventional NMF methods.

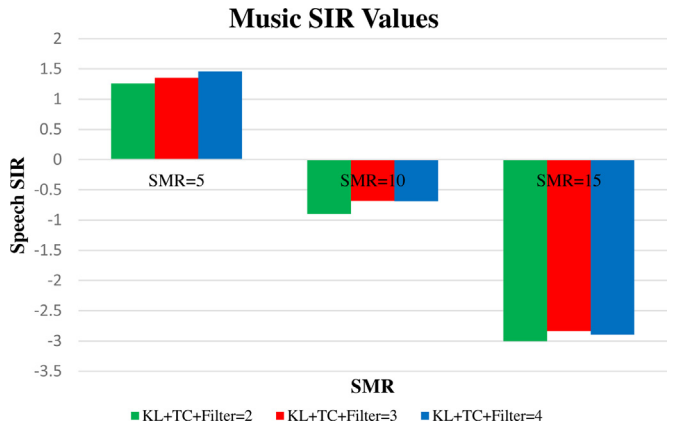

Fig. 10. Music SIR values vs. SMR values for different *p*.

References

[1] E.M. Grais, H. Erdogan, Regularized non-negative matrix factorization using gaussian mixture priors for supervised single channel source separation, Comput. Speech Lang. (2013) 746–762.
[2] E.M. Grais, H. Erdogan, Source separation using regularized NMF with MMSE estimates under GMM priors with online learning for the uncertainties, Digital Signal Process. (2014) 20–34.
[3] M.H. Radfar, W. Wong, R.M. Dansereau, W.Y. Chan, Scaled factorial hidden markov models: a new technique for compensating gain differences in model-based single channel speech separation, IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP) (2010) 1918–1921.
[4] E.M. Grais, M. Umut Sen, H. Erdogan, Deep neural networks for single channel source separation, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2014, May) 3734–3738.
[5] X. Jauregui berry, E. Vincent, G. Richard, Multiple-order non-negative matrix factorization for speech enhancement, Annual Conference of the International Speech Communication Association (INTERSPEECH) (2014, June) 2839–2842.
[6] E.M. Grais, Nonnegative Matrix Factorization for Audio Source Separation (PHD thesis), Sabanci University, 2013.
[7] D.D. Lee, H.S. Seung, Learning the Parts of Objects with Nonnegative Matrix Factorization, 1999, pp. 788–791.
[8] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, Neural Inf. Process. Syst. (2001) 556–562.
[9] C. Févotte, N. Bertin, J. Durrieu, Non-negative matrix factorization: with the Itakura-Saito divergence with application to music analysis, Neural Comput. 21 (2009, March) 793–830.
[10] C. Févotte, Majorization-minimization algorithm for smooth Itakura-Saito non-negative matrix factorization, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2011, May) 1825–1828.
[11] T. Virtanen, A.T. Cemgil, S. Godsill, Bayesian extensions to non-negative matrix factorization for audio signal modeling, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2008, April) 1980–1983.
[12] T. Virtanen, Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria, IEEE Trans. Audio Speech Lang. Process. 15 (3) (2007, March) 1066–1074.
[13] N. Boulanger-Lewandowski, G.J. Mysore, M. Hoffman, Exploiting long-term temporal dependencies in NMF using recurrent neural networks with application to source separation, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2014) 6969–6973.
[14] T. Virtanen, Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria, IEEE Trans. Audio Speech Lang. Process. 15 (3) (2007, March) 1066–1074.
[15] T. Virtanen, Sound source separation using sparse coding with temporal continuity objective, in: Proc. Int. Comput. Music Conf., Singapore, 2003, pp. 231–234.
[16] K.W. Wilson, B. Raj, P. Smaragdis, Regularized non-negative matrix factorization with temporal dependencies for speech denoising, Annual

Conference of the International Speech Communication Association (INTERSPEECH) (2008) 411–414.
[17] C. Vaz, D. Dimitriadis, S.S. Narayanan, Enhancing audio source separability using spectro-temporal regularization with NMF, Annual Conference of the International Speech Communication Association (INTERSPEECH) (2014, September) 855–859.
[18] E.M. Grais, H. Erdogan, Single channel speech music separation using nonnegative matrix factorization with sliding windows and spectral masks, IEEE International Conference on Digital Signal Processing (DSP) (2011) 1–6.
[19] E. Vincent, R. Gribonval, C. Fevotte, Performance measurement in blind audio source separation, IEEE Trans. Audio Speech Lang. Process. 14 (4) (2006, July) 1462–1469.
[20] M. Emad Grais, H. Erdogan, Single channel speech music separation using nonnegative matrix factorization and spectral masks, in: Faculty of Engineering and Natural Sciences, Sabanci University, Orhanli Tuzla, 34956, Istanbul, 2011.
[21] http://marsyas.info/download/data_sets/.