NMF based speech and music separation in monaural speech recordings with sparseness and temporal continuity constraints

Ming Tu[1], **Xiang Xie**[1], **Yishan Jiao**[1]

Abstract. This paper proposes a semi-supervised approach of speech and music separation in monaural speech recordings based on non-negative matrix factorization (NMF). Considering the scenario that the genre of background music is known, music basis vectors are randomly picked from the magnitude of short time fourier transform (STFT) of training music, while speech basis vectors are estimated by executing NMF on the magnitude of STFT of polluted speech signal. Moreover, we apply sparseness and temporal continuity constraints to speech and music respectively and evaluate how different constraints can influence the separation performance. The test set contains 10 Mandarin speech utterances from 10 speakers mixed with music in different speech-music ratios (SMR). The baseline is semi-supervised separation system with no constraint. The results reveal that adding temporal continuity constraint can improve the separation performance compared with the baseline and separation system with only sparseness constraint.

Keywords: non-negative matrix factorization • speech and music separation • sparse coding • temporal continuity • semi-supervised learning

1 Introduction

Speech and music separation belongs to one specific problem of audio source separation, the applications of which include: speech enhancement when talking to mobile phone in loud music background, online video transcript interfered by music due to the amateurism of the uploader, in-car speech recognition with the background music from CD player or FM radio, etc. Although the research on audio source separation has achieved good results in these years and some of the methods also show feasibility on separation of speech and music, the problem of

[1] Ming Tu (✉), Xiang Xie, Yishan Jiao
Reseach Institue of Communication Technology, Beijing Institute of Technology
e-mail: tuming90@gmail.com

This work is supported by the National Natural Science Foundation of China (Grant No. 91120015, No. 90920304, No. 11161140319).

separating speech and music in monaural speech materials has been difficult due to the nonstationarity and diversity of music. A lot of dedication has been made to improve the performance of separation. These methods include: statistical modeling, such as Gaussian Mixture Model (GMM) [1, 8], vector quantization (VQ) [6], discrete energy separation algorithm (DESA) [12], computational auditory scene analysis (CASA) [11], NMF [3, 5, 9], etc. This paper mainly concerns about the NMF based method.

NMF is a matrix factorization paradigm with the criterion that all the matrixes in the factorization are non-nonegative. It inspires a method to decompose audio signals into weighted combination of several non-negative basis vectors. This characteristic facilitates NMF to be used in audio source separation. Among the methods mentioned above, NMF has obtained good results in terms of intelligibility, which makes the separated speech more understandable to listeners. In [3], an exemplar-based supervised method of suppressing background music was proposed. It trained the speech and music basis vectors using magnitude of STFT of the corresponding speech and music signals. The weighting matrix was estimated through factorization of magnitude of STFT of the mixture. The results showed that the method improved the performance of automatic speech recognition in terms of word error rate (WER). In [5], the author presented a semi-supervised speaker-dependent algorithm based on NMF. It fixed the speech basis vectors and estimated music basis vectors and weighting matrix through iteration with sparseness constraint on music basis vectors and corresponding weighting vectors. The experiment obtained outstanding performance. However, on one hand, in [5], the speaker-dependent system makes it limited in practical application because before separated, clean speech data was needed to train speech basis vectors for each speaker. In the full supervised algorithm of [3], the situation was worse. On the other hand, both the two methods did not make full use of a-priori information of speech and music under the situation that we have already known the two components were speech and music, even though both the methods were supervised or semi-supervised. Based on the above analysis, we extend the NMF based method to a more loosely limited situation. Besides the sparseness constraint, we add the temporal continuity constraint to music.

In this paper, the music basis vectors are fixed considering the situation that the genre of the background music is known. To make use of a-priori information of speech and music, we investigate the difference between speech and music in time-frequency domain. Because the energy of speech signal mainly focuses on some narrow bands and is sparse in time-frequency domain, we add the sparseness constraint to speech component of the weighting matrix. Further, in time-frequency domain, the spectrogram of music signal shows better continuity than speech signal because of its stronger harmonicity. Thus we add the temporal continuity constraint to music component of the weighting matrix. We extend both criteria to semi-supervised scenario in contrast to the unsupervised method in [9]. A speaker-independent test set with 10 speakers from the ASCCD corpus [2] is mixed with music fragments randomly picked from GTZAN Genre Collection. In this paper, two styles of music (classical and jazz) as background are evaluated. The proposed method gives an expectant result and the constraint of temporal continuity is proved to be valid. Section 2 introduces the NMF-based model. Experiment setup and the corresponding result are illustrated in section 3. Finally, in section 4, we conclude our work and make some prospects for future work.

2 NMF based Separation Methods

2.1 NMF based model

In our case, the speech and music signal are added together to construct the mixed signal. So, in time-frequency domain, it is easy to get

$$\mathbf{V}_t(f) = \mathbf{S}_t(f) + \mathbf{M}_t(f) . \qquad (1)$$

$\mathbf{V}_t(f)$, $\mathbf{S}_t(f)$, $\mathbf{M}_t(f)$ are signals' STFT correspondingly. The non-negative magnitude of $\mathbf{V}_t(f)$ can be factorized into two non-negative matrixes as following

$$\mathbf{V} = \mathbf{WH} , \qquad (2)$$

where $\mathbf{V}$ is a $m \times n$ matrix, $\mathbf{W}$ is a $m \times r$ matrix and $\mathbf{H}$ is a $r \times n$ matrix. r denotes the NMF dimensionality.

According to matrix multiplicative rules, equation (2) can be rewritten to

$$\mathbf{v}_j = \sum_{i=1}^{r} h_{i,j} \mathbf{w}_i . \qquad (3)$$

$\mathbf{v}_j$ is the jth column of $\mathbf{V}$, so is $\mathbf{w}_j$. $h_{i,j}$ is the element of row i, column j in matrix $\mathbf{H}$. Further, $\mathbf{w}_j$ can be deemed to be the basis vectors that represent one building block of the magnitude spectrogram of signal $v(t)$, and $h_{i,j}$ constructs the corresponding weighting matrix. r is the column number of $\mathbf{W}$. If the phase of every signal is overlooked, (1) can be expressed as following

$$\mathbf{V} = \mathbf{W}^{(s)}\mathbf{H}^{(s)} + \mathbf{W}^{(m)}\mathbf{H}^{(m)} + \varepsilon . \qquad (4)$$

$\mathbf{W}^{(s)}$, $\mathbf{H}^{(s)}$,$\mathbf{W}^{(m)}$, $\mathbf{H}^{(m)}$ denote the basis vectors of speech component in mixture and the corresponding weighting matrix, the basis vectors of music component and the corresponding weighting matrix respectively. ε is the reconstruction error. To factorize **V** in (4), we can use the multiplicative update rules first presented in [4], which is not the fastest but effective enough for our method.

2.2 Proposed method

In our method, we assume the genre of the mixture's background music is known. We first obtain $\mathbf{W}^{(m)}$ from the magnitude of STFT of the music training set. Then, during the iteration, $\mathbf{W}^{(m)}$ is fixed and finally we get non-negative $\mathbf{W}^{(s)}$, $\mathbf{H}^{(s)}$ and $\mathbf{H}^{(m)}$ through iterating. So, the following cost function is iterated to be minimized:

$$c(\mathbf{W}^{(s)},\mathbf{H}) = c_r(\mathbf{W}^{(s)},\mathbf{H}) + \lambda c_s(\mathbf{H}^{(s)}) + \mu c_t(\mathbf{H}^{(m)}), \tag{5}$$

where c_r represents the reconstruction error as non-negative least square. λ corresponds to the weight of sparseness constraint on $\mathbf{H}^{(s)}$. The sparseness constraint has been used in some audio source separation systems and in some cases the sparseness constraint improves the separation quality. The sparseness constraint can be formulated as

$$c_s(\mathbf{H}^{(s)}) = \sum_{i=1}^{r_s}\sum_{j=1}^{n} h_{i,j} / \sigma_i, \tag{6}$$

where σ_i is the standard deviation of row *j* of $\mathbf{H}^{(s)}$. r_s is the number of speech basis vectors in **W**. μ in (5) corresponds to the weight of temporal continuity criterion on $\mathbf{H}^{(m)}$. The temporal constraints can result in a better representation of music considering its harmonicity. Temporal continuity constraint of the music components is measured as following:

$$c_t(\mathbf{H}^{(m)}) = \sum_{i=1}^{r_m}\frac{1}{\sigma_i^2}\sum_{j=2}^{n}(h_{i,j} - h_{i,j-1})^2 \tag{7}$$

r_m is the number of music basis vectors in **W** accordingly.

The cost function (5) is minimized by applying multiplicative update rules to $\mathbf{W}^{(s)}$, $\mathbf{H}^{(s)}$ and $\mathbf{H}^{(m)}$ referring to the gradient algorithm proposed in [9]. The detailed algorithm will not be presented here.

After the iteration, the magnitude spectrogram of the speech can be filtered out by

$$\mathbf{S} = \mathbf{V} \otimes \frac{\mathbf{W}^{(s)}\mathbf{H}^{(s)}}{\mathbf{W}^{(s)}\mathbf{H}^{(s)} + \mathbf{W}^{(m)}\mathbf{H}^{(m)}} . \qquad (8)$$

The $\otimes$ symbol and the division above are both element-wise operation.

Finally, we can get the estimated clean speech signal by doing inverse short time fourier transform (iSTFT) to **S**.

3 Experimental Setup and Analysis

3.1 Experiment setup

The speech section of test set was uttered by 10 different speakers from ASCCD corpus, five females and five males. Each speaker had one utterance about 3 seconds long in Mandarin and the content involved news review, narration and so on. The background music fragments were from GTZAN Genre Collection. In the experiment, we used the classical and jazz genre for training and testing. For each genre, we selected 10 pieces in the collection same length with test speech signal. Then, the music fragments were artificially mixed with speech utterances in different SMR ranging from +10dB to 0dB in intervals of 5dB. Then we got the testing set, which includes 60 mixtures for different music genre and different SMR.

The music training set was completely derived from the music section of test set. Then STFT was applied to the total 30 seconds music pieces to make a cross-validation testing method. Because we used the magnitude of STFT of the training material straightly as the basis vectors, it was an exemplar-based method as in [7]. The training method in [5], which trained the specific speaker's basis vectors by doing NMF to training speech, did not works well in our method, since the random picking of basis vectors might result in large mismatch between test music and picked basis vectors. The STFT's parameters were empirically set to window size 60ms long (for the consideration of temporal continuity), window period 45ms long, and the window type was Hamming window.

In the experiments, the mixtures' spectrograms were first computed. The parameters of STFT were identical with the settings when doing STFT to the training material. Noting that in (8), the result **S** is only the estimated magnitude spectrogram of the clean speech without any information about the phase. But we can use the phase of mixed signal to synthesize the speech signal, which has been proved working well [3]. Therefore, in fact, we iterated to minimize the Euclidean distance between the magnitude spectrogram of mixed signal and **WH**. But when synthesizing clean speech signal, we made use of the phase of mixed signal.

Heuristically, the NMF dimensionality parameter *r* was set to 60 and the number of speech basis vectors is 36, resulting in a 3:2 speech to music basis vector ratio. In the experiment of proposed method, we changed the value of λ and μ in (5) to measure the performance of different combination of λ and μ. In order to evaluate the performance, we investigated signal-to-distortion ratio (SDR), source-to-interference ratio (SIR) and source-to-artifact ratio (SAR) of the estimated speech signal as the measurement of the quality of separation as in [5]. SDR measures the overall separation quality. SIR measures the suppression of undesired music signals. SAR measures degradation of speech quality by the separation. Larger values represented better result. During this procedure, measurements were carried out using the PEASS toolkit [10], a matlab toolbox for measurement of audio source separation.

3.2 Experiment analysis

All the indicators mentioned below are the average of the result of mixture signals in test set with same SMR and same music genre. The results of SIR represent the difference between the SIR of separated speech and SMR of corresponding original mixture signal.

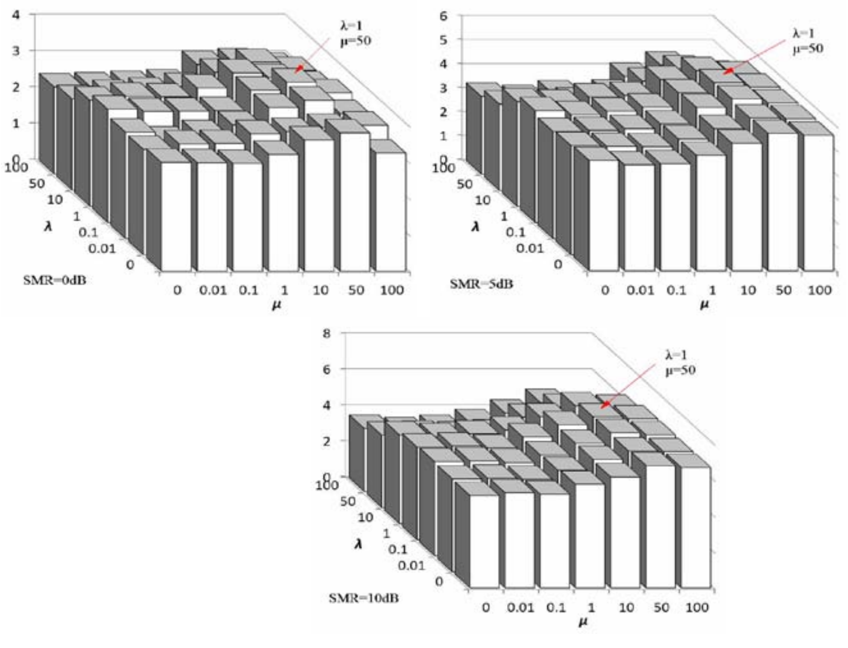

Fig. 1 The SDR result of different SMR when background music is classical. The red arrow in each figure indicates the performance of λ =1 and μ =50.

First, we evaluated the effect of changing the value of λ and μ for different SMR and different genre of background music in our proposed method. The result is shown in Fig. 1. Here we only present the SDR result of mixture signals with classical music, which measures the overall separation quality. From the result we can find explicitly that when SMR is 10dB, the SDR always increases with μ in all values of λ, including λ is 0. When μ reaches 50, the improvement is the most. In the situation that SMR equals 5dB or 0dB, when μ is 0.01 or 0.1, sometimes there is slight decrease of SDR compared with the value when μ is 0. However, after that tiny decrease, SDR begins to increase dramatically. Further, in the result of all SMRs, if μ is too high, the performance starts to decrease. This is mainly because the temporal continuity of music is overestimated. On the contrary, increase of λ does not improve the SDR obviously and when the λ is too high, the performance degrades a lot. The result proves powerful evidence that with the adding of temporal continuity constraints, the performance of separation improves a lot. In most cases, the best performance occurs when λ is 1 and μ is 50.

Second, we compared our proposed method with baseline and the separation system with only sparseness constraint on speech. In all the experiments, the parameter setting was the same, including the STFT parameter in the training phase and NMF parameter in the testing phase. The result is shown in table 1. λ in sparseness only system is set to 1. λ and μ are set to 1 and 50 for the proposed system based on the analysis in the above paragraph. From the result, we can find that the overall performance of the proposed method is superior to the baseline and sparseness only system in two different music genres. This again proves that the temporal continuity constraint we use is valid to improve the separation of speech and music.

Table 1 The performance comparison among proposed method, baseline and sparseness constraint only system in different SMR.

(a) The background music is classical.

	baseline			*sparseness only*			*proposed*		
SMR	0dB	5dB	10dB	0dB	5dB	10dB	0dB	5dB	10dB
SDR	3.0	4.5	5.1	3.1	4.7	5.1	**3.9**	**5.8**	**6.8**
SIR	1.6	3.4	4.1	2.2	3.8	4.4	3.2	4.2	5.8
SAR	15.9	17.5	18.1	15.0	17.0	17.3	15.3	17.4	18.3

(b) The background music is jazz.

	baseline			*sparseness only*			*proposed*		
SMR	0dB	5dB	10dB	0dB	5dB	10dB	0dB	5dB	10dB
SDR	2.9	4.5	5.2	2.8	4.4	5.2	**3.5**	**5.8**	**6.8**
SIR	1.2	2.3	3.4	1.2	2.0	3.8	2.0	2.9	4.6
SAR	15.7	17.4	18.1	15.0	17.0	17.3	15.6	18.0	18.7

4 Conclusions

We have shown a semi-supervised method with sparseness and temporal continuity constraints for speech and music separation, which performed better compared with baseline and sparseness only system. We studied the influence of the strength of sparseness and temporal continuity constraints on the performance. The result showed that temporal continuity criterion could improve the speech and music separation system, while sparseness constraint only improves the performance a little or not. However, in our proposed method, both the strength of sparseness and temporal continuity constraints could not be set too high in order to avoid degradation of performance.

Future work could focus on trying to apply this method to real-time application of suppression of background interference, which will be a challenging but meaningful work.

References

1. A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE transaction on signal processing*, vol. 59, no. 7, pp. 3155–3167, JULY 2011.
2. "ASCCD: Read discourse corpus with prosodic, segmental and syntactic annotation," Phonetics Lab, Institute of Linguistics, Chinese Academy of Social Sciences, Tech. Rep. [Online]. Available: http://ling.cass.cn/yuyin/english/resc6.htm.
3. B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Nonnegative matrix factorization based compensation of music for automatic speech recognition," in *Interspeech*, 2010, pp. 717–720.
4. D. D.Lee and H. Seung, "Algorithms for nonnegative matrix factorization," in *NIPS*, vol. 13, 2000, pp. 556–562.
5. F. Weninger, J. Feliu, and B. Schuller, "Supervised and semi-supervised suppression of background music in monaural speech recordings," in *ICASSP*, 2012, pp. 61–64.
6. M. Asgari, M. Fallah, E. A. Mehrizi, and A. Mostafavi, "A vq-based single-channel audio separation for music/speech mixtures," in *11th International Conference on Computer Modelling and Simulation*, 2009, pp. 223–227.
7. P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *NIPS*, 2009, pp. 1705–1713.
8. T. Hughes and T. Kristjansson, "Music models for music/speech separation," in *ICASSP*, 2012, pp. 4917–4920.
9. T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. ASLP*, vol. 15, no. 3, pp. 1066–1074, MARCH 2007.
10. V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. ASLP*, vol. 19, pp. 2046–2057, SEPTEMBER 2011.
11. Yipeng Li, Deliang Wang. "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Trans. ASLP*, vol. 15, pp. 1475-1487, 2007.
12. Y. Litvin, I. Cohen, and D. Chazan, "Monaural speech/music source separation using discrete energy separation algorithm," *Signal Processing*, vol. 90, pp. 3147–3163, 2010.