ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

A Two-step NMF Based Algorithm for Single Channel Speech Separation

Shuo Wang[1], *Wenjun Wu*[2]

[1, 2]State Key Lab of Software Development Environment,
Department of Computer Science and Engineering, Beihang University, Beijing, China
wangshuo@nlsde.buaa.edu.cn, wwj@nlsde.buaa.edu.cn

Abstract

Nonnegative Matrix Factorization (NMF) has become an increasingly popular method in the field of non-stationary speech denoising. However most NMF-based algorithms assume prior knowledge about the background noise, which is often not available in time-varying and mobile environments. In this paper, we propose a two-step NMF based speech-noise separation algorithm to address this issue. This algorithm takes the outcome of the first NMF separation as the dataset to train the basis vectors for the background noise, which will be used for the second-step NMF separation with fixed speech and noise basis vectors. Experimental results show that the proposed algorithm could achieve better results than other NMF algorithms for speech-noise separation.

Index Terms: single channel speech separation, nonnegative matrix factorization, voiced/unvoiced sound classification, Wiener filter.

10.21437/Interspeech.2012-532

1. Introduction

Single channel speech separation (SCSS) is a very challenging problem in blind source separation (BSS), as it only has one mixed signal available, which raises the difficulty of the separation. There are many methods proposed to solve this problem recently, among which nonnegative matrix factorization (NMF) is one of the most promising ones. NMF method was first introduced in [1] and widely used in image and speech processing subsequently. In [2], the NMF method was employed to separate the drum music from the mixed music, using a support vector machine (SVM) that classifies the basis vectors of the correspondent source signals. Mixed speech signal separation by sparse NMF was proposed in [3]. In [4], a NMF based algorithm was designed to decompose the mixed audio data with and without fixed trained basis vectors, where it requires human interaction for clustering the basis vectors. In [5], an extended NMF method initially decomposes the mixed data by fixed trained basis vectors for both speech and music, and then estimates the source signals through a Wiener filter, which is built as a mask with the decomposition results.

In terms of speech denoising and enhancement, especially in the scenarios such as mobile communication, NMF-based SCSS could separate the speech voice from the complex and time-varying background noise so as to improve the quality of the speech. In this paper, we propose a SCSS algorithm based on NMF, which only requires the speech voice to be trained before the separation, and then the target voice could be estimated from the mixed signal with complex background noise. This work includes three main stages: firstly, the NMF method was used to train the speech voice and to complete the first round of separation with fixed speech basis vectors. Secondly, in order to train the background noise basis vectors, the algorithm classified audio frames collected from the mixed signal into the noise dataset according to a threshold determined by the estimated result signals. Thirdly, the second round of separation of the mixed signal was performed using NMF with fixed speech basis vectors and background noise vectors, and the estimated speech signal was finally obtained through a Wiener filter composed as a soft mask by the separated signals.

The main contribution of this paper is the extension of NMF SCSS to separate the speech signal from the complex and time-varying background noise, which only requires the training data of the speaker's voice. It could work well in the chaotic noise environment where the noise signal keeps changing all the time, such as the situation of talking via a mobile phone, or in the case when the noise data are hard to capture.

The remainder of this paper is organized as follows: In section 2, a brief overview of NMF is given. We describe how to use the NMF method to train the basis vectors for the source signal and complete the first separation in section 3. In section 4, the detail of our algorithm is explained on the second-step separation of the mixed signal using determined threshold and unvoiced sound detection. Experimental results are presented in section 5 and conclusions are given in section 6.

2. Nonnegative Matrix Factorization

Non-negative matrix factorization is an algorithm that is used to decompose a nonnegative matrix X into a nonnegative basis matrix B and a nonnegative weight matrix W. Nonnegative refers to the property that each element in the matrix is nonnegative.

$$X_{n\times m} = B_{n\times r} W_{r\times m} \tag{1}$$

Matrix B and W can be obtained by minimizing the objective function:

$$\min D(X \,||\, BW) \tag{2}$$

where

$$D(X \,||\, BW) = \sum_{i,j} (X_{i,j} \log \frac{X_{i,j}}{(BW)_{i,j}} - X_{i,j} + (BW)_{i,j})$$

Equation (2) is the divergence of X from BW, which has been proved to work well in audio source separation in [6]. The NMF solution for equation (2) can be computed by alternating updates of B and W as follows:

$$B \leftarrow B \cdot (\frac{X}{BW} W^T \Big/ 1W^T), \tag{3}$$

$$W \leftarrow W \cdot (B^T \frac{X}{BW} \Big/ B^T 1), \tag{4}$$

where $\mathbf{1}$ is a matrix of ones with the same size of X, the operation $\cdot$ and all divisions are element-wise multiplications and divisions respectively.

3. Training the bases and signal separation

3.1. Training the basis vectors

Given a set of training data for voice signals of a single speaker, the short time Fourier transform (STFT) is computed for getting the magnitude spectrogram of the signal. Then the NMF algorithm decomposes the magnitude spectrogram matrix to basis vector matrix *B* and weight matrix *W*.

$$S_{n\times m} = B_{n\times r} W_{r\times m} \tag{5}$$

Where $S_{\mathrm{n\times m}}$ is the magnitude spectrogram of the training speech signal, *m* is the number of data frames, *n* is the size of each frame and *r* is the number of the basis vectors. The updated versions in equations (3) and (4) are used to solve equation (5). At the end of the iteration, the obtained matrix *B* will be the trained speech basis vectors identified as B_{speech}.

3.2. Mixed signal separation

NMF method is used again here to separate the magnitude spectrogram matrix *X* of the mixed signal with fixed speech basis vectors obtained through the training process.

$$X \approx BW = \begin{bmatrix} B_{speech} B_{noise} \end{bmatrix} \begin{bmatrix} W_{speech} \\ W_{noise} \end{bmatrix} \tag{6}$$

Where B_{speech} is the trained basis vector of the speech and B_{noise} represents the basis vectors for background noise. W_{speech} and W_{noise} are the weights vectors of estimated speech and noise respectively. Here B_{speech} is fixed while the update role in equation (3) and (4) are used to solve equation (6), the noise basis vectors B_{noise} and the weights matrix W get updated during the iteration. At the end of the iteration, the spectrogram of the estimated speech signal and noise signal are respectively calculated by multiplying the bases matrix with its corresponding weights matrix.

$$S = B_{speech} W_{speech} \tag{7}$$

$$N = B_{noise} W_{noise} \tag{8}$$

S and *N* represent the magnitude spectrogram of the estimated speech signal and noise signal respectively after the first NMF separation.

4. Second NMF separation

Generally speaking, the estimated speech signal can be obtained by *S* directly, but the result of the target voice is not always satisfactory. The defect of the algorithm mentioned above is due to using the trained speech basis vectors exclusively in the decomposition while missing the noise prior knowledge because of its complexity or unavailability.

To solve the problem mentioned above, this paper proposed a two-step NMF solution. After the first NMF separation process, the decomposition results are subsequently used to determine a threshold which helps to collect the background noise as a data set. Then, based on this noise dataset, we can train the noise basis vectors, and perform the second round of NMF separation process with the fixed speech and noise basis vectors. This method can further reduce the interference of ambient noise on the target speech and to some extent alleviate the decomposition error of the separation results for the first time.

4.1. Threshold calculation

In the previous section, NMF was used to obtain the magnitude spectrogram *S* for the estimated speech signal and *N* for noise. The short-term energy (STE) ratio of the estimated speech and noise is calculated in the unit of frames. If the resulting value is very small, it means that the energy of the target speech occupies relatively lower portion in the mixed signal, and it may be categorized as a noise frame with higher possibility. On the contrary, a large result value suggests the opposite situation. Thus, a threshold could be determined by the ratio to classify the mixed signal frame into a noise signal data set. Based on these collected noise data, we could train the noise basis vectors for the second NMF separation.

$$STE_t = \frac{1}{N}\sum_{k=1}^{N} |X_t(k)|^2 \tag{9}$$

$$R_t = STE_t^S / STE_t^N = \sum_{k=1}^{N} |S_t(k)|^2 \Big/ \sum_{k=1}^{N} |N_t(k)|^2 \tag{10}$$

Where *t* represents the frame index and *k* is the frequency-index, STE_t^s and STE_t^n represent the STE of speech and noise respectively, R_t is the short-term energy ratio of frame *t*.

In order to trade off the impact of the maximum and minimum ratio values on the threshold selection, logarithm is used for the ratio and takes the mean value of the results as the threshold:

$$Threshold = \frac{1}{T}\sum_{t=1}^{T} \log_{10} R_t \tag{11}$$

where T is the total number of frames of the mixed signal.

4.2. Unvoiced sound detection

Frame *t* of the mixed signal could be collected exactly as the background noise training data if $log_{10}R_t < Threshold$ is satisfied. This assumption is based on the fact that most of the results generated by the first NMF separation are correct. However, sometimes the target speech signal may be wrongly classified as the background noise, which could be mingled into the noise training set. Thus, in order to clean the collected noise dataset, we introduce a new filtering mechanism based on the different characteristics of voiced sound and the background noise that can be considered as unvoiced sound. Therefore, we can screen the candidate noise dataset and drop these frames identified as voiced sound data to avoid the interference of the mistakenly introduced speech voice for the second NMF decomposition.

The loss function to determine frame *t* as an unvoiced sound is:

$$L_t^0 = \max[R_t(k)] / E_t \tag{12}$$

where

$$R_t(k) = \sum_{n=0}^{N-1} m_t(n) m_t[\mathrm{mod}(n+k), N], 0 < k < N$$

$$E_t = \sum_{n=0}^{N-1} m_t(n)^2$$

$R_t(k)$ is the autocorrelation function, $\max R_t(k)$ means the maximum magnitude, m_t is the noise training data collected based on the threshold, frame size is *N*, E_t represents the energy of frame *t*. As $\max R_t(k)$ could better reflect the periodicity of the speech frame, it is chosen as the judgment of the voiced level. The larger value of L_t^0 suggests the less likelihood for the frame *t* to be identified as unvoiced sound or noise, and vice versa.

The loss function to mark frame *t* as the voiced sound is:

$$L_t^1 = \begin{cases} 1 & L_t^0 \le 0.5 \\ 1 - L_t^0[1 - \frac{\arg(\max[R_t(k)])}{N}], & L_t^0 > 0.5 \end{cases} \quad (13)$$

where $\arg(\max[R_t(k)])$ is the index corresponding to the maximum magnitude. If $L_t^0 \le 0.5$, the maximum value of $R_t(k)$ is too small so that the fundamental period may not exist, so the current frame could be considered as an unvoiced sound where $L_t^1 > L_t^0$. Otherwise, the current frame may be a voiced sound or noise, and the index of $\max R_t(k)$ should be considered at the same time. The larger value of L_t^0 and little index value makes the smaller value of L_t^1, which indicates higher probability for the signal in frame *t* to be judged as voiced sound. If the result of $\arg(\max[R_t(k)])/\mathrm{N}$ is relatively large, it may be caused by noise, and equation (13) may balance the loss and make $L_t^1 > L_t^0$, which helps to judge the frame as noise [7].

4.3. Second NMF separation

After the classification of voiced and unvoiced signals, the unvoiced sounds were included as the data set for training the basis vectors of the background noise accompanied in the mixed signal. Then, the second time NMF decomposition for the mixed signal could be implemented with fixed basis vectors both for speech and noise.

$$X \approx \begin{bmatrix} B_{speech} & B_{noise} \end{bmatrix} W \quad (14)$$

Where B_{speech} and B_{noise} are the basis vectors for speech and noise respectively. Here only the update rule in equation (4) is used to solve (14), and the base matrix is fixed. The magnitude spectrogram of speech and noise could be obtained through equation (7) and (8). They are respectively identified as $\tilde{S}$ and $\tilde{N}$ in order to distinguish from the results of the first separation. Finally, the Wiener filter is used as a soft mask, which is built by $\tilde{S}$ and $\tilde{N}$, to eliminate the approximation error caused during the iteration process with a fixed base matrix, and the estimated time domain speech signal can be retrieved through IFFT.

5. Experiments and Results

The algorithm proposed in this paper was simulated on a collection of speech and noise data at a 8kHz sampling rate. For training speech data, 56 seconds of voice from a single speaker were used and another 10 seconds data for testing. Some background noise data were downloaded online including sirens, street noise, babble, wind and white noise. Some mixed signals are a linear combination of speech and noise data, the others are captured in a real scenario, in which the speaker was standing next to a noisy street with a lot of traffic. The magnitude spectrograms for training data were calculated through the STFT, the frame size was set to 512 and NMF was run for 100 iterations.

Table 1 shows the accuracy of the determined threshold for noise data collection. The selection radius means that the current frame will be selected if and only if all the other frames within the radius are under the threshold. The larger radius can result in higher accuracy while narrowing the range of options of the noise signal. In order to achieve a balance, the radius was chosen to be 2, which could lead to relatively good separation results for our data set.

Table 1. *The accuracy of the noise data collection based on the determined threshold.*

Selection Radius	Noise frames	Total frames collected	Accuracy
1	712	803	88.67%
2	536	585	91.62%
3	475	511	92.95%
4	427	449	95.10%

Table 2. *The accuracy of the unvoiced sound classification.*

Selection Radius	Unvoiced noise frames	Unvoiced frames	Accuracy
1	581	598	97.16%
2	427	436	97.94%
3	409	415	98.55%
4	373	378	98.68%

Table 3. *Quality measures in dB for the speech separation with different NMF methods*

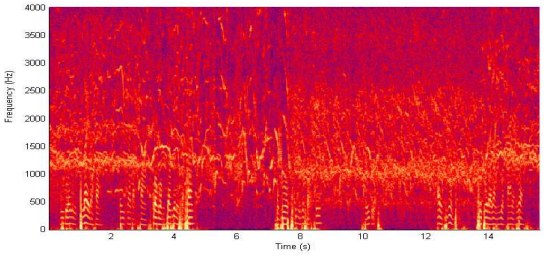
Quality Measures	openBliSSART toolkit	Our Method
SDR	5.25	16.14
SIR	25.86	27.01
SAR	5.29	16.53

Table 2 shows that the accuracy of the classification for voiced/unvoiced sound is high enough to produce clean noise data for achieving better results in the second NMF separation with fixed basis vectors both for speech and noise.

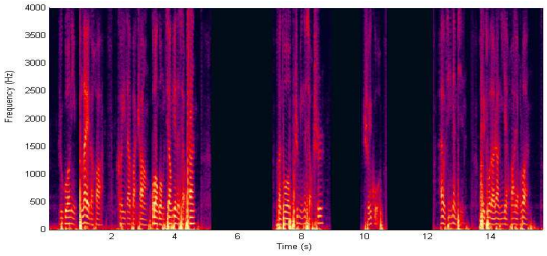
The tests of the algorithm proposed in this paper are compared with openBliSSART[8], a toolkit for blind source separation in audio recognition tasks. The separation results of openBliSSART toolkit were obtained by using the supervised NMF method integrating prior knowledge into the separation process by training both the speech and noise data in advance.

Table 3 illustrates the separation performance of using NMF with the openBliSSART toolkit and the algorithm proposed in this paper. Measures like signal-to-distortion ratio (SDR), the signal-to-interference ratio (SIR), and the signal-to-artifacts ratio (SAR) defined in [9], are used for the evaluation of separation quality. SDR determines the overall sound quality of the estimated signal, SIR measures the interference of other sources in the target sound source and SAR calculates the artifacts present in the separated signal [10]. All three quality measures of our method outperform those of openBliSSART, which indicates better separation performance of our algorithm.

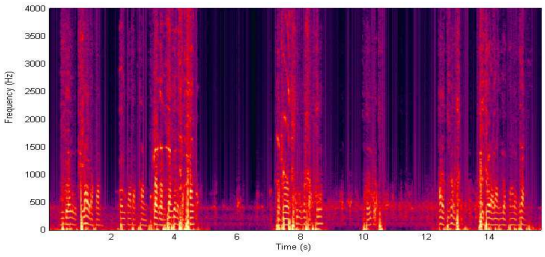
Fig. 1 shows an intuitive separation result of a mixed signal. As we can see from the figure, the proposed algorithm successfully suppresses the background noise from the mixed signal, and yields a better approximation of the speech signal than only taking the NMF method with fixed speech basis vectors. The openBliSSART toolkit can also obtain a good separation result, but the blurred spectrogram reduces the intelligibility of the speech signal. There are still some distortions in the estimated speech signal, especially at low frequencies which may be caused by the similarity of the speech and noise signals.



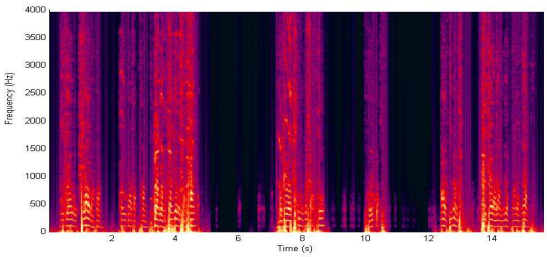
(a) The spectrogram of the mixed signal



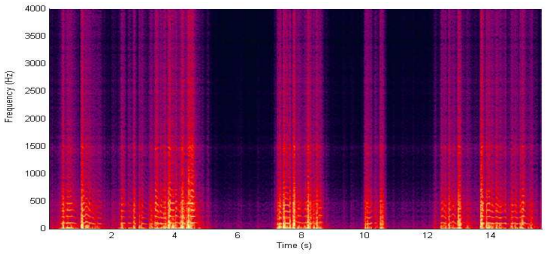
(b) The spectrogram of the original speech signal



(c) The spectrogram of the estimated speech signal after the first NMF separation



(d) The spectrogram of the estimated speech signal after the second NMF separation



(e) The spectrogram of the estimated speech using openBliSSART toolkit

Figure 1: *The spectrograms of the mixed signal, the original speech signal, the estimated speech signal after first NMF separation, the estimated speech signal after second NMF separation and the estimated speech signal using openBliSSART toolkit. The mixed signal is a linear combination of the speech signal of a speaker and babble.*

6. Conclusion

In this paper, we proposed a new algorithm to separate speech and background noise signals based on the NMF method with Wiener filter. In the case when only speech training is available for NMF decomposition, the noise training data were collected automatically depending on the determined threshold and further unvoiced sound detection. The proposed algorithm gives better results than the original one for further suppressing the background noise from the mixed signal and obtaining a purer speech signal.

7. Acknowledgements

This work was supported by the State Key Laboratory of Soft ware Development Environment Funding No.SKLSDE-2012ZX-19.

8. References

[1] D. Lee and H. S. Seung, "Algorithms for Non-negative Matrix Factorization," *Advances in Neural Information Processing System*, 2000, pp.556-562.

[2] M. Heln and T. Virtanen, "Separation of drums from polyphonic music using nonnegative matrix factorization and support vector machine," *in Proc. Eur. Signal Process. Conf.*, Istanbul, Turkey, 2005

[3] Mikkel N. Schmidt and Rasmus K. Olsson, "Single channel speech separation using sparse non-negative matrix factorization," *in International Conference on Spoken Language Processing (INTERSPEECH)*, 2006.

[4] B. Wang and M. D. Plumbley, "Investigating single channel audio source separation methods based on nonnegative matrix factorization," *in Proceedings of the ICA Research Network International Workshop*, 2006.

[5] Emad M. Grais and Hakan Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," *International Conference on Digital Signal Processing (DSP)*, 2011.

[6] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," *in Proc. of ICASSP*, 2008.

[7] LI Hao and TANG Chao-jing, "Voiced Sound Endpoint Detection Based on Circular Autocorrelation Function," *Trans. Computer Engineering*, Vol. 37, No.22, Nov. 2011.

[8] F. Weninger, A. Lehmann and B. Schuller, "OpenBliSSART: design and evaluation of a research toolkit for blind source separation in audio recognition tasks," *in Proc. of ICASSP*, 2011.

[9] E. Vincent, R.Gribonval, and C. Fevotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Audio, Speech and Language Processing*, 2006.

[10] R. Jaiswal, D. FitzGerald, D. Barry, E. Coyle and S. Rickard, "Clustering NMF basis functions using shifted NMF for monaural sound source separation," *in Proc. of ICASSP*, 2011