INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2021.15.16 Volume 15, 2021

Music Signal Separation Using Supervised Robust Non-Negative Matrix Factorization with β-divergence

Feng Li, Hao Chang
Department of Computer Science and Technology,
Anhui University of Finance and Economics,
Caoshan Road, Bengbu 233030, China

Received: May 29, 2020. Revised: January 30, 2021. Accepted: February 16, 2021. Published: February 22, 2021.

Abstract—We propose a supervised method based on robust non-negative matrix factorization (RNMF) for music signal separation with β-divergence called supervised robust non-negative matrix factorization (SRNMF). Although RNMF method is an effective method for separating music signals, its separation performance degrades due to has no prior knowledge. To address this problem, in this paper, we develop SRNMF that unifying the robustness of RNMF and the prior knowledge to improve such separation performance on instrumental sound signals (e.g., piano, oboe and trombone). Application to the observed instrumental sound signals is an effective strategy by extracting the spectral bases of training sequences by using RNMF. In addition, β-divergence based on SRNMF be extended. The results obtained from our experiments on instrumental sound signals are promising for music signal separation. The proposed method achieves better separation performance than the conventional methods.

Keywords—Music signal separation; Robust; Non-negative matrix factorization (RNMF); β-divergence.

I. Introduction

In In recent years, music signal separation methods have attracted considerable interest and been intensively studied [1-3], There are many real-life applications of audio signal processing in the different fields [4-6]. However, these methods have a critical problem regarding separation performance in which several signals are mixed together and the objective is to recover the original clean signal from the mixture. Many methods have been proposed with the goal of overcoming the difficulty in separation including independent component analysis (ICA) [7], sparse decomposition [8], robust principal component analysis (RPCA) [9] and weighted RPCA (WRPCA) [10]. Non-negative matrix factorization (NMF) [11], which is a type of sparse representation method, has shown impressive results in source separation. The methods of source separation for audio signals based on NMF can be roughly categorized into two types according to whether they require prior knowledge, namely, unsupervised methods and supervised methods. The former can separate the source signals directly whereas the latter require prior knowledge to separate such signals.

Since unsupervised methods are used to attempt to separate without using any prior knowledge, they are particularly useful in separating unknown sources. NMF [11] is a very typical example of such methods and proposed by Lee and Seung who decomposed a non-negative matrix into a non-negative basis matrix and a non-negative activation matrix using multiplicative update rules by minimizing a cost function. Although NMF has been proven to be a useful tool in source separation, one drawback is that the separation performance tends to be poor in the case of noise. Robust non-negative matrix factorization (RNMF) [12] can be used to improve the robustness of NMF, which decomposed the non-negative matrix as the summation of the product of two non-negative matrices and one sparse error matrix. However, RNMF incurs a risk of degrading the separation performance in audio signals owning to the lack of prior knowledge.

In contrast to unsupervised methods, supervised methods are particularly noteworthy in that they can obtain better separation performance with prior knowledge. This is because we can use additive and useful training sequences in advance during separation processing from the mixed sound signals. Supervised non-negative matrix factorization (SNMF) [13] is an example of such methods and has attracted much attention in recent years. The SNMF separates the target sound signals using a prior

training signal for source separation, which is provided better separation results than NMF. Nevertheless, the critical problem is that separation accuracy degrades due to the simultaneous generation of similar spectral patterns between the trained basis and target sound signals. Pablo et.al [2] proposed a new supervised NMF method that can improve the separation performance of music signals, which used the deformation with an all-pole model of a spectral supervision basis trained. Although this method has made great progress in source separation under certain conditions, the disadvantage is that the separation results become poor when noise existing. This implies that we need a more robust method for separating source signals.

As stated above, unsupervised methods (e.g., RNMF) and supervised methods (e.g., SNMF) have their own advantages and disadvantages in source separation tasks. To address these disadvantages and combine the advantages of them, in this paper, we propose a supervised method called supervised RNMF (SRNMF) for unifying the robustness of RNMF and the advantages of supervised methods to separate source signals.

The remainder of this paper is organized as follows. In Section II, we review conventional methods, e.g., non-negative matrix factorization (NMF), supervised non-negative matrix factorization (SNMF) and robust non-negative matrix factorization (RNMF), respectively. In Section III, we describe the proposed method supervised robust non-negative matrix factorization (SRNMF) with β-divergence. Experiments on instrumental sound signals are conducted in Section IV. And finally, we draw conclusions in Section V.

II. RELATED WORK

In this section, we discuss the extension methods of NMF-based in the context of source separation.

A. Non-negative matrix factorization

NMF [11] [14] is a type of sparse representation method for source separation of music signals and has also exhibited separation performance improvement in recent years. The NMF method for acoustical signals decompose an input spectrogram into a product of a spectral basis matrix and its activation matrix. The following equation represents the decomposition model of NMF

$$V \approx WH, \tag{1}$$

where $V(V \in R_{m\times n})$ is an observed non-negative matrix that represents an amplitude spectrogram of sound source signals, $W(W \in R_{m\times k})$ is a non-negative basis matrix of a sound signal as column vectors, $H(H \in R_{k\times n})$ is a non-negative activation matrix that corresponds to the activation of each basis vector of W, m and n are the rows and columns of observed sound signals, respectively. And k is the number of supervised signal basis vectors. Usually, we choose $m\times k + k\times n \prec m\times n$; hence reducing the dimensions of the input data.

In NMF, the multiplicative update rules for W and H have been derived to minimize each of the three divergences and without the need for constraints to enforce non-negativity. In order to reduce dimension, commonly, set to a small number, which results in NMF being a low-rank matrix approximation method. Therefore, the multiplicative update rules are derived as follows for the Euclidean distance (EUC),

$$W \leftarrow W \otimes \frac{VH^T}{WHH^T}, \quad H \leftarrow H \otimes \frac{W^T V}{W^T WH}, \tag{2}$$

Kullback-Leibler divergence (KL),

$$W \leftarrow W \otimes \frac{\frac{V}{WH}H^T}{1H^T}, \quad H \leftarrow H \otimes \frac{W^T\frac{V}{WH}}{W^T 1}, \tag{3}$$

and Itakura-Saito divergence (IS)

$$W \leftarrow W \otimes \frac{\frac{V}{(WH)^2}H^T}{\frac{1}{WH}H^T}, \quad H \leftarrow H \otimes \frac{W^T\frac{V}{(WH)^2}}{W^T\frac{1}{WH}}, \tag{4}$$

note that the operator $\otimes$ denotes element-wise multiplication of two matrices (Hadamard product), $\frac{V}{WH}$ denotes element-wise division, $(WH)^2$ denotes element-wise exponentiation, and 1 denotes a matrix of ones of appropriate dimension. NMF plays a vital role in audio source separation, but the disadvantage is that the separation performance tends to be poor in the case of noise.

B. Supervised non-negative matrix factorization

Supervised non-negative matrix factorization (SNMF) [12] is developed from NMF, which contains two processes: prior knowledge training and observed signal separating. For the prior knowledge training process, it requires sample sounds that should be trained to achieve signal separation from the sound signals. For the observed signal separating process, the observed sound signals are separated using the prior knowledge of the training process.

In SNMF, the training sound signals are required in advance, but separation performance is better than NMF from the observed sound signals with prior knowledge. The decomposition of SNMF with trained supervised sound signals can be expressed as

$$V \approx WH + FQ \tag{5}$$

where $V(V \in R_{m\times n})$ is an observed non-negative matrix, $W(W \in R_{m\times k})$ is a non-negative basis matrix trained in advance, $H(H \in R_{k\times n})$ is an activation matrix that corresponds to the observed matrix V. $F(F \in R_{m\times k})$ is the residual spectral pattern matrix $Q(Q \in R_{m\times k})$ is an activation matrix that corresponds to F. The notations H and F are non-negative

matrices, m and n are the rows and columns of observed sound signals, respectively. And k is the number of supervised signal basis vectors. With SNMF, the matrices H, F and Q are optimized under the condition that W is known in advance from the prior training process. Therefore, WH represents the target training instrumental sound signals, and FQ represents the observed instrumental sound signals from the sound signal data. The SNMF method can extracts the target sound signals, particularly in the case of a small number of source signals. Although SNMF can obtains better separation results than NMF, the separation accuracy degrades owning to the simultaneous generation of similar spectral patterns between basis and target sound signal.

C. Robust non-negative matrix factorization

Robust non-negative matrix factorization (RNMF) [12] is an extension of NMF and effective for source separation. Assuming entries of a data matrix may be arbitrarily corrupted, but the corruption is sparse; therefore, the decomposition model is expressed as

$$V \approx WH + E, \tag{6}$$

where $V(V \in R_{m\times n})$ is a given non-negative matrix, $W(W \in R_{m\times k})$ is a non-negative basis matrix, $H(H \in R_{k\times n})$ is a non-negative activation matrix. $E(E \in R_{m\times n})$ is the residue or noise non-negative matrix representing the approximation error, m and n are the rows and columns of observed sound signals, respectively. And k is the number of supervised signal basis vectors.

From the above decomposition model of source separation, we can see when E is 0, the RNMF method is the conventional form of NMF. Optimal W, H and E can be obtained by minimizing the approximation error. Therefore, we need to define a new cost function for RNMF to separate the target sound signals. RNMF can improves the robustness of NMF; nevertheless, the separation performance degrades due to the lack of prior knowledge.

III. Proposed Method

In this section, we explain the proposed method and its application by the multiplicative update rules with β-divergence for instrumental sound signals separation. In addition, we give an example of spectrograms from the observed instrumental sound signals (e.g., the mixture of oboe and piano).

A. Supervised robust non-negative matrix factorization

Supervised robust non-negative matrix factorization (SRNMF), which is a supervised method based on RNMF method. We can firstly obtain the basis matrices of instrumental sound signals (e.g., oboe and piano) using RNMF in advance. Then, use the obtained prior knowledge in the training processing to separate the instrumental signals from the observed mixture of sound signals (e.g., mixture of oboe and piano). The separation model of SRNMF can be expressed with the supervised trained sound signals as

$$V \approx W_1H_1 + W_2H_2 + E, \tag{7}$$

where $V(V \in R_{m\times n})$ is an observed non-negative matrix of instrumental sound signals, $W_1(W_1 \in R_{m\times k})$ and $W_2(W_2 \in R_{m\times k})$ are the prior knowledge non-negative basis matrices, which include spectral patterns of the target signals as column vectors, $H_1(H_1 \in R_{k\times n})$ and $H_2(H_2 \in R_{k\times n})$ are non-negative activation matrices by W_1 and W_2, and $E(E \in R_{m\times n})$ is the residue or noise non-negative matrix representing the approximation error, m and n are the rows and columns of observed sound signals, respectively. And k is the number of supervised sound signal basis vectors.

B. β-divergence

The β-divergence [15] [16] is a family of cost functions parameterized by a signal shape parameter β and can be defined as

$$D_\beta(y \mid x) = \begin{cases} \dfrac{y^\beta + (\beta-1)x^\beta - \beta y x^{\beta-1}}{\beta(\beta-1)}, \beta \in \Re \setminus \{0,1\} \\ \dfrac{y}{x} - \log\dfrac{y}{x} - 1, (\beta = 0) \\ y\log\dfrac{y}{x} + x - y. (\beta = 1) \end{cases} \tag{8}$$

Generally, the cost functions in NMF can be calculated by the following three distances: Itakura-Saito divergence (IS: β = 0), Kullback-Leibler divergence (KL: β = 1), and Euclidean distance (EUC: β = 2). The corresponding formulas are given as

$$D_\beta(y \mid x) = \begin{cases} \dfrac{y}{x} - \log\dfrac{y}{x} - 1, (\beta = 0) \\ y\log\dfrac{y}{x} + x - y, (\beta = 1) \\ \dfrac{1}{2}(y-x)^2. (\beta = 2) \end{cases} \tag{9}$$

C. Multiplicative update rules

The multiplicative update rules for SRNMF are derived in a similar manner to the original NMF update rules [11] [14], we can then obtain the following rules (13), (14), and (15), respectively.

$$W \leftarrow W \otimes \frac{V(WH+E)^{(\beta-2)}H^T}{(WH+E)^{(\beta-1)}H^T} \tag{10}$$

$$H \leftarrow H \otimes \frac{W^T V(WH+E)^{(\beta-2)}}{W^T (WH+E)^{(\beta-1)}} \quad (11)$$

$$E \leftarrow E \otimes \frac{V(WH+E)^{(\beta-2)}}{(WH+E)^{(\beta-1)}+I} \quad (12)$$

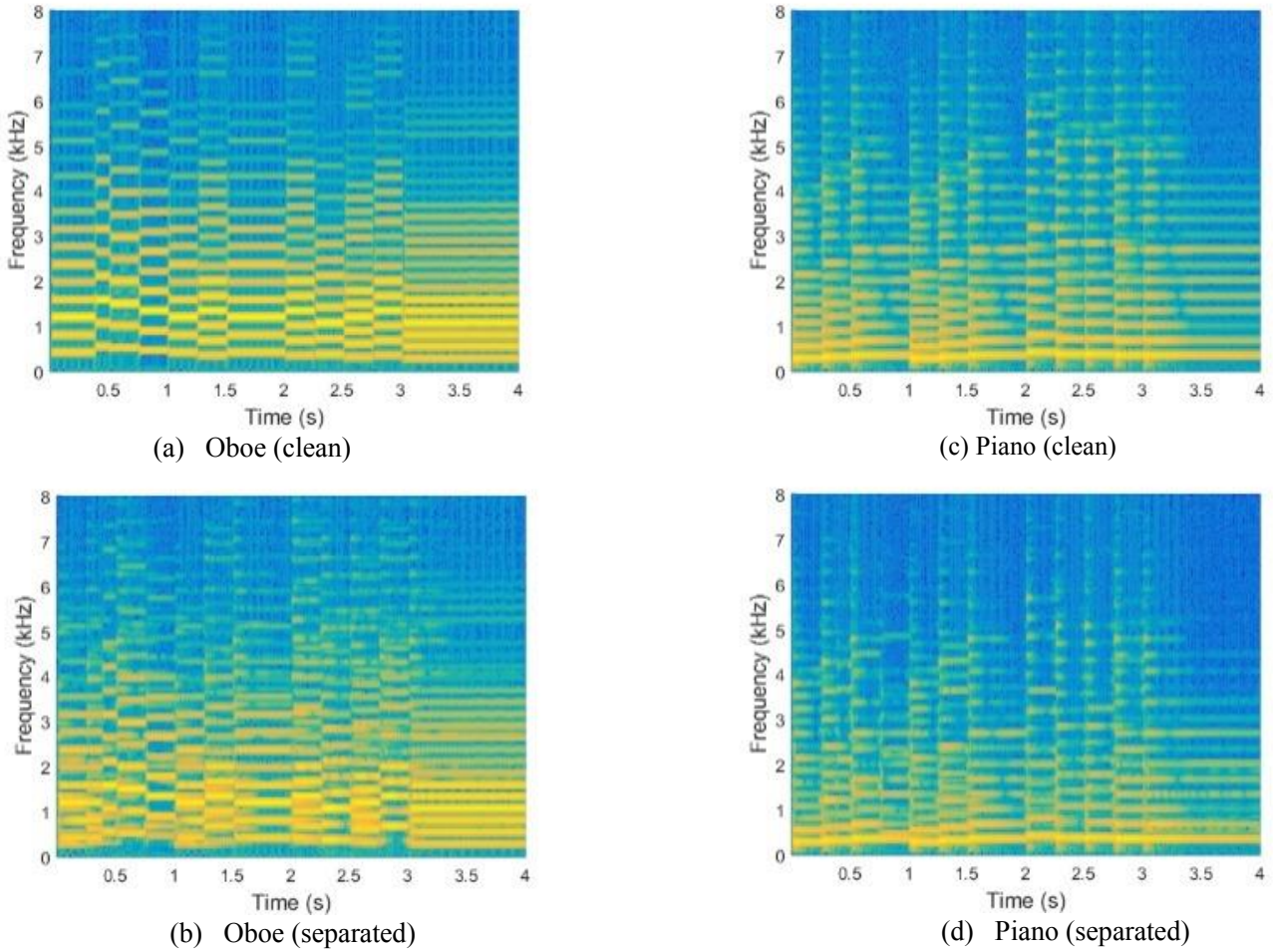
where W, H and E are all non-negative matrices. Note that all multiplications and divisions are carried out in an element-wise manner. The operator $\otimes$ denotes element-wise multiplication of two matrices (Hadamard product). The multiplicative update rules are easily implemented by alternating update rules, and there are not need to do any interference during the process of separating target sound signals.

D.Mask estimation

After obtaining the update rules by RNMF, the estimated spectrograms W_1H_1 and W_2H_2 are used to compute soft masking M_1 (e.g., oboe) and M_2 (e.g., piano) due to it can provides less artifacts in the resynthesize while increases the amount of interference among of them. The mask estimation M_1 and M_2 can be defined as

$$M_1 = \frac{W_1H_1}{W_1H_1 + W_2H_2} \quad (13)$$

$$M_2 = \frac{W_2H_2}{W_1H_1 + W_2H_2} \quad (14)$$



(a) Oboe (clean) (c) Piano (clean)

(b) Oboe (separated) (d) Piano (separated)

Figure 1. Spectrograms of instrumental sound signals (oboe and piano): (a) and (c) are original sources. In contrast, (b) and (d) are separated by using SRNMF from the mixture signals.

In our work, we separate the observed mixture of instrumental sound signals using multiplicative update rules with different values of β. And set β = 0, 1, and 2, respectively. Figure 1 is the spectrograms of instrumental sound signals by

using SRNMF. (a) and (c) are the original of clean instrumental sound signals (oboe and piano), after the separation of the oboe instrumental sound signal from the mixture by using SRNMF with KL-divergence (β = 1), the corresponding results of separated instrumental sound signals are (b) and (d), respectively.

IV. EXPERIMENTAL EVALUATION

In this section, we conduct experiments to evaluate our SRNMF method with the different values of β and compare it with the conventional methods based on the performance of separating instrumental sound signals (e.g., piano, oboe, and trombone).

A. Experiment conditions

We evaluate our proposed method using the three instrumental sound data, a piano (Pf), oboe (Ob), and trombone (Tb). The three melodies depicted in Figure 2 are created using Microsoft GS Wavetable SW Synth software (as artificial MIDI sounds). In order to separate one instrumental sound signal from a mixture of two instrumental sound signals in our experiments. All instrumental data are monaural and sampled at 44.1 kHz. We set k to 30. And the experiments are run for 1000 iterations. The input feature we used is calculated using STFT (short-time Fourier transform) and ISTFT (inverse STFT) with 1024-points window size and a hop size is 512-points. In our experiments, we firstly separate the instrumental sound signals using RNMF method to obtain the pre-trained non-negative basis matrices W_1 and W_2, then use the prior knowledge to separate the observed mixture of sound signals. And finally, we can obtain the target instrument sound signals.

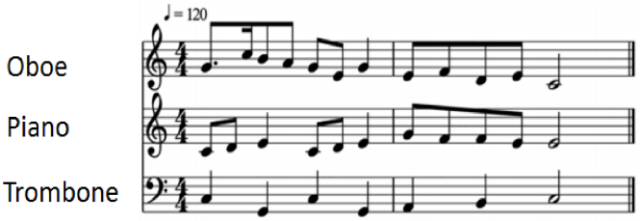


Figure 2. Scores of each instrument.

To confirm the effectiveness of the proposed SRNMF method, the quality of separation is assessed in terms of source-to-distortion ratio (SDR), source-to-artifact ratio (SAR), and source-to-interference ratio (SIR) by using the BSS-EVAL 3.0 metrics [17] and the normalized of SDR (NSDR). The estimated signal $S(t)$ is defined as

$$S(t) = S_{target}(t) + S_{interf}(t) + S_{artif}(t). \quad (15)$$

where $S_{target}(t)$ is the allowable deformation of the target sound, $S_{interf}(t)$ is the allowable deformation of the sources that account for the interferences of the undesired sources, and

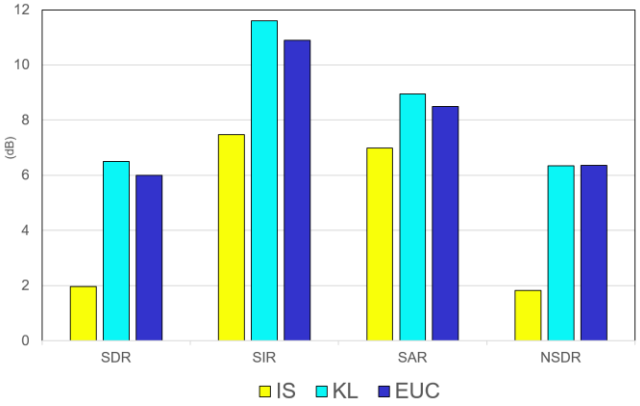


Figure 3. Experimental results regarding SDR, SIR, SAR, and NSDR for instrumental sound signals separation by using SRNMF with β-divergence (β = 0, 2, and 1).

$S_{artif}(t)$ is an artifact term that may correspond to the artifact of the separation method. The formulas for SDR, SIR, and SAR are defined as

$$SDR = 10\log_{10}\frac{\sum_t S_{target}(t)^2}{\sum_t \{e_{interf}(t) + e_{artif}(t)\}^2}. \quad (16)$$

$$SIR = 10\log_{10}\frac{\sum_t S_{target}(t)^2}{\sum_t e_{interf}(t)^2}. \quad (17)$$

$$SAR = 10\log_{10}\frac{\sum_t \{S_{target}(t) + e_{interf}(t)\}^2}{\sum_t e_{artif}(t)^2}. \quad (18)$$

The higher values of SDR, SIR and SAR represent the method that exhibits better separation performance of source separation. The SDR represents the quality of the separate target sound signals, SAR represents the absence of artificial distortion, and SIR represents the degree of separation between the target and other sound signals. In addition, the NSDR is the normalized SDR can be defined as

$$NSDR(u, v, x) = SDR(u, v) - SDR(x, v). \quad (19)$$

where u is the resynthesized instrumental sound signals, v is the original clean signal, and x is the mixture of two instrumental sound signals (e.g., the mixture of piano and oboe). The NSDR is used to estimate the improvement in the SDR between x and u. All the metrics are expressed in dB.

B. Experiments results

In our experiments, we firstly evaluate SRNMF method with different values of β on the three instrumental sound signals. Figure 3 shows the experiment results by using SRNMF method based on β-divergence. From the experiment results, we can see that KL (β = 1) is better than IS (β = 0) and EUC (β = 2) regarding SDR, SIR, SAR, and NSDR. However, the IS (β = 0) is very poor results in the performance of separation for all four evaluation standards.

Additionally, we compare our proposed method with SNMF and RNMF. And also compare with the different values of β. Because SDR indicates the total evaluation criteria of separation performance that involves SIR and SAR, we compare the proposed method based on SDR. Table 1 lists the results of SDR based on SRNMF method with the different values of β. We extract the target instrumental sound signal (the first of two mixed sounds) from each combination of the instrumental sound signals. The first is the target instrumental sound signal and the second is the non-target instrumental sound signal as shown in Table 1. The IS, EUC, and KL are the SRNMF method with Itakura-Saito divergence, Euclidean distance, and KL-divergence (β = 0, 2, and 1), respectively. From the experimental results in Table 1, we can confirm that RNMF obtains poorly, while the SRNMF method performs well regarding separation performance on the instrumental sound signals. Moreover, the KL-divergence can obtain best results than Euclidean distance and Itakura-Saito divergence on instrumental sound signals separation task. However, we also can see that the separation performance of Itakura-Saito divergence (IS) is not as good as that of Euclidean distance (EUC) and KL-divergence (KL) as shown in Table I, particularly for the mixture of instrumental sound signals of piano and oboe.

V. Conclusion

In this paper, we proposed a supervised method called SRNMF for music signal separation from monaural audio recordings. In addition, we discussed the different values of β for extracting instrumental sound signals. Experimental results show clearly that the proposed method outperforms the conventional methods on instrumental sound signals separation, especially for the KL-divergence.

Conflict of Interest

The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

Feng Li and Hao Chang conceived and designed the experiments; Feng Li implemented the models, performed the experiments, analyzed the experiment data and wrote the paper; Hao Chang fine-tuned the paper and gave some precious advances.

Acknowledgment

This work was supported in part by the Natural Science Foundation of the Higher Education Institutions of Anhui Province under grant No. KJ2020A0011, the Science Research Project of Anhui University of Finance and Economics under grant No. ACKYB20012, the Natural Science Foundation of China under grants No. 61704001, Anhui Provincial Natural Science Foundation under grant No.1808085QF196.

References

[1] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," in Proc. ISMIR, pp. 375-378, 2007.

[2] P. Sprechmann, A. M. Bronstein, G. Sapiro, "Supervised non-negative matrix factorization for audio source separation," Excursions in Harmonic Analysis, Volume 4. Birkhäuser, Cham, 2015, pp. 407-420.

[3] E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley, and F.R. Stoter, "Musical source ¨separation: An introduction," IEEE Signal Processing Magazine, vol. 36, no. 1, 2019, pp.31-40.

[4] M. Zabcikova, Z. Koudelkova, R. Jasek, "Examining the Efficiency of Emotiv Insight Headset by Measuring Different Stimuli," WSEAS Transactions on Applied and Theoretical Mechanics, Volume 14, 2019,, pp. 235-242.

[5] H. Bagheri, M. Sajjadi, R. Chimeraad, "Empirical investigation of noise reduction filter for a flow-based spirometer accuracy improvement," Engineering World, Vole 1, 2019, pp. 58-63.

[6] J. Glover, V. Lazzarini and J. Timoney, "Real-time detection of musical onsets with linear prediction and sinusoidal modeling," EURASIP Journal on Advances in Signal Processing, Volume 68, 2011, pp. 1-13.

[7] M. E. Davies and C. J. James, "Source separation using single channel ICA," Signal Process., vol. 87, no. 8, pp. 1819-1832, 2007.

[8] M. Zibulevsky and B. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," Neural Comput., 2001.

[9] P. S. Huang, S. D. Chen, P. Smaragdis, and M. H. Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in Proc of ICASSP, pp.57-60, 2012.

[10] F. Li and M. Akagi, "Weighted Robust Principal Component Analysis with Gammatone Auditory Filterbank for Singing Voice Separation," in Proc of ICONIP 2017(6):849-858.

[11] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in Adv. NIPS, pp. 556-562, 2000.

[12] L. Zhang, Z. Chen, M. Zheng, and X. He, "Robust non-negative matrix factorization," Frontiers of Electrical and Electronic Engineering in China, 6:192-200, 2011.

[13] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semisupervised separation of sounds from single-channel mixtures," in Proc. 7th International Conference on Independent Component Analysis and Blind Signal Separation (ICA), UK, pp. 414-421, 2007.

[14] C. Fevotte, N. Bertin, and J. L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," Neural computation, vol.21, no.3, pp. 793-830, 2009.

[15] A. Cichocki, R. Zdunek, and S. Amari, "Csiszars divergences for nonnegative matrix factorization: Family of new algorithms," in Proc. 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA), SC, USA, pp. 32-39, 2006.

[16] C. Fevotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," Neural Computating, vol. 23, no. 9, pp. 2421-2456, 2011.

[17] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," IEEE Transactions on Audio, Speech, and Language Processing, vol.14, no.4, pp. 1462-1469, 2006.

Sources of funding for research presented in a scientific article or scientific article itself

Report potential sources of funding if there is any