Accompaniment Music Separation Based on 2DFT and Image Processing (Workshop)

Tian Zhang(✉), Tianqi Zhang, and Congcong Fan

Chongqing University of Posts and Telecommunications, Chongqing, China
284148541@qq.com, zhangtq@cqupt.edu.cn, 2669432120@qq.com

Abstract. For the difficulty of separation of accompaniment from mono music, image filtering was applied into a novel approach to separate accompaniment music. Our approach presents how single channel music manifests in the 2D Fourier Transform spectrum. In image domain, the position of periodic peak energy was determined by image filtering, and then masking matrix was constructed by rectangular window to extract the constituent of the accompaniment music. We find that our system is more robust and very simple to describe. The simulation experiments show that the method in this work has an advantage over other separation algorithm.

Keywords: Accompaniment separation · 2 dimension fourier transform · Time-frequency mask · Image processing

1 Introduction

In the information age, the demand for music signal processing technologies such as music annotation, retrieval, and identification, under massive digital music is growing. However, the correlation between accompaniment music and human voice makes it difficult for accompaniment and vocals to be extracted separately, which brings huge obstacles to music processing. The separation of vocal accompaniment in the music signal, as a pre-treatment of these techniques, has drawn increasing attention and has important research value.

In recent years, many experts have conducted in-depth research on music separation. Li and Hsu et al. used pitch estimation [1–3] to generate a sound music template, and Li used amplitude and phase information to further estimate the pitch [2] to generate a more accurate template, and then used the template to extract singing voice from the mixed music.

Using the sparseness of the vocal signal and the low rank property of the music accompaniment, Huang separates the mixed signal amplitude spectrum into a sparse matrix and a low rank matrix [4], and then uses the binary mask to realize the separation of music. REPET used the beat spectrum to extract background music, based on the priori knowledge of musical accompaniment with a certain periodicity [5–8]. Raffi based on local self-similarity of the accompaniment music, proposed adaptive method [9]. And

H. Gao et al. (Eds.): ChinaCom 2019, LNICST 313, pp. 414–424, 2020.
https://doi.org/10.1007/978-3-030-41117-6_34

similar matrixes [10] are used to extract the model of the repeated background music, which further improves the accuracy of the separation. The separation methods studied by the above scholars can separate the accompaniment music to a certain extent, but the robustness of the algorithm is poor, and the separation effect of different music segments is different.

In music information retrieval, 2DFT (2 Dimension Fourier Transform) has been used for song recognition [11, 12] and music segmentation [13]. Stöter and Fabian Robert [14] and others used 2DFT transform as input for sound source separation. Pishdadian et al. [15] used a multi-resolution 2D patches instead of a fixed-size 2D patches to further improve this separation method. Both approaches focus on distinguishing different characteristics of the sound source (such as vibrato, etc.), to separate sources with the same fundamental frequency from one to another in short audio. Both need to create more complex multi-resolution filter banks using 2DFT. At present, most scholars use the sparseness and periodicity of signals to separate music while few scholars use 2DFT transform to separate the accompaniment music.

Based on the above analysis, we explored a novel accompaniment separation method using image processing based on 2DFT. Our system is with high robustness and very easy to describe and implement and competitive to existing music separation method.

2 Proposed Method

2.1 The 2D Fourier Transform on Single Channel Musical Signals

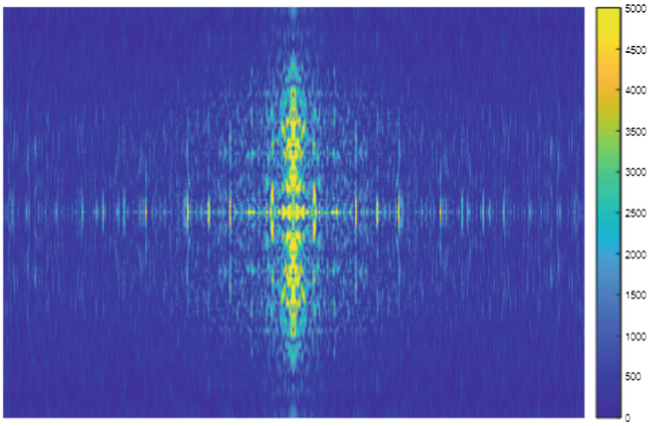
The 2DFT, like the IDFT in music analysis, is a popular technique in digital image processing, and is used for image denoising and compression, among other things, but 2DFT cannot be applied on single channel audio. By taking the magnitude of the 2DFT on the STFT (Short-Time Fourier Transform), we obtain a key-invariant representation of the audio.

Let $\mathbf{X}_{\omega,\tau}$ denote the constant Q transform (CQT) of the music signal $f(t)$, $\mathbf{W}_{\omega,\tau} = \left|\mathbf{X}_{\omega,\tau}\right|$ is its amplitude spectrum, where and are variables representing frequency and time respectively. The 2D Fourier transform is expressed as follows

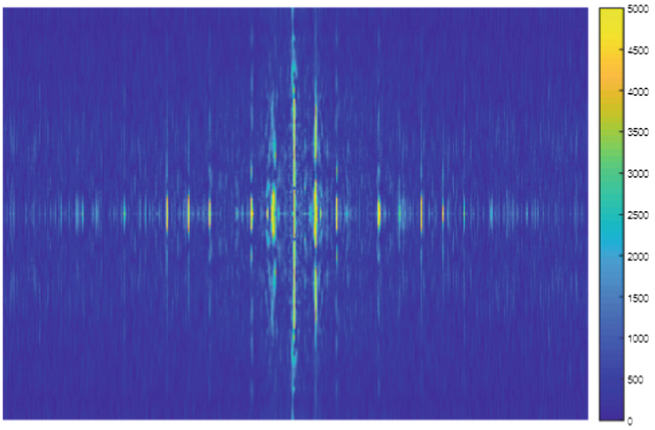
$$F(u,v) = \frac{1}{MN}\sum_{\omega=0}^{M-1}\left[\sum_{\tau=0}^{N-1} W_{\omega,\tau} exp(-j2\pi v\tau/N)\right]\exp(-j2\pi u\omega/M) \quad (1)$$

The vertical dimension and horizontal dimension of 2DFT domain are called scale and rate. These terms are borrowed from studies of the auditory system in mammals [16–18]. Figure 1 presents 2DFT spectrum of different music signals.

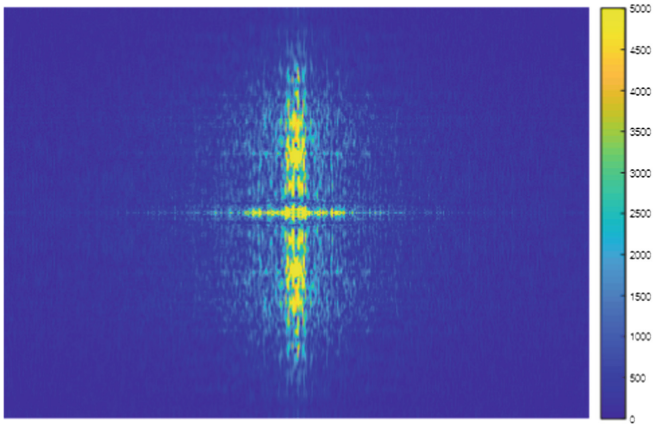
It can be seen from Fig. 1(b) that the energy of the pure singing voice mainly concentrates on the central region of the 2DFT transform spectrum, while the energy of the singing voice is striped from the center of the 2DFT transform spectrum to the two sides as shown in Fig. 1(c). The 2DFT transform spectrum of mixed music is a superposition of pure vocal and pure music spectra as shown in Fig. 1(a). If the 2DFT transform spectrum of the mixed music can be separated into the forms of the two figures (b) and (c) of Fig. 1, the accompaniment and the singing voice can be separated.



(a) The 2DFT spectrum of mixed music



(b) The 2DFT spectrum of pure singing



(c) The 2DFT spectrum of pure accompaniment

Fig. 1. 2DFT spectrum of different music signals

2.2 Accompaniment Music Separation

We separate accompaniment music by constructing time-frequency masking. In 2DFT spectrum, we set 1 to the position of bright stripe, and 0 otherwise. By this way, we obtain time-frequency masking. To pick bright stripe positions, we compare the difference between the maximum and minimum magnitude values over a neighborhood surrounding each point in the scale-rate domain with a certain threshold. When the difference is greater than the threshold, the maximum point existing in the neighborhood is recorded. This means there is a sharp increase in energy compared to other points.

In this work, we design our neighborhood shape to be a rectangle whose size along the scale is 1. The size of our rectangle neighborhood along the rate axis varies from 15 to 50. If the neighborhood is too large, the singing will be easily leaked into the accompaniment. On the contrary, the separated accompaniment will be more easily leaked into the singing voice.

We denote the center point of our rectangle neighborhood by $C = (s_C, r_C)$, and represent the length of the neighborhood. Set the standard deviation γ of $W_{\omega,\tau}$ for the threshold and α for the difference between the maximum and minimum magnitude values over a neighborhood, that is

$$\alpha = \max_N |F(s,r)| - \min_N |F(s,r)| \tag{2}$$

The vocal masking matrix can be derived from the following formula:

$$M_{\mathrm{fg}}(s_C, r_C) = \begin{cases} 1 \; \alpha > \gamma & |F(s,r)| = \max_N |F(s,r)| \\ 0 & otherwise \end{cases} \tag{3}$$

Figure 2 shows the masking matrix of the singing voice.

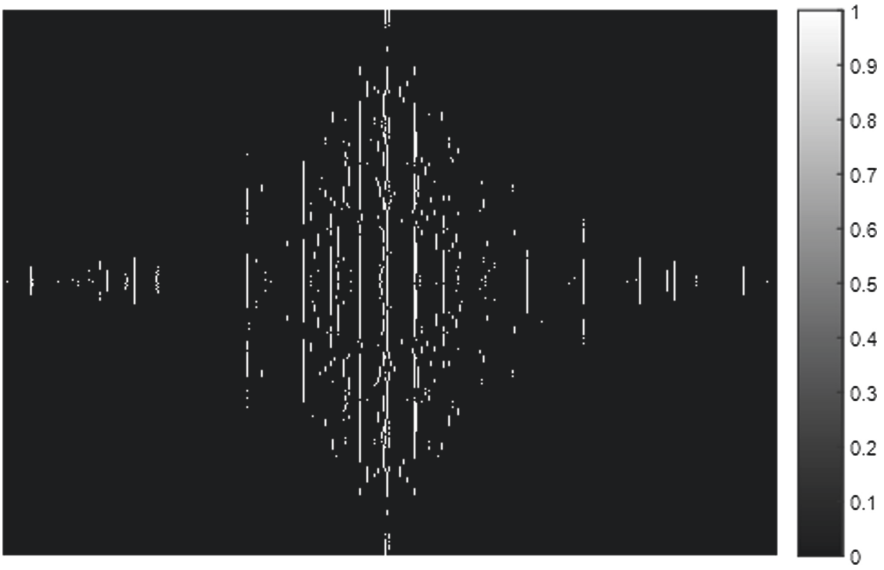


Fig. 2. The masking matrix of the song

It can be seen in Fig. 2, the positions of the value 1 are basically consistent with the positions of the singing value in mixed music matrix. Comparing masking matrix and 2DFT spectrum in Fig. 1, we find the energy of accompaniment music is mainly concentrated in the position of the center of the spectrum, and the vocal energy at this position is relatively low. Therefore, the masking matrix is processed to remove the center. The masking matrix after processing is shown in Fig. 3.

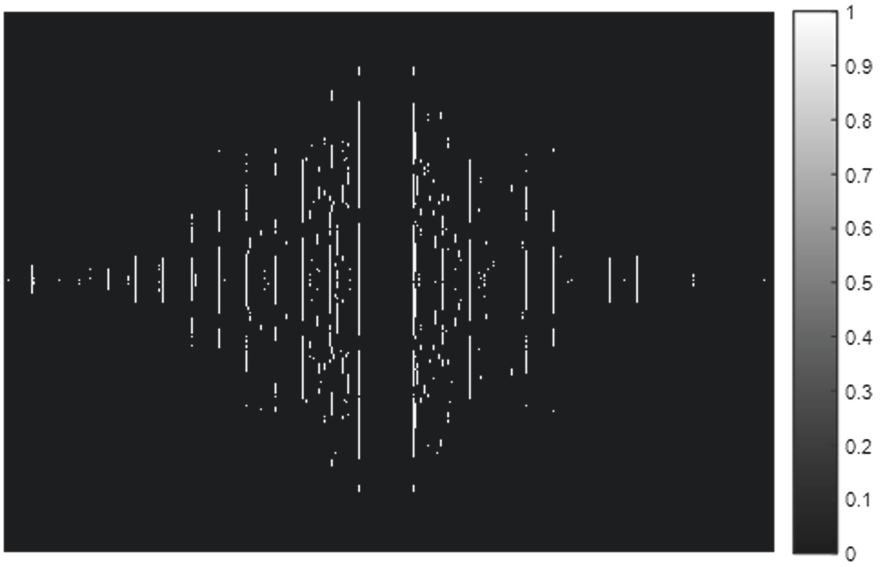


Fig. 3. Remove the center band of masking matrix

The masking matrix of accompaniment music can be calculated by the following formula:

$$M_{bg}(s, r) = 1 - M_{fg}(s, r) \tag{4}$$

Where represents inverse 2D Fourier transform, and denotes element-by-element multiplication. The time-frequency masking can obtain by comparing the magnitude spectrogram of accompaniment and singing.

$$M_{bg}(\omega, \tau) = \begin{cases} 1 & \left|\boldsymbol{X}_{bg}(\omega, \tau)\right| = \left|\boldsymbol{X}_{fg}(\omega, \tau)\right| \\ 0 & otherwise \end{cases} \tag{5}$$

The short-time Fourier spectrum of the accompaniment can be obtained by the two-dimensional inverse Fourier transform by the masking matrix of the singing voice and the 2DFT transform spectrum of the mixed music.

Finally, the time domain signal of the accompaniment can be obtained by time-frequency masking $M_{bg}(\omega, \tau)$ and the time-frequency spectrogram $\mathbf{X}(\omega, \tau)$ of the mixed signal.

$$x_{bg}(t) = ICQT\left\{M_{bg}(\omega, \tau) \cdot \boldsymbol{X}(\omega, \tau)\right\} \tag{6}$$

Where $ICQT\{\cdot\}$ is the inverse constant Q transform.

3 Evaluations

The music data set in the experiment uses the music data set MIR-1 K [19] published by Hsu Lab. The data set consists of 1,000 song clips in the form of split stereo WAVE files sampled at 16 kHz, extracted from 110 karaoke Chinese pop songs, performed mostly by amateurs, with the music and voice recorded separately on the left and right channels, respectively. The duration of the clips ranges from 4 to 13 s.

In order to quantitatively evaluate the separation effect of the method in this work, the Févotte Blind Source Separation Evaluation (BSS_EVAL) [20] was used to measure the performance of the improved algorithm. The toolbox provides a set of measures that intend to quantify the quality of the separation between the source signal and its estimate. The principle is to decompose the estimated signal as follows:

$$\hat{s}(t) = s_{t\arg et}(t) + e_{interf}(t) + e_{artif}(t) + e_{noise}(t) \tag{7}$$

Where $s_{t\arg et}(t)$ is the portion of the estimated signal that belongs to the source signal, and $e_{interf}(t)$ is the estimated error caused by the other signal source, that is, the portion of the estimated signal that is a mixed signal but does not belong to the source signal. $e_{artif}(t)$ represents the system noise error due to the algorithm itself, and denotes the noise interference error contained in the observed signal.

Since the effects of noise can be ignored in most music separations, $e_{noise}(t)$ can be omitted directly. Therefore, we use the following performance indicators, namely source-to-interference ratio (SIR) and source-to-artifacts ratio (SAR), which are defined as follows:

$$SIR = 10\lg\left[\frac{\left\|s_{t\arg et}(t)\right\|^2}{\left\|e_{interf}(t)\right\|^2}\right] \tag{8}$$

$$SAR = 10\lg\left[\frac{\left\|s_{t\arg et}(t) + e_{interf}(t)\right\|^2}{\left\|e_{artif}(t)\right\|^2}\right] \tag{9}$$

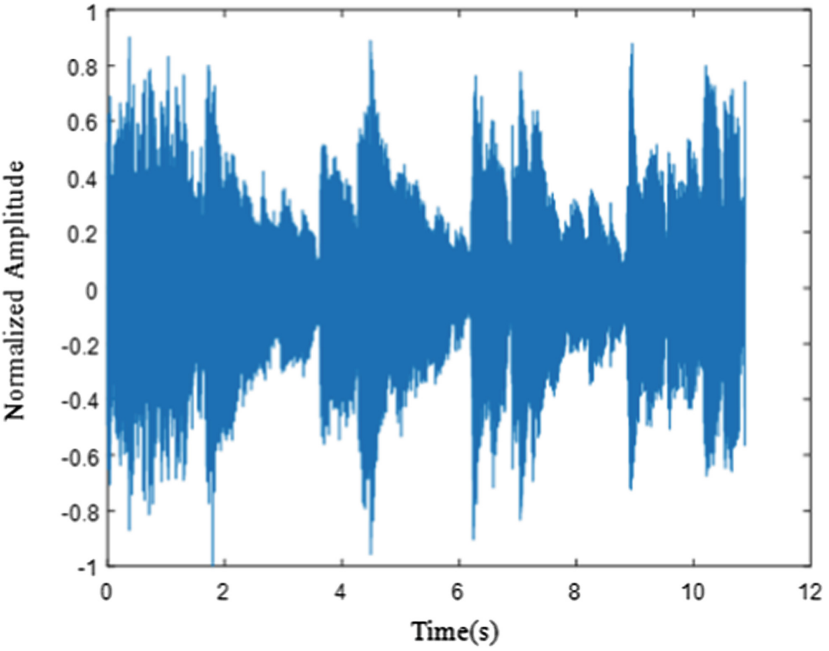
SIR represents the resolution of the algorithm, SAR represents the robustness of the algorithm, and the higher the values of SIR and SAR, the better the performance of the algorithm.

3.1 Comparative Results

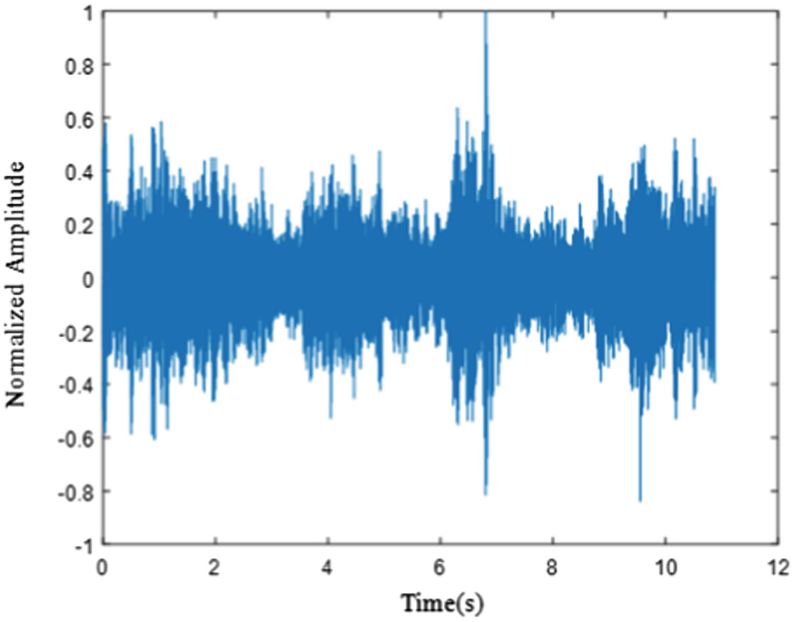
A piece of music (geniusturtle_5_01.wav) in MIR-1 K was randomly selected, and the accompaniment was separate by the method in this work. Waveforms comparison before and after separation are shown in Fig. 4.

It can be seen from Fig. 4 that the waveforms of the separated accompaniment and the original accompaniment are basically the same in shape, but the amplitude of the separated accompaniment waveform is reduced, which is caused by the neighborhood length being too small. It can be improved by adjusting the length of the rectangular neighborhood.

After verifying that the method can effectively separate the accompaniment music, the advantages of the method are explained. Five pieces of music in the MIR-1 K

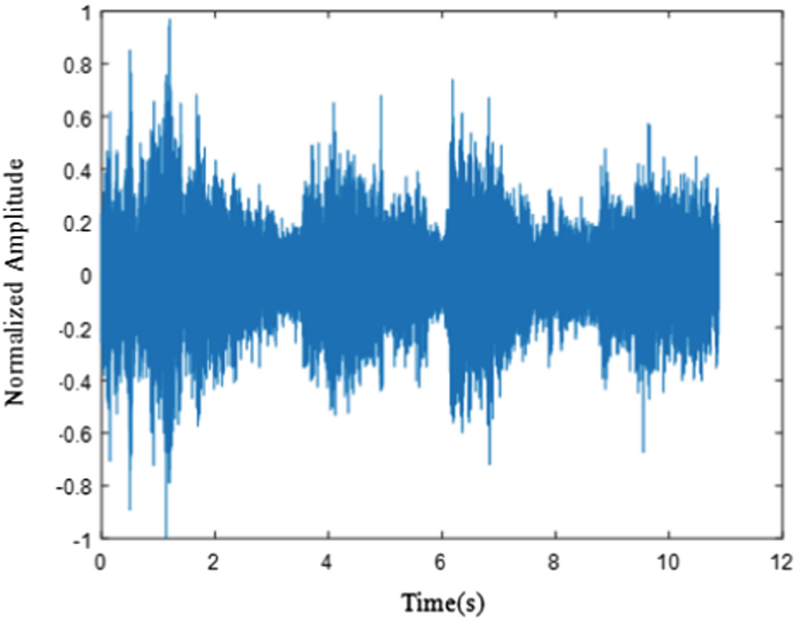


(a) Original accompaniment waveform

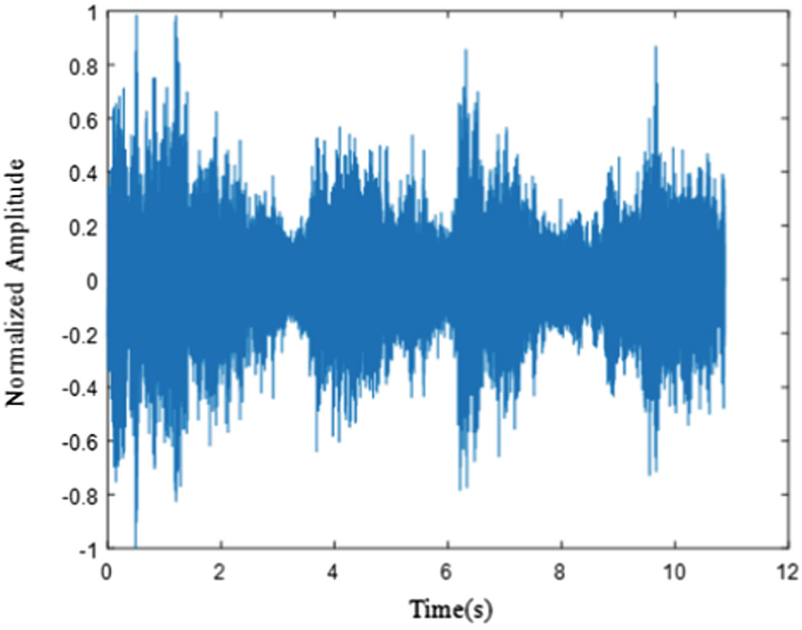


(b) Separated accompaniment with neighborhood is 15.

Fig. 4. Comparison of waveforms before and after separation



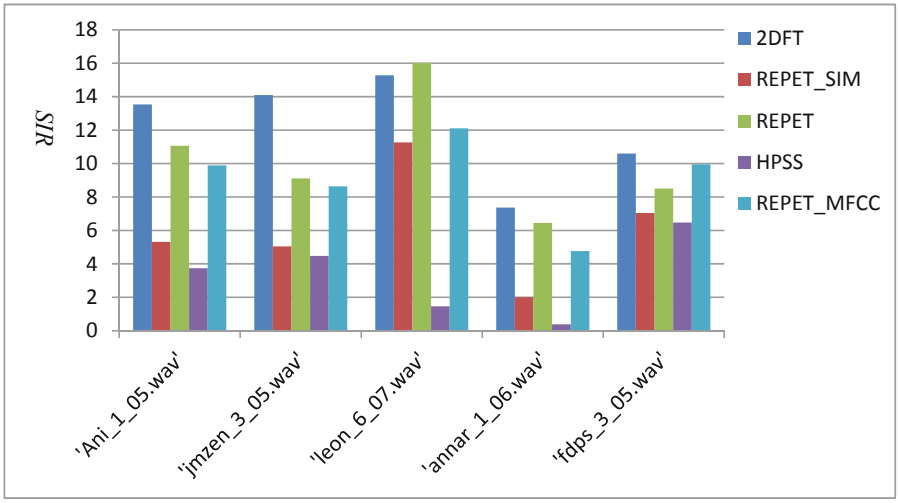
(c) Separated accompaniment with neighborhood is 35.



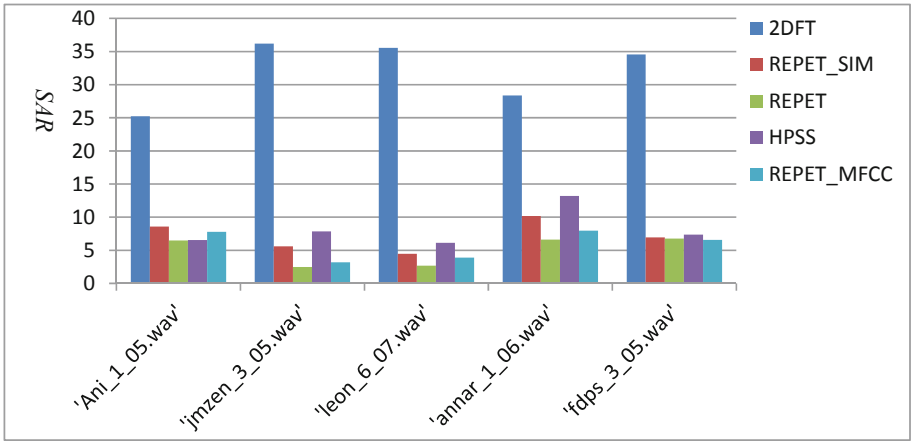
(d) Separated accompaniment with neighborhood is 50.

Fig. 4. (*continued*)

dataset are randomly selected and separated, and the SIR and SAR values are calculated. The extracted segment is 'Ani_1_05.wav. ', 'jmzen_3_05.wav', 'leon_6_07.wav', 'annar_1_06.wav', 'fdps_3_05.wav'), the result is shown in Fig. 5.



(a)Accompaniment separation indicator SIR (dB) contrast diagram



(b) Accompaniment separation indicator SAR (dB) contrast diagram

Fig. 5. Comparison diagram of random 5 music segment separation performance indexes

It can be seen from Fig. 5, our approach is superior to HPSS and REPET and its improved algorithm in separating indicators SIR and SAR when separating music accompaniment. In this work, proposed method has at least 2 dB improvement on SIR compared with other traditional algorithms. In SAR, our method keeps about 30 dB, which is at least 15 dB better than other algorithms.

The 500 pieces of music in MIR-1 K were separated by the 2DFT algorithm, and the SIR and SAR averages were calculated and compared with HPSS algorithm, REPET and its improved algorithm. The results are shown in Table 1.

Table 1. Separation performance (Average result)

Method	Accompaniment	
	SIR(dB)	SAR(dB)
HPSS	4.989	8.078
REPET	8.121	4.572
REPET-SIM	4.672	6.068
REPET-MFCC	7.346	5.278
2DFT	9.290	31.177

It can be seen from Table 1 that the 2DFT algorithm in this paper is about 4 dB higher than HPSS in SIR when separating accompaniment. Compared with REPET and its improved algorithm, SIR is improved by about 0.9–4 dB. In terms of SAR, the SAR of this algorithm is 27 dB. Better than other algorithms.

4 Conclusion

Aiming at the accompaniment separation in music separation, we proposed an accompaniment separation approach based on 2DFT transform. The method firstly transforms the single-dimensional music signal into a two-dimensional domain by 2D Fourier transform, and then uses the image filtering method to process the spectrogram. Thus we used the rectangular neighborhood to pick the position of the energy peak, and constructed the masking matrix to extract the music accompaniment component. Finally time-domain accompaniment was recovered by inverse transformation. Simulation experiments show that the music accompaniment separation method based on 2DFT does not need to create a complex filter bank and very easy to implement. We find that our system is competitive with existing unsupervised music separation approaches that leverage similar assumptions.

References

1. Li, Y., Wang, D.L.: Separation of singing voice from music accompaniment for monaural recordings. IEEE Trans. Audio Speech Lang. Process. **15**(4), 1475–1487 (2007)
2. Li, Y., Woodruff, J., Wang, D.L.: Monaural musical sound separation based on pitch and common amplitude modulation. IEEE Trans. Audio Speech Lang. Process. **17**(7), 1361–1371 (2009)
3. Hsu, C.L., Jang, J.S.R.: On the improvement of singing voice separation for monaural recordings using the MIR-1 K dataset. IEEE Trans. Audio Speech Lang. Process. **18**(2), 310–319 (2010)

4. Huang, P.S., Chen, S.D., Smaragdis, P.: Singing-voice separation from monaural recordings using robust principal component analysis. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 57–60 (2012)
5. Rafii, Z., Pardo, B.: A simple music/voice separation method based on the extraction of the repeating musical structure. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 221–224 (2011)
6. Rafii, Z., Pardo, B.: Repeating pattern extraction technique (REPET): a simple method for music/voice separation. IEEE Trans. Audio Speech Lang. Process. **21**(1), 73–84 (2013)
7. Zhang, T., Xu, X., Wu, W.: Music/vice separation based on the multi-repeating structure of mel-frequence cepstrum coefficients. Acta Acustica **2016**(1), 134–142 (2016). (in Chinese)
8. Rafii, Z., Pardo, B.: Online REPET-SIM for real-time speech enhancement. In: IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE (2016)
9. Liutkus, A., Rafii, Z., Badeau, R.: Adaptive filtering for music/voice separation exploiting the repeating musical structure. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 53–56 (2012)
10. Rafii, Z., Pardo, B.: Music/voice separation using the similarity matrix. In: 13th International Society for Music Information Retrieval(ISMIR), pp. 583–588 (2012)
11. Seetharaman, P., Rafii, Z.: Cover song identification with 2D fourier transform sequences. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 616–620 (2017)
12. Bertin-Mahieux, T., Ellis, D.P.: Large-scale cover song recognition using the 2D Fourier transform magnitude. In: 13th International Society for Music Information Retrieval Conference (2012)
13. Nieto, O., Bello, J.P.: Music segment similarity using 2D-fourier magnitude coefficients. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 664–668 (2014)
14. Stöter, F.-R., Liutkus, A., Badeau, R., et al.: Common fate model for unison source separation. In: IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). IEEE, pp. 126–130 (2016)
15. Pishdadian, F., Pardo, B., Liutkus, A.: A multi-resolution approach to common fate-based audio separation. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 566–570 (2017)
16. Chi, T., Ru, P., Shamma, S.A.: Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Am. **118**(2), 887–906 (2005)
17. Patterson, R.D., Allerhand, M.H., Giguere, C.: Time-domain modeling of peripheral auditory processing: a modular architecture and a software platform. J. Acoust. Soc. Am. **98**(4), 1890–1894 (1995)
18. Ru, P., Shamma, S.A.: Representation of musical timbre in the auditory cortex. J. New Music Res. **26**(2), 154–169 (1997)
19. http://sites.google.com/site/unvoicedsoundseparation/mir-1k[OL]
20. Vincent, E., Gribonval, R., Fevotte, C.: Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)