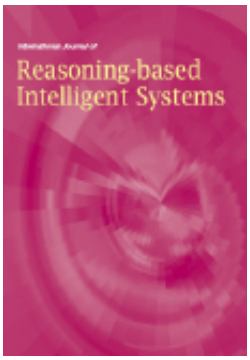
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Reasoning-based Intelligent Systems

ISSN online: 1755-0564 - ISSN print: 1755-0556
https://www.inderscience.com/ijris

Separating voice and background music based on 2DFT transform

Maoyuan Yin, Li Pan

DOI: 10.1504/IJRIS.2023.10057742

Article History:

Received:	10 March 2023
Last revised:	27 April 2023
Accepted:	27 April 2023
Published online:	18 March 2025

Copyright © 2025 Inderscience Enterprises Ltd.

Separating voice and background music based on 2DFT transform

Maoyuan Yin and Li Pan*

School of Music and Dance,
Mudanjiang Normal University,
Mudanjiang, 157011, China
Email: 0502009@mdjnu.edu.cn
Email: 0502017@mdjnu.edu.cn
*Corresponding author

Abstract: A new separation method for human voice and background music is designed to address the problems of large positioning errors, large feature extraction errors, and low separation accuracy in existing methods. Firstly, a microphone array is setup in the virtual space to complete signal denoising, and a generalised cross correlation function is introduced to achieve signal localisation. Then, construct a signal time spectrum graph, calculate the position change of signal energy on the frequency axis, and extract components in the sound signal frequency band and time frame. Finally, hamming window function is introduced to improve the 2DFT transform algorithm and build a signal separation model. The test results show that when the proposed method is applied, the localisation error of human voice is only 0.50% when the frame rate of human voice in audio is 1,000 kbps, and the error of background music feature extraction is only 0.05% when the sample audio sampling rate is 60 KHz. The separation accuracy of human voice and background music remains above 95%, with a maximum of nearly 99%. The application effect is good.

Keywords: 2DFT; voice; background music; separation; generalised cross correlation function; microphone array; separation model.

Reference to this paper should be made as follows: Yin, M. and Pan, L. (2025) 'Separating voice and background music based on 2DFT transform', *Int. J. Reasoning-based Intelligent Systems*, Vol. 17, No. 1, pp.50–57.

Biographical notes: Maoyuan Yin completed his Master of Arts, an Associate Professor of the School of Music and Dance of Mudanjiang Normal University, the President, and a supervisor of Master's students. His research directions are vocal music singing and teaching, music technology, and musical theory.

Li Pan completed his Doctor of Education, an Associate Professor of the School of Music and Dance of Mudanjiang Normal University, a Supervisor of Master's students, the Director of vocal music performance, and a Visiting Professor at the Mudanjiang University. His research direction is vocal music singing and teaching.

1 Introduction

In recent years, electronic information technology has been widely used in many fields of society, constantly enriching people's life (Büker and Hanili, 2021), and changing the mode of people's life and production. However, voice is still the most important way of communication for people. It can transmit information and is an important symbol that carries human emotions and thoughts. In the era of modern industrialisation, people need more information through pure voice. Based on this, it is of great significance to carry out research on the separation of human voice and background music (Li and Xiang, 2022). The so-called separation of voice and background music is to separate the target voice from the background noise, which is the basic task in signal processing. Music is formed by different sound source signals at the same time and has certain diversity, which makes it difficult to retrieve and separate music signals. Among them, the mixed signals contained in the background music are relatively complex (Garg and Sahu, 2022). A variety of complex and irregular signals make it difficult to quickly separate the voice in the background music. In recent years, there has been more and more research on this method. Researchers have designed many methods to solve this problem, and their results are as follows:

Mirbeygi et al. (2021) proposed a method for separating voice and music based on RPCA. This method designs a new random singular value decomposition algorithm in the non-convex optimisation environment to significantly reduce the complexity of the previous RPCA method. The feasibility of this separation method is verified by comparing the experimental results of different datasets

Copyright © 2025 Inderscience Enterprises Ltd.

with the most advanced methods. In the application of this method, there is a problem of large positioning error between voice and background music signals. As studied in Zhang et al. (2021a), an UNet music source separation method combining SE and BiSRU was proposed. Firstly, the Demucs model is improved and an end-to-end network UNet-SE-BiSRU is proposed. Then, attention mechanisms are introduced in the generalised encoding and decoding layers, and squeezing excitation blocks are used to selectively extract features based on the type of audio to be separated; the bidirectional short- and long-term memory network is improved into a bidirectional simple cycle unit to improve the parallelism of learning. This completes the design of the music source separation model. This method has a significant error in extracting background music source features during its application, and further improvement is needed. As shown in Deng (2021), a music separation method based on the combination of harmony and percussion sound source separation was studied. This method separates music signals at high resolution through a harmonic and percussion sound source separation algorithm, preserving the harmonic sound source; separate the impact sound source again using a flexible window non negative matrix separation algorithm. This method has the problem of low separation accuracy during application and cannot be applied on a large scale.

On the basis of the above methods, in order to improve the separation effect of voice and background music, this paper designs a new separation method of voice and background music based on 2DFT transform. Its technical route is as follows:

1 Setup an audio virtual space, and setup a microphone array to locate the human voice signal in the virtual space; remove noise signals based on near-field and far-field models, and introduce a generalised cross correlation function to locate the target background music signal.

2 To construct a time-frequency spectrum of human voice and background music signals, calculate the position change of sound signal energy on the frequency axis, and extract components in the sound signal frequency band and time frame based on the determined amplitude change of the time-frequency spectrum of human voice and music signals.

3 In order to improve the separation accuracy and efficiency, hamming window function is introduced to process the initial signal, then determine the one-dimensional signal set of voice and background music, reduce the dimension to process the one-dimensional time sequence characteristics of the sound signal, convert the signal into a two-dimensional signal with the aid of 2DFT transform algorithm, and build a separation model to achieve the separation of voice and background music.

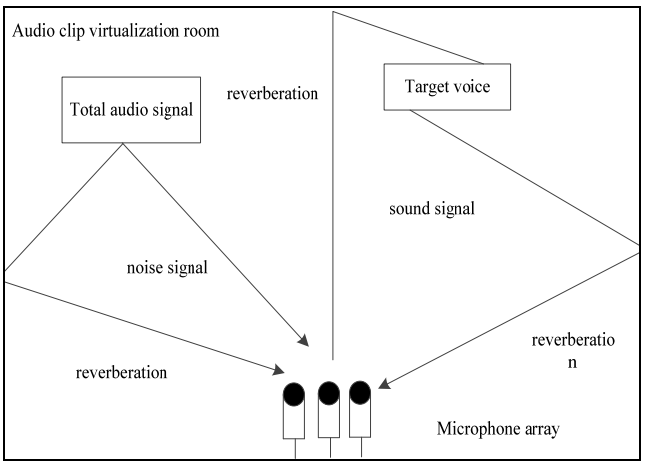
2 Research on localisation and feature extraction of voice and background music signals

2.1 Signal positioning

In the separation of voice and background music, it is necessary to determine the position of voice and background music. Due to the particularity of voice and background music, in order to improve the accuracy of their separation, this paper locates the signal positions of the two, determines different voice and background music, and lays the foundation for subsequent separation (Ge et al., 2021).

The characteristics of human voice are reflected in the vibration of sound. If there is vibration in a segment of audio, it means that the signal is human voice. The vibration of vocal cords in human voice signals is very important, and also includes part of vocal cords with silent vibration, which can be ignored in this study. The vocal cord vibration signal is complex and has no periodicity to follow, and its positioning process is similar to the process of positioning noise. In vocal positioning, it is assumed that the voice presents a linear invariant feature, but when the voice appears, it presents a time-varying feature (Benito-Gorron et al., 2021). Therefore, in the localisation of human voice signals, we mainly locate the short and stable human voice signals. In this location, simulate a virtual room of audio clip, and set microphone array in this room to determine the voice signal. The simulation process is shown in Figure 1.

Figure 1 Simulation diagram of vocal positioning process in audio virtual room



As shown in Figure 1, in the voice location in the audio virtual room, the voice is received through the microphone array. In the analogue space, except for the target voice, other sounds are considered as noise. The positioning model built during the positioning process is:

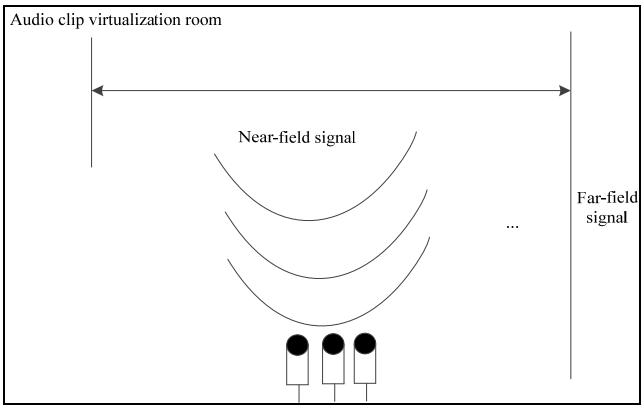
$$a_i(n) = b_i(n)c(n - d_i) + v_i(n) \tag{1}$$

In formula (1), $a_i(n)$ represents the localisation model, $b_i(n)$ represents the acoustic response value in the space unit localised by the i microphone, d_i represents the time required for the acoustic signal to reach the microphone, $v_i(n)$ represents the noise value of the i microphone

localisation, and c represents the interference coefficient in the voice localisation.

Based on the above vocal signal positioning, the background music signal positioning is still carried out in the virtual room. After removing the signal after the above vocal localisation, the remaining sound signal includes background music signal and noise signal. First, the noise signal is determined in the virtualisation space (Al-Dhief et al., 2021). Because the noise signal is different from the background music signal, this can simplify the research process of the method. The noise signal in the virtual space shows a trend of spherical wave with the diffusion of the medium to the surrounding discontinuities. When the signal approaches the microphone array, it forms a near-field signal, which is easier to be detected than the far-field signal. The near-field and far-field schematic diagram of the signal is shown in Figure 2.

Figure 2 Noise audio signal positioning simulation process



As shown in Figure 2, when the near-field noise signal is close enough to the microphone, the position of the signal can be directly determined; when the signal is far away from the microphone array, it needs to be located comprehensively according to the signal attenuation. The location formula is:

$$R = \frac{2g^2}{\alpha} \tag{2}$$

In formula (2), R represents the position result of the noise signal, g^2 represents the maximum diameter around the array, and α represents the wavelength of the noise signal.

When the noise signal is located successfully, the generalised cross-correlation function (Cheng et al., 2021) is used to locate the target background music signal. Assuming that the background music signal exists in an ideal state in the virtual space, the mathematical model of the background music signal located by the array is:

$$\begin{Bmatrix} x_1(n) = h(n-t_1) + v_1(n) \\ x_1(n) = h(n-t_2) + v_3(n) \\ \cdots \\ x_n(n) = h(n-t_n) + v_n(n) \end{Bmatrix} \tag{3}$$

In formula (3), $x_n(n)$ represents the target background music signal, $v_n(n)$ represents the background music signal located by each array, and h represents the coefficient value of the cross-correlation function.

In the location of voice and background music signal, set the audio virtual space, and set the microphone array in the set virtual space to locate the voice signal; according to the near-field and far-field models, the noise signal is removed, and the generalised cross-correlation function is introduced to locate the background music signal of the target to realise the positioning research.

2.2 Signal feature extraction

According to the above positioning of voice and background music signals, in order to achieve effective separation of voice and background music, it is also necessary to carry out effective separation according to its characteristics. For this reason, this chapter mainly extracts the features of voice and background music signals to provide key data support for subsequent separation.

In the feature extraction of human voice and background music, these sound signals are used to construct time-frequency spectra. The frame length of the time-frequency spectra is set as a fixed value, and this value does not affect the stability of feature extraction. The adjacent frames in the sound signal are overlapped by three quarters (Gimeno et al., 2021). When the time frame of the constructed sound signal is longer, the time spectrum map is given a lower time resolution to ensure the accuracy of feature extraction. The time scaling of the sound signal is easy to change the speed of the original sound signal. After the time scaling, the constructed time spectrum diagram should remain stable on the frequency axis, but because of its constant change, the speed of these signals constantly shifts (Qian et al., 2021). Therefore, in order to make the constructed time-frequency spectrum stable, the translation relationship between them is expressed before extracting features. Set the frequency of any sound signal, and its energy is distributed around the sound signal sub-band, namely:

$$y(f) = 12 \times \log_2 \frac{f}{f_i} + 1 \tag{4}$$

In formula (4), f represents the frequency of any sound signal, f_i represents the initial frequency of any sound signal, and $y(f)$ represents the result of the offspring energy distribution of the sound signal.

When the signal components around it change due to the translation relationship, calculate the position change of the sound signal energy on the frequency axis (Nguyen et al., 2021), namely:

$$y(1+k)f - y(f) = 12 \times \log_2(1+k) \tag{5}$$

In formula (5), k represents the translation component of the sound signal on the frequency axis.

On the basis of the above calculation, the variation amplitude of the time-frequency spectrum diagram of the vocal and music signals is determined, namely:

$$s(k,t)=\log\left|y(k,t)\right| \tag{6}$$

In formula (6), $s(k, t)$ represents the change amplitude result of the frequency spectrum of human voice and music signals, and t represents the change time.

According to the amplitude change of the time-frequency spectrum of the determined voice and music signals, the components are extracted in the frequency band and time frame of the sound signal (Bahmei et al., 2022), that is, the key features of the voice and music signals are extracted, and the extraction results are as follows:

$$W(k,t)=\frac{1}{n}\frac{s(k,t)-\min(s)}{\max(s)-\min(s)} \tag{7}$$

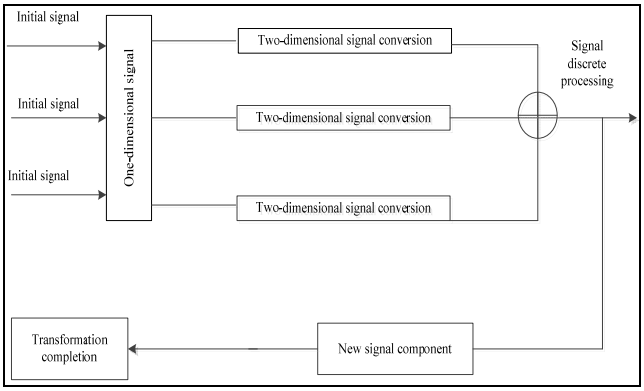
In formula (7), $W(k, t)$ represents the human voice and musical signal features, max(s) represents the maximum eigenvalue, and min(s) represents the minimum eigenvalue.

In the feature extraction of human voice and background music signals, the time-frequency spectrum of human voice and background music signals is constructed, and the position change of sound signal energy on the frequency axis is calculated. According to the determined change of the time-frequency spectrum of human voice and music signals, the components are extracted in the frequency band and time frame of the sound signal respectively, so as to realise the feature extraction research of human voice and background music signals.

3 Design of separation method for vocal and background music

2DFT transform is a two-dimensional discrete Fourier transform commonly used to process the spatial frequency characteristics of two-dimensional digital signals. Its algorithm can transform signals from time domain to frequency domain, as well as from frequency domain to time domain. When the signal changes from time domain to frequency domain, it is necessary to first perform a one-dimensional discrete Fourier transform on a given two-dimensional digital signal matrix in rows, and then perform a one-dimensional discrete Fourier transform on the results of the Fourier transform in columns. When the signal changes from frequency domain to time domain, it is necessary to first perform a one-dimensional discrete Fourier inverse transform on a given two-dimensional frequency domain image in columns; then, perform a one-dimensional discrete Fourier inverse transform on the results obtained from the inverse transform in rows (Elsayed et al., 2021). At present, this algorithm is widely used in fields such as image and signal processing. This algorithm has properties such as separability, periodicity, conjugate symmetry, translation, rotation, etc. It can quickly and effectively process two-dimensional signals, and the processing process is simple (Comanducci et al., 2021). Therefore, this paper introduces the 2DFT transform algorithm to achieve the final separation. The basic principle of the algorithm is shown in Figure 3.

Figure 3 Schematic diagram of basic principle of 2DFT transformation algorithm



First, in the separation of voice and background music, in order to improve the operation speed of the 2DFT transform algorithm, this study introduces the window function to improve it. First, the original signal is expanded to an integer power of 2 by zeroing, and the hamming window function is selected as follows:

$$W_i=0.54-0.46\times\cos\frac{2\pi n}{N} \tag{8}$$

In equation (8), n represents the sequence number of points in the sequence, and N represents the total length of the sequence. Complete the windowing of the initial function using the above equation. Afterwards, determine the one-dimensional signal set of vocals and background music, and set the one-dimensional time series of vocals and background music signals (Zhang et al., 2021b) as follows:

$$q_i=(q_0,q_1,\ldots,q_n) \tag{9}$$

$$p_i=(p_0,p_1,\ldots,p_n) \tag{10}$$

In formula (9) and (10), q_i represent the one-dimensional time series set of human voice signals and p_i represent the one-dimensional signal set of background music.

Then, the dimensionality of the signal characteristics of the one-dimensional time series determined above is reduced. This is because the dimension of the sound signal is constantly changing due to external influence during the collection and storage. Before the transformation, it needs to be dimensionally reduced. The result of the processing is:

$$U_i=\int(q_i,p_i)\sum_{i=1}^{n}\frac{l_i}{u_i} \tag{11}$$

In formula (11), U_i represents the dimension reduction result of one-dimensional temporal sequence, l_i represents the dimension reduction, and u_i represents the actual dimension change value of human voice and background music signals.

Thirdly, on the basis of the dimension-reduced voice and background music signal mentioned above, the signal is

converted into a two-dimensional signal with the help of the 2DFT transformation algorithm. The converted two-dimensional voice and background music signal (Sadeghi and Alameda-Pineda, 2021) can be expressed as follows:

$$q_i^{'} = \sum_{i=1}^{n} q_n F_n^T, T > 0 \tag{12}$$

$$p_i^{'} = \sum_{i=1}^{n} p_n F_n^T, T > 0 \tag{13}$$

In formulas (12) and (13), $q_i^{'}$ represents the transformed two-dimensional voice signal, $p_i^{'}$ represents the transformed music background signal, F represents the two-dimensional voice and background music signal conversion factor (Zha et al., 2021), and T represents the conversion substitution factor.

Finally, the voice and background music signals converted by 2DFT transform algorithm are separated, and the separation model is as follows:

$$Z_k\left(q_i^{'}, p_i^{'}\right) = \sum_{k=1}^{n-1}\left(\tau_n^k\right) + r_i \int \circ F_n^T \tag{14}$$

In formula (14), $Z_k(q_i^{'}, p_i^{'})$ represents the separation result output of human voice and background music signals, r_i represents the two-dimensional complex vector, and τ_n^k represents the inverter vector of human voice and background music signals (Tuncer, 2021).

In the process of separating voice and background music, determine the one-dimensional signal set of voice and background music, reduce the dimension and process the one-dimensional time series characteristics of sound signal, transform the signal into two-dimensional signal with the help of 2DFT transform algorithm, build a separation model, and realise the separation of voice and background music.

4 Experimental analysis

4.1 Experimental scheme design

To verify the effectiveness of the proposed method in separating human voice and background music, experimental testing and research were conducted. This experiment takes the audio database in NetEase Cloud platform as the research object, selects several classic audio clips in the database, converts the audio signal to the. wav format, and pre-processes the signal such as noise reduction and downsampling, compiles it into a dataset, which occupies 5.7 GB of memory, then selects some as the training set for model training, and the rest as the test set to facilitate the performance test of the proposed method. The specific test parameters are shown in Table 1.

Table 1 Experimental parameters

Parameter	*Details*
Test system	Windows 10
Testing software	Programming
Sample audio length/min	5
Sample audio voice frame rate/kbps	<1,411
Sample audio format	Little
Sample audio bit rate/kbps	1,411.2
Sample audio channel	2
Audio sampling rate/KHz	65
Audio bit depth/bit	16

4.2 Setting of experimental indicators

Conduct experimental analysis based on the set experimental parameters, and compare the proposed method, Mirbeygi et al. (2021) method, and Zhang et al. (2021a) method in the analysis. To ensure the accuracy of the experiment, maintain a consistent experimental environment during the test. Select indicators for testing such as positioning error of vocal and background music signals, background music feature extraction error, and separation accuracy of vocal and background music. The meaning and calculation process of the set experimental indicators are as follows:

- Indicator 1: Positioning error of voice and background music signal. This indicator mainly reflects the accurate position of the human voice signal in the sample audio clip separation, and takes this position as the separated human voice position. Therefore, this indicator is more critical. The formula for calculating the positioning error of human voice is as follows:

$$D_i^t = \sum_{i=1}^{n} \frac{d_i}{d_j} \times 100\% \tag{15}$$

  In formula (15), D_i^t represents the calculation result of acoustic localisation error, d_i represents the standard position of acoustic signal, d_j represents the actual position of the acoustic signal.

- Indicator 2: Background music feature extraction error. This indicator reflects the key features of background music. According to this signal feature, background music and voice are extracted and separated. The calculation formula of this indicator is:

$$E_{ij} = \frac{1}{n} \sum \frac{e_i}{e_j} \tag{16}$$

  In formula (16), E_{ij} represents the error result value of background music feature extraction, n represents the feature extraction times, e_i represents the key features of background music, and e_j represents the total features of background music.

- Indicator 3: Separation accuracy of voice and background music. This indicator reflects the quality of the separation method as a whole. This indicator is displayed as a percentage. The value range is [1–100]%. The closer the value is to 100%, the better the separation effect is.

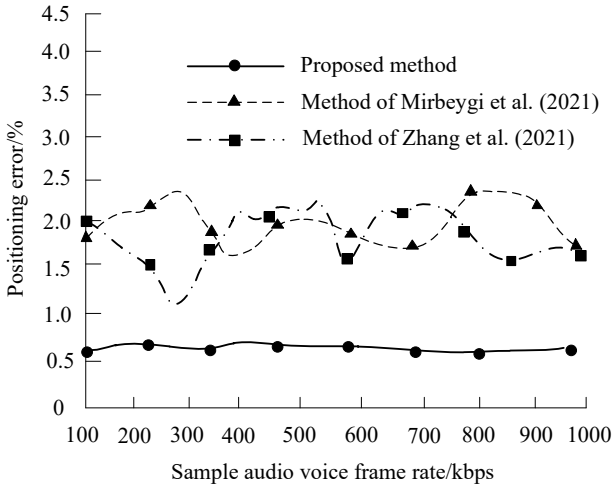
  Based on this, conduct research on the separation testing of vocal and background music signals.

4.3 Analysis of experimental results

4.3.1 Analysis of positioning error between human voice and background music signals in audio clips

Proposed method, Mirbeygi et al. (2021) method and Zhang et al. (2021a) method are used to locate the human voice in the sample audio clip, and the positioning error analysis results are shown in Figure 4.

Figure 4 Analysis of positioning error results of vocal and background music signals in sample audio clips



It can be seen from Figure 4 that there are some differences in the results of positioning errors after applying the proposed method, Mirbeygi et al. (2021) method and Zhang et al. (2021a) method. When the human voice frame rate in the audio is 200 kbps, the results of the human voice positioning error of the three methods are about 0.51%, 2.10% and 1.51% respectively; when the frame rate of human voice in audio is 600 kbps, the results of human voice positioning error of the three methods are about 0.50%, 1.95% and 1.68% respectively; when the frame rate of human voice in audio is 1,000 kbps, the results of human voice positioning error of the three methods are about 0.50%, 1.53% and 1.52% respectively; through the change of different frame rates, the positioning errors of the three methods are not the same, and the positioning errors of the three methods are the lowest, which shows that the proposed method is feasible.

4.3.2 Error analysis of background music feature extraction in audio clips

The proposed method, Mirbeygi et al. (2021) method and Zhang et al. (2021a) method are used to extract background music features in sample audio clips, and the error analysis results are shown in Table 2:

Table 2 Background music feature extraction error in audio clips with different methods

Audio sampling rate/KHz	*Background music feature extraction error/%*		
	Mirbeygi et al. (2021) method	*Zhang et al. (2021a) method*	*Proposed method*
10	0.12	0.16	0.02
20	0.17	0.18	0.03
30	0.22	0.19	0.03
40	0.25	0.25	0.04
50	0.28	0.27	0.04
60	0.31	0.28	0.05

Analysing the data in Table 2, it can be seen that there are certain differences in the error results when the proposed method, the proposed method, Mirbeygi et al. (2021) method and Zhang et al. (2021a) method are used to extract background music features. When the sample audio sampling rate is 10 KHz, the error results of the proposed method, the proposed method, Mirbeygi et al. (2021) method and Zhang et al. (2021a) method for background music feature extraction are 0.02%, 0.12% and 0.16% respectively; when the sample audio sampling rate is 30 KHz, the error results of background music feature extraction using the proposed method, the proposed method, Mirbeygi et al. (2021) method and Zhang et al. (2021a) method are 0.03%, 0.17% and 0.19% respectively; when the sample audio sampling rate is 60 KHz, the error results of background music feature extraction using the proposed method, Mirbeygi et al. (2021) method and Zhang et al. (2021a) method are 0.05%, 0.31% and 0.28% respectively; by comparing the data in the table, it can be seen that the error of the results obtained by the proposed method for background music feature extraction is the smallest, followed by the method in Zhang et al. (2021a) method, which verifies the feasibility of the proposed method.

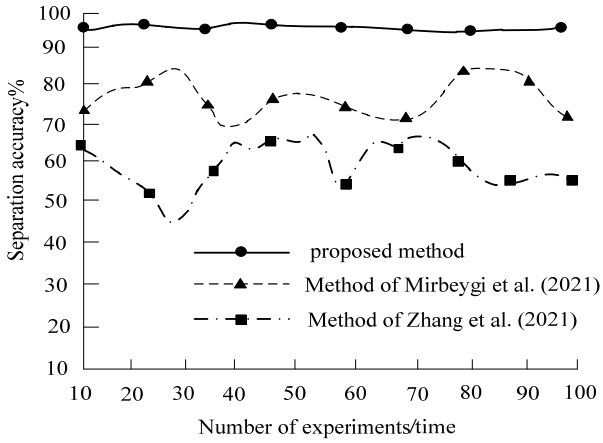
4.3.3 Analysis of the separation accuracy of vocal and background music in audio fragments

The proposed method, Mirbeygi et al. (2021) method and Zhang et al. (2021a) method are used to test the separation accuracy of human voice and background music, and the results are shown in Figure 5.

By analysing the results in Figure 5, it can be seen that the separation accuracy of the proposed method between voice and background music is always above 95%, and the highest is close to 99%; however, the separation accuracy of the other two methods was once greatly distorted, and the separation accuracy was always lower than that of the

proposed method. This can prove that the proposed method can well separate the voice and background music of the sample audio clip, and verify the feasibility of this method. The reason for this phenomenon is that the 2DFT transform algorithm is applied in this research, and hamming window function is used to improve it, which improves the separation accuracy and efficiency, optimises the method, and makes it obtain better application effect.

Figure 5 Separation accuracy results of vocal and background music of sample audio clip



5 Conclusions

The separation of voice and background music is related to the development of audio signal processing. In view of the shortcomings of existing methods, this study proposes and designs a separation method of voice and background music based on 2DFT transform. This method optimises and improves it from three aspects:

1 Audio virtual space is set, and microphone array is set in the virtual space to locate the voice signal; according to the near-field and far-field models, the noise signal is removed, and the generalised cross-correlation function is introduced to locate the target background music signal to improve its positioning accuracy.

2 The time-frequency spectrum of voice and background music signal is constructed, and the position change of sound signal energy on the frequency axis is calculated. According to the determined change of the time-frequency spectrum of voice and music signal, the components are extracted in the frequency band and time frame of the sound signal respectively to achieve the feature extraction of voice and background music signal.

3 Determine the one-dimensional signal set of voice and background music, reduce the dimension and process the one-dimensional time sequence characteristics of the sound signal, transform the signal into two-dimensional signal with the help of 2DFT transform algorithm, and build a separation model to realise the separation of voice and background music.

The implementation results show that the proposed method can reduce the positioning error of voice and background music signal. When the frame rate of voice in audio is 1,000 kbps, the positioning error of voice is about 0.50%; it can reduce the background music feature extraction error in audio clips. When the sample audio sampling rate is 60 KHz, the background music feature extraction error is 0.05%; it can improve the separation accuracy of voice and background music signals, and its separation accuracy is always above 95%, which is superior to the comparison method, and has great research value.

Acknowledgements

This work is supported by the Basic Scientific Research Projects of Heilongjiang provincial universities in 2022 under Grant Nos. 1452cxy004 and 1452ZD003.

References

Al-Dhief, F.T., Baki, M.M., Latiff, N. et al. (2021) 'Voice pathology detection and classification by adopting online sequential extreme learning machine', *IEEE Access*, Vol. 32, No. 7, pp.543–557.

Bahmei, B., Birmingham, E. and Arzanpour, S. (2022) 'CNN-RNN and data augmentation using deep convolutional generative adversarial network for environmental sound classification', *IEEE Signal Processing Letters*, Vol. 29, No. 11, pp.2032–2045.

Benito-Gorron, D.D., Ramos, D. and Toledano, D.T. (2021) 'A multi-resolution CRNN-based approach for semi-supervised sound event detection in DCASE 2020 challenge', *IEEE Access*, Vol. 14, No. 9, pp.2987–2996.

Büker, A. and Hanili, C. (2021) 'Deep convolutional neural networks for double compressed AMR audio detection', *IET Signal Processing*, Vol. 15, No. 4, pp.265–280.

Cheng, L., Peng, R., Li, A., Zheng, C. and Li, X. (2021) 'Deep learning-based stereophonic acoustic echo suppression without decorrelation', *The Journal of the Acoustical Society of America*, Vol. 150, No. 2, pp.816–829.

Comanducci, L., Bestagini, P., Tagliasacchi, M., Sarti, A. and Tubaro, S. (2021) 'Reconstructing speech from CNN embeddings', *IEEE Signal Processing Letters*, Vol. 21, No. 19, pp.1–14.

Deng, X. (2021) 'Research on music separation method based on the combination of harmony and percussion sound source separation', *Techniques of Automation and Applications*, Vol. 40, No. 8, pp.66–69.

Elsayed, N.E., Tolba, A.S., Rashad, M.Z., Belal, T. and Sarhan, S. (2021) 'A deep learning approach for brain computer interaction-motor execution EEG signal classification', *IEEE Access*, Vol. 9, No. 10, pp.101513–101529.

Garg, A. and Sahu, O.P. (2022) 'Deep convolutional neural network-based speech signal enhancement using extensive speech features', *International Journal of Computational Methods*, Vol. 19, No. 8, pp.2142005–2142012.

Ge, W., Zhang, T., Fan, C. and Zhang, T. (2021) 'Monaural noisy speech separation combining sparse non-negative matrix factorization and deep attractor network', *Acta Acustica*, Vol. 46, No. 1, pp.55–66.

Gimeno, P., Mingote, V., Ortega, A., Miguel, A. and Lleida, E. (2021) 'Generalizing AUC optimization to multiclass classification for audio segmentation with limited training data', *IEEE Signal Processing Letters*, Vol. 29, No. 8, pp.11–21.

Li, J. and Xiang, S. (2022) 'Audio-lossless robust watermarking against desynchronization attacks', *Signal Processing*, Vol. 198, No. 12, pp.108561–108573.

Mirbeygi, M., Mahabadi, A. and Ranjbar, A. (2021) 'RPCA-based real-time speech and music separation method', *Speech Communication*, Vol. 126, No. 13, pp.22–34.

Nguyen, D., Nguyen, D.T., Zeng, R., Nguyen, T.T., Tran, S.N., Nguyen, T., Sridharan, S. and Fookes, C. (2021) 'Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition', *IEEE Transactions on Multimedia*, Vol. 12, No. 12, pp.1001–1012.

Qian, X., Liu, Q., Wang, J. and Li, H. (2021) 'Three-dimensional speaker localization: audio-refined visual scaling factor estimation', *IEEE Signal Processing Letters*, Vol. 12, No. 9, pp.4–12.

Sadeghi, M. and Alameda-Pineda, X. (2021) 'Mixture of inference networks for VAE-based audio-visual speech enhancement', *IEEE Transactions on Signal Processing*, Vol. 13, No. 2, pp.1098–1101.

Tuncer, T. (2021) 'A new stable nonlinear textural feature extraction method based EEG signal classification method using substitution box of the Hamsi hash function: Hamsi pattern', *Applied Acoustics*, Vol. 17, No. 2, p.107607.

Zha, X., Xu, S., Liu, K. and Huang, L. (2021) 'A fuzzy radial basis inference network with grouped signal feature embedding and its application in multi-source signals classification', *IEEE Access*, Vol. 17, No. 21, pp.119–131.

Zhang, R.F., Bai, J.T., Guan, X. et al. (2021a) 'Music source separation method based on UNet combining SE and BiSRU', *Journal of South China University of Technology (Natural Science Edition)*, Vol. 49, No. 11,. pp.106–115+134.

Zhang, Y., Zhang, K., Wang, J. and Su, Y. (2021b) 'Robust acoustic event recognition using AVMD-PWVD time-frequency image', *Applied Acoustics*, Vol. 178, No. 5, p.107970.