Mixing or Extracting? Further Exploring Necessity of Music Separation for Singer Identification

Yuxin Zhang*, Yatong Xiao†, Wei-Qiang Zhang†[1], Xu Tan‡, Ling Lei*, Shengjin Wang†
* Key Laboratory of Media Audio & Video, Ministry of Education, Communication University of China, Beijing 100024, China
E-mail: zhang.yuxin@cuc.edu.cn leiling@cuc.edu.cn
† Beijing National Research Center for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
E-mail: xiaoyt19@mails.tsinghua.edu.cn wqzhang@tsinghua.edu.cn wgsgj@tsinghua.edu.cn
‡ Microsoft Research Asia
E-mail: xuta@microsoft.com

***Abstract*—One song has two major acoustic components that are singing vocals and background accompaniment. Although identifying singers is similar to speaker identification, it is challenging due to the influence of background accompaniment on the singer-specific information in singing vocals. In past work on singer identification, studies on smaller datasets have considered the introduction of audio-source separation to remove the accompaniment to be beneficial for singer identification. In our work, this was not found to be absolutely valid for identification accuracy on a larger dataset with a wider variety of acoustic environments. Moreover, to further illustrate the necessity of removing accompaniment in the singer identification problem, we collected three characteristic datasets focusing on backing tracks for publicly released songs, cover songs, and multiple songs per singer. And general and specific system performance example results are given to reveal the effectiveness and reliability of removing the accompanying sound.**

I. Introduction

The development of singer identification enables the effective management and exploration of large amounts of music data based on singer similarity. With this technology, songs performed by a particular singer can be automatically clustered for easy management or searching. Several studies in singer identification pay attention to features extraction directly from the songs [1][2]. In popular music, the singing voice is often interwoven with the accompaniment. The presence of background music and chorus increases the uncertainty of the task.

In a part of the past work related to singer identification, researchers are aware of this problem. The proposed methods fall into two categories: one represents the singer's timbre by extracting local features from the mixture of singing and accompaniment [3], and another is to remove the accompaniment by audio source separation [4][5]. In their studies, both believe that the presence of accompaniment interferes with singer identification and negatively affects identification accuracy. At the same time, there is another part of the study that does not consider this issue [6].

[1]Corresponding author.

In our work, we focus on singer identification by extracting the singing voice through music separation. The separation was performed for the first time on JukeBox [7], the largest dataset currently available for the singer identification problem, to our knowledge. On the modified JukeBox test set, the Equal Error Rate (EER) is improved relatively by 15.7% on the i-vector-PLDA system but decreased obviously on the x-vector-PLDA system.

It seems that music separation does not result in performance gains for all systems. Hence, we tend to think that the presence of accompaniment assisted in the stronger identification system. Thus, although better identification accuracy was obtained in the system without separation, this result was not reliable. Further, the concept of music separation is useful in singer identification, but existing methods would cause impairment and distortion to the data, leading to worse results.

Therefore, to further reveal the availablity and reliability of extracting vocals, three datasets were designed as test sets for the experiments to illustrate this point from different perspectives. These three datasets are:

- Cover songs dataset. Specific examples will illustrate that the presence of accompaniment interferes with the system's identification due to the similarity brought by the accompaniment.

- Multiple songs per singer dataset. When the vast majority of songs in the evaluation set use different accompaniment from the enrollment set, EER increase relatively 2.8% on the i-vector-PLDA system and 11.1% on the x-vector-PLDA system, which means that the singer recognition system has better generalization capability after adding music separation.

- Backing tracks publicly released datasets. Through this experiment, we found that using the idealized way of extracting vocals improved the EER by at least 42.4% compared to the system without separation, but using the existing separation method, the EER was even 3 times higher than in indealized situation.

II. Related Works

Research on singer identification (SID) approaches has been divided into three categories: i) treating the singer identification problem as a speaker identification problem [8], ii) ignoring the influence of the accompaniment to identify singers directly through raw waveform audio data [3][9][10], and iii) removing the accompaniment by a specified method and then performing singer identification [11]. Our method belongs to the third one and shows experimentally that the results obtained by the other two methods are not always reliable.

Researchers used the i-vector to extract song-level descriptors from the frame-level timbre features Mel-frequency cepstral coefficients (MFCC). I-vectors, which are the result of a factor analysis procedure applied to frame-level features, provide a low-dimensional and fixed-length representation of each song [8]. This method was pioneered in speaker identification. Their experimental results achieved an F1 score of 0.846 on Artist20 [6].

There is also some research that extracts features directly from raw waveform audio data for singer identification. The fusion of timbre and vibrato features with different attributes was used to describe the vocal characteristics of the singer and found spectral roll-off correlating to singer characteristics by providing distinct band of roll-off frequencies for each singer, which obtained an accuracy of 0.805 on a 23-singer-dataset [9]. A WaveNet classifier was introduced that directly models the features from a raw audio waveform. The WaveNet takes the waveform as an input and then distinguishes the artists by several downsampling layers. Their experimental results achieved a F1 score of 0.854 on Artist20 [10]. The KNN-NET for singer identification was proposed, which is a deep neural network model with the goal of learning local timbre feature representation from the mixture of singer voice and background music. And an attention mechanism is introduced to highlight crucial timbre features for identification, which improved the results by 4% [3].

Several studies have focused on singer identification after removing the interference of background music by using music separation method. The Gaussian Mixture Method (GMM) and Support Vector Machine (SVM) have obtained accuracies of 96.42% and 81.23% with a dataset of 100 songs of 10 singers, which used Robust Principal Component Analysis (RPCA) to separate voice and accompaniment [4]. An effective system of singer identification with human voice separated from original music, which used Robust Principal Component Analysis (RPCA) to music separation with its high performance, and then the Linear Predictive Coding (LPC) method was chosen as the experimental method for feature extraction. Finally, the singer would be identified by Gaussian Mixture Model (GMM) with 63.6% accuracy in a dataset of 100 singers [5]. In addition, different audio-source separation methods were also compared in the singer identification task [11], and the researchers demonstrated that the state-of-the-art audio-source separation method brought the strongest gain in the accuracy of the singer identification system.

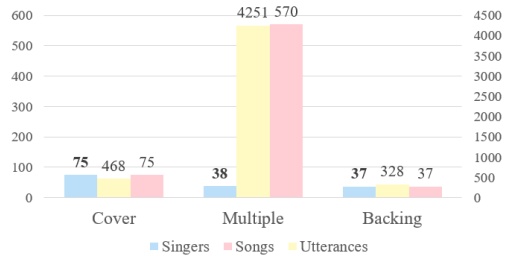

Fig. 1: Scale of three supplementary specific datasets.

Audio-source separation is also used for data augment in singer identification, accompaniments were shuffled and remix with vocals, which obtained better performance than original audio recordings. However, as researchers find in this experiment, sometimes vocal-only dataset obtain worse results than original songs [12]. This situation also appears in our experiments, and this paper is dedicated to explaining this problem by designing a series of specific datasets.

III. Datasets

A. JukeBox

This work chooses JukeBox, the largest known dataset available for singer identification, as the primary experimental source. In the train set of JukeBox, 670 singers have been included, with an average of 10 songs per singer. The whole train set consists of 38600 audio files. And for the test set of JukeBox, which consists of 1875 audio files, 98 singers have been included, and each singer has two songs.

However, in our experiments, some troubled audio files in the test set were removed to form a subset of JukeBox as our experimental test set. For example, we manually deleted some audio files without singing voice at all, as well as some duplicate audio files. The experiment finally keeps a total of 1772 data for 95 singers in the sub-test set[1].

B. Supplementary Datasets

1) Motivation and Overview:

Cover songs dataset. As we know, there are many similarities between a song and its cover version, but the singer's voice must be very different. Therefore, we designed the cover song dataset to explore the singer identification performance of the system where the song information is similar while the singer information was completely different.

Multiple songs per singer dataset. Consider that in the JukeBox test set and Cover song test set, the number of songs per artist is limited to no more than two. Therefore, most of the music data in the evaluation set shares similar accompaniment information with the enrollment set, including acoustic environment characteristics and the orchestration of the backing tracks, etc., and perhaps the system without

[1]The list of modified test set could be found at https://github.com/Valak0/Modified-Jukebox

separation will benefit from it in the identification system. Therefore, we created the Multiple songs per singer dataset, increasing the number of songs per singer.

Backing tracks publicly released dataset. In order to figure out whether accompaniment affects the results of the singer identification task, the experiment collect some songs, which all of them were published with the backing tracks of the songs released at the same time. And by using phase cancellation to capture the "true value" of the vocals in the song, we tested the singer identification system with an idealized method of removing the accompaniment.

2) Data-Mining Pipeline:

Listening and downloading songs of interest. We first listed the artists of interest (AOI) and songs of interest (SOI), and then downloaded these songs from QQ Music and NetEase Cloud Music. For cover songs dataset, we find 35 songs with their cover versions for a total of 75 songs from 75 singers. For multiple songs per singer dataset, 38 singers and their respective 15 songs were selected for the experiment. For backing tracks publicly released datasets, we have found 37 songs from 37 artists and downloaded each of these songs and their public accompaniment.

Changing audio settings. After downloading these songs, we clip each of these songs into multiple 30-second segments, then remove the segments that are less than 30 seconds. After this, each audio file was downsampled to 16 kHz, the bit depth was set to 16 bits, and changed to mono, just as set up in the JukeBox dataset. It should be mentioned that the backing tracks publicly released dataset was kept in stereo, for now, to prepare for acquiring vocals through phase cancellation.

Removing non-singing segments. We enhanced the method used in the JukeBox dataset by making our human listeners listen to 3 equally separated 1-second-long audio segments in every 30-second clip to make their decision, effectively avoiding the non-vocal utterance caused by prelude or interlude longer than 30 seconds.

3) Mined Datasets: Using the above data-mining pipeline, we obtain three supplementary datasets.[1] The scale of the dataset is shown in Fig. 1. These songs were processed to the same settings as JukeBox and were used as the test sets in our experiments.

C. Datasets Processing

For the backing tracks publicly released dataset, we processed the preserved stereo songs and backing tracks to generate four types of songs: *origin*, *accompaniment*, *vocal*, *separation*.

Origin: the original song.

Accompaniment: the backing tracks of the song released from the songs' publisher.

Vocal: the vocals obtained by overlaying the stereo original song and the stereo backing tracks inverted.

Separation: the vocals are extracted from the original song by audio-source separation.

[1]We release the supplementary datasets to public on http://github.com/Valak0/SPSID

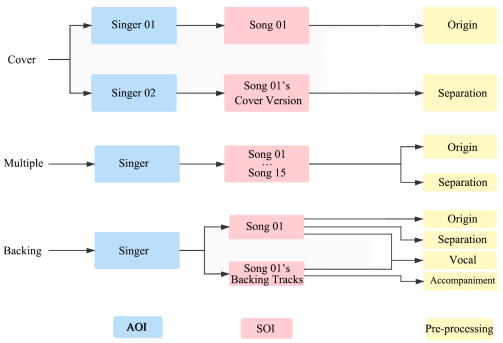

Fig. 2: Pre-processing on the supplementary datasets.

Note for *vocal*, when the inverted stereo backing tracks is overlayed on the stereo original song, most of the accompanying instruments have equal intensity in the left and right channels, and these sources would be canceled out due to phase cancellation [13]. The vocals, on the other hand, are not presented in the backing tracks and will therefore be preserved. However, this method may also preserve a very small portion of the accompaniment, but the vocals are not damaged in any way, which is different from audio-source separation method, the *vocal* obtained here can be approximated as "true value". After the above steps, all four types of songs are also converted to mono.

Besides, both the Cover song dataset and the Multiple songs per singer dataset were audio-source separated. Information about the datasets we designed and the preprocessing of them is presented in Fig. 2

IV. Implementation

In our work, we propose a singer identification system consisting of two key stages. The first stage separates the voice and accompaniment of the songs in datasets. In the second stage, original songs and separated voices are in comparison for the accuracy of identification results. As shown in Fig. 3.

A. Music Separation

For all of the datasets, the experiment uses the U-Net-based audio source separation toolkit spleeter[2] [14], the best performing open-source toolkit for music separation, using pretrained state-of-the-art model for separation: vocal/ accompaniment separation (2 stems) to remove the accompaniment.

B. Identification Model

1) I-Vector Model: In our work, the first model in each set of experiments is a traditional GMM-UBM i-vector model, which is based on the Kaldi recipe[3] [15]. We use the default setting in the Kaldi, the universal background model (UBM) involves 1024 Gaussian components, and the dimensionality of the i-vector is 400.

[2]http://github.com/deezer/spleeter
[3]http://github.com/kaldi-asr/kaldi

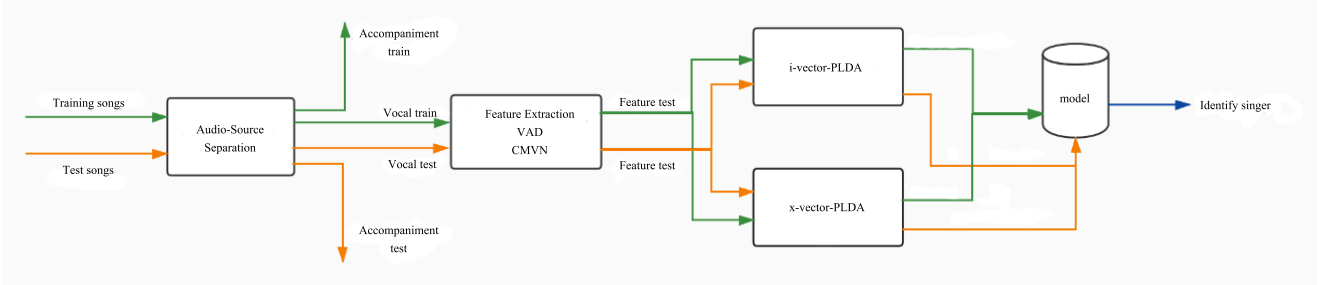

Fig. 3: The framework of the system for our experiments: a singer identification system with audio-source separation module.

2) X-Vector Model: The second model of each group of our experiments is the Kaldi x-vector model, as detailed in [16] and the Kaldi toolkit, the architecture is shown in Table. I. The DNN consists of three components: frame-level, statistics-level, and segment-level components. And the frame-level component is composed of the layers 1 to 5. These layers are constructed with a Time-Delay Neural Network (TDNN), assuming t is the current time step. The statistics-level component is an essential component that converts the variable length speech signal into a vector of fixed dimension. The statistics level is composed of one layer: the statistics-pooling, which aggregates over frame-level output vectors of DNN (layer 6) and calculates their mean and standard deviation. The segment-level component maps the segment-level vectors to speaker identities. The mean and standard deviation are concatenated together and forwarded to two additional hidden layers (layers 7 and 8), and finally to a SoftMax output layer (layer 9). In addition, this experiment does not use any data augmentation in the current work, which is worth exploring in the future.

3) Probabilistic Linear Discriminant Analysis (PLDA): In the singer identification task, given the two vectors η_1 as enrollment and η_2 to evaluate, we are interested in testing two alternative hypotheses H_s , that both η_1 and η_2 share the same speaker identity latent variable, or H_d that the i-vectors were come from different space. The verification score can be computed as

$$Score = log \frac{p(\eta_1, \eta_2|H_s)}{p(\eta_1|H_d)p(\eta_2|H_d)}$$

TABLE I: Architecture of the x-vector model in experiments.

Layer	Layer Type	Context	Size
1	TDNN-ReLu	$t-2:t+2$	512
2	TDNN-ReLu	$t-2, t, t+2$	512
3	TDNN-ReLu	$t-3, t, t+3$	512
4	Dense-ReLu	t	512
5	Dense-ReLu	t	1500
6	Pooling(mean+stddev)	t	10000
7	Dense(Embedding)- ReLu	t	512
8	Dense-ReLu	t	512
9	Dense-SoftMax	t	$NumSpk$

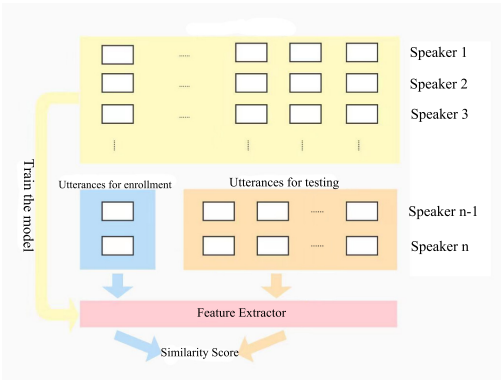

Fig. 4: Utterances for training, enrollment and testing.

the log-likelihood ratio for this hypothesis test as the higher the score, the more likely it is that two utterances from the same singer [17].

C. Other Experimental Setup

In all experiments in this paper, the test set will be split into enrollment and evaluation sets, an utterance from each singer in the test set will be randomly selected to become enrollment data, and the rest are going to be evaluated, as indicated in Fig. 4. Especially, the setup of the experiments on the backing tracks publicly released dataset will be indicated in Table IV.

V. Results and Discussion

A. Studies on JukeBox dataset

Table. IIa. indicates the performance of the proposed framework with audio-source separation and its comparison to baseline in JukeBox. It observes that the proposed framework with audio-source separation helps to improve the performance in the i-vector-PLDA system, but in the x-vector-PLDA system, it even leads to performance degradation. We believe that the explanations include the help of accompaniment and the distortion cause by music separation.

B. Studies on cover song dataset

Results of the experiments test on cover song dataset show in Table. IIb. In the experiments test on the cover song dataset, the results of the separated system do not perform better than

TABLE II: Identification result on JukeBox [J1], the cover song dataset [D1], and the multiple songs per singer dataset [D3]. [J2], [D2], [D4] represent separated [J1], [D1], [D3].The Models we used are i-vector-PLDA [M1] and x-vector-PLDA [M2].

(a) JukeBox

Exp	Train set/ Test set	Models	EER(%)
1	J1/J1	M1	27.53
2	J2/J2	M1	**23.22**
3	J1/J1	M2	**9.42**
4	J2/J2	M2	13.75

(b) Cover songs

Exp	Train set/ Test set	Models	EER(%)
1	J1/D1	M1	6.95
2	J2/D2	M1	8.93
3	J1/D1	M2	**1.24**
4	J2/D2	M2	2.23

(c) Multiple songs per singer

Exp	Train set/ Test set	Models	EER(%)
1	J1/D3	M1	24.35
2	J2/D4	M1	**23.65**
3	J1/D3	M2	14.06
4	J2/D4	M2	**12.50**

TABLE III: Both on the original music data system and separated music data system, similarity score between A-Lin's utterance from Chinese song named 《给我一个理由忘记》and her enrollment, and similarity score between A-Lin's utterance and Joyce Cheng's enrollment from the Cantonese song《忘记的理由》, which share identical accompaniment.

System without music separation			**System with music separation**		
Singer	**Utterance**	**Similarity Score**	**Singer**	**Utterance**	**Similarity Score**
S10001	S100010001	7.497327	S10001	S100010001	**15.33988**
S10060	S100010001	**9.062836**	S10060	S100010001	1.696868
S10001	S100010002	2.235353	S10001	S100010002	**12.37363**
S10060	S100010002	**7.536013**	S10060	S100010002	-6.050038
S10001	S100010003	4.02524	S10001	S100010003	**23.75219**
S10060	S100010003	**5.424552**	S10060	S100010003	-3.723748
S10001	S100010005	0.5797107	S10001	S100010005	**13.59131**
S10060	S100010005	**10.78016**	S10060	S100010005	1.931807
S10001	S100010010	1.3162	S10001	S100010011	**12.8726**
S10060	S100010010	**4.953076**	S10060	S100010011	-4.131832

the non-separated system. And yet, the singer identification results for some segments also suggest that the identification was impacted with by the accompaniment in a system without separation. In reviewing the trials-out documents, we found that the system with separation would not readily identify the singer of the song as the singer of the cover version of the song, as the system without separation did. Experiments without separation obtained several times that the similarity score of the cover singer was higher than the actual singer. One of the most typical examples, in Table. III. where S10001 is the label of singer A-Lin and S10060 is the label of vocalist Joyce Cheng for the Cantonese version of A-Lin's song in the test set, which accompanied by the exact same backing tack. In the system without separation, for the A-Lin's song fragment, Joyce Cheng gets higher similarities than A-Lin several times, while the results from the system with separation, the two singers' voices alone are not very similar, the system can easily determine that it is A-Lin's singing voice. In this case, the presence of the accompaniment apparantely becomes a distraction for the singer identification.

C. Studies on multiple songs per singer dataset

Multiple songs per singer dataset's experimental results confirm that the results using the separated data becomes better after reducing the proportion of songs with high similarity in song characteristics to the enrollment utterance in the test set. After the proportion of test data for which the accompaniment could be helpful was significantly reduced, the performance of the system with music separation was much better. The results of experiments are shown in Table. IIc.

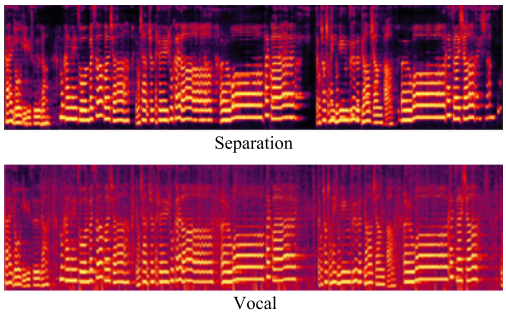

Fig. 5: The waveforms of *vocal* obtained by phase cancellation and *separation* obtained by spleeter.

D. Studies on backing tracks publicly released dataset

Besides, after the three experiments mentioned above, we believe the better performance because not only accompaniment as an assistant in identification systems, but also the impact of unstable and imperfect audio-source separation methods on vocal information. The instability and imperfection of the audio-source separation method cause distortion of the vocal information as illustrated in the spectrograms of *vocal* and *separation* in Fig. 5. The *vocal* we acquired in datasets processing does not have such a problem, they could be approximated as authentic vocal tracks. The experimental results for the four datasets as a test set are shown in Table. IV.

Accompaniment could be helpful. The results of this experiment corroborate our previous assumptions in several dimensions. The results of Exp 2. and Exp 3. show that

TABLE IV: Performance comparison of four types of pre-processed datasets accompaniment publicly released dataset.

Exp	Train set/ Enroll/ Eval	EER (%) in M1	EER (%) in M2
1	J1/ori/ori	10.65	2.41
2	J1/ori/acc	30.58	2.75
3	J1/acc/acc	20.08	2.36
4	J1/ori/voc	8.09	2.55
5	J1/ori/sep	17.18	7.56
6	J1/voc/voc	**7.58**	**2.02**
7	J1/sep/sep	18.50	6.69
8	J2/ori/ori	17.53	4.81
9	J2/ori/voc	11.49	**0.85**
10	J2/ori/sep	13.06	4.47
11	J2/voc/voc	**10.10**	1.01
12	J2/sep/sep	16.93	4.33

accuracy even when identifying only the accompaniment is not to be underestimated on the few samples test set. This explain the reason that on test sets such as JukeBox and Cover Song, where an artist's utterance is edited from no more than two songs, could obtain better performance without separation module, which does not stand on Multiple songs per singer dataset. This is not the essential purpose of singer identification and leads to worse generalization capability.

Removing accompaniment is better. A comparison of the results between *vocal* and *origin* in identification system, including Exp1&4&6 and Exp8&9&11, show that removing the accompaniment is extremely effective for singer identification in an optimal situation.

Separation method is unstable. Most importantly, each of the experiments on *separation* as a test set is significantly worse than on *vocal*, which suggests that the impairment and distortion of the extracted voice due to the music separation toolkit has a significant negative impact on identification accuracy. In this way, singer identification accuracy perhaps could indirectly measure the music separation performances in realistic utterances. But using phase cancellation to extract vocals requires both a stereo song and its counterpart stereo accompaniment, which is not realistic in practical applications. A more appropriately method of removing accompaniment for singer identification will be explored in future research.

VI. Conclusion

In this paper, we focus on audio-source separation for singer identification. Based on the experiments, we demonstrated that incorporating music separation in singer identification is necessary, even if the overall accuracy is slightly worse in some cases, but we also show that it has stronger generalization ability. We also proved that removing the accompaniment in an ideal condition is effective for the improvement of singer identification accuracy. But it must be mentioned here that there are inherent problems with existing vocal separation methods that result distortion the extracted vocals, which negatively affect the identification of the system. In future work, we are determined to explore more appropriate methods to make the singer identification system focus identifying of singers' vocals and obtaining both efficient and reliable results.

Acknowledge

This work was supported by the National Natural Science Foundation of China under Grant No. U1836219, and in part by the National Key RD Program of China, and the Institute for Guo Qiang of Tsinghua University under Grant No. 2019GQG0001, and the Cross-Media Intelligent Technology Project of Beijing National Research Center for Information Science and Technology (BNRist) under Grant No. BNR2019TD01022.

References

[1] Swe Zin Kalayar Khine, Tin Lay Nwe, and Haizhou Li. Exploring perceptual based timbre feature for singer identification. In *International Symposium on Computer Music Modeling and Retrieval*, pages 159–171. Springer, 2007.

[2] Jialie Shen, John Shepherd, Bin Cui, and Kian-Lee Tan. A novel framework for efficient automated singer identification in large music databases. *ACM Transactions on Information Systems (TOIS)*, 27(3):1–31, 2009.

[3] Xulong Zhang, Jiale Qian, Yi Yu, Yifu Sun, and Wei Li. Singer identification using deep timbre feature learning with knn-net. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3380–3384. IEEE, 2021.

[4] Shiteng Yang. *Statistical approaches for signal processing with application to automatic singer identification*. Rochester Institute of Technology, 2016.

[5] Lu Xing. Singer identification of pop music with singing-voice separation by rpca. 2017.

[6] Daniel PW Ellis. Classifying music audio with timbral and chroma features. 2007.

[7] Anurag Chowdhury, Austin Cozzo, and Arun Ross. Jukebox: A multilingual singer recognition dataset. *arXiv preprint arXiv:2008.03507*, 2020.

[8] Hamid Eghbal-Zadeh, Bernhard Lehner, Markus Schedl, and Gerhard Widmer. I-vectors for timbre-based music similarity and music artist classification. In *ISMIR*, pages 554–560, 2015.

[9] Deepali Y Loni and Shaila Subbaraman. Timbre-vibrato model for singer identification. In *Information and Communication Technology for Intelligent Systems*, pages 279–292. Springer, 2019.

[10] Xulong Zhang, Yongwei Gao, Yi Yu, and Wei Li. Music artist classification with wavenet classifier for raw waveform audio data. *arXiv preprint arXiv:2004.04371*, 2020.

[11] Bidisha Sharma, Rohan Kumar Das, and Haizhou Li. On the importance of audio-source separation for singer identification in polyphonic music. In *INTERSPEECH*, pages 2020–2024, 2019.

[12] Tsung-Han Hsieh, Kai-Hsiang Cheng, Zhe-Cheng Fan, Yu-Ching Yang, and Yi-Hsuan Yang. Addressing the confounds of accompaniments in singer identification. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2020.

[13] Dan Barry, Bob Lawlor, and Eugene Coyle. Sound source separation: Azimuth discrimination and resynthesis. In *7th International Conference on Digital Audio Effects, DAFX 04*, 2004.

[14] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154, 2020.

[15] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011.

[16] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. Deep neural network embeddings for text-independent speaker verification. In *Interspeech*, pages 999–1003, 2017.

[17] Daniel Garcia-Romero and Carol Y Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Twelfth annual conference of the international speech communication association*, 2011.