IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 January 2023, accepted 18 February 2023, date of publication 6 March 2023,
date of current version 18 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3253207

RESEARCH ARTICLE

Vocal92: Audio Dataset With a Cappella Solo Singing and Speech

ZHUO DENG AND RUOHUA ZHOU
Department of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
Corresponding author: Ruohua Zhou (zhouruohua@bucea.edu.cn)

ABSTRACT Singer recognition plays a vital role in music information retrieval systems. Most songs in the singer recognition system are mixed audios of music and voice. In contrast, there is a lack of labeled a cappella solo singing data suitable for singer recognition. Text-independent singer recognition systems successfully encode audio features such as voice pitch, intensity, and timbre to achieve good performance. Most such systems are trained and evaluated using data from music with accompaniment. However, due to the influence of background music, the performance of the singer recognition model was limited. Contrarily, a powerful singer identification system can be trained and evaluated using a cappella solo singing voice with a clear and broad range of qualities. There needs to be labeled clear singing data suitable for singer recognition research. To address this issue, we present Vocal92, a multivariate a cappella solo singing and speech audio dataset spanning around 146.73 hours sourced from volunteers. Furthermore, we use three models to construct the singer recognition baseline system. In experiments, the singer recognition model developed by a cappella solo singing data performs well in both single-mode and cross-modal verification data, significantly improving related works. The dataset is accessible to everyone at https://pan.baidu.com/s/1Pn62DHfal2OOZ_5JqgGBdQ with jnz5 as the validation code. For non-commercial use, the dataset is available free of charge at the IEEE DataPort (https://ieee-dataport.org/documents/vocal92-multimodal-audio-dataset-cappella-solo-singing-and-speech).

INDEX TERMS Singer recognition, deep learning, singing voice dataset.

I. INTRODUCTION

Singing is an exclusive sound art produced by a rhythmic combination of one or more vocal organs [1]. Singing voice has always been an exciting and abundant area of research. There is a variety of singing styles and techniques. Different singing styles can produce proper coordination and control of vocal organs such as the lungs, throat, pharynx, nose, and mouse [2]. To analyze and study different singing styles, we need to analyze the generalization process of singing sincerely. The analysis of singing sounds is challenging. It enables exploring various areas of study (e.g., song emotion analysis, lyric recognition, separation of singing sounds, classification of singing types, singer identification, and singer tracking in duet songs) only through songs [3]. Singer recognition is one of the leading research areas of the singing voice. Speech and singing are different expressive entities of human beings. Even though the organs involved in producing sound are the same, their extended frequency domain information is different. Speech is a natural use of vocal organs, but singing involves precise control of various organs. E. A. Zveglic proposed a comprehensive study of the relationship between speech and singing [1]. The research shows that singing stretches or lengthens acoustic features, while speech sacrifices acoustic features. Medeiros et al. invited three speakers and three singers to give a lecture and sing on a book of modern Brazilian literature [4]. They tested the hypothesis that singing is more stable than speech, particularly pitch and duration. Livingstone et al. observed that singing exhibits longer duration, higher pitch, and greater sound intensity than speech [5].

Singer recognition based on speaker recognition requires comparing two audio samples and evaluating whether the voices belong to the same person [6]. Most research in singer

The associate editor coordinating the review of this manuscript and approving it for publication was Sunil Karamchandani.

recognition focuses on modeling the features of singers from hybrid entities of music and voice [7]. However, accompanied songs only exhibit a limited range of the singer's possible dynamic vocal range [8]. As a result, such singer recognition systems can be less generalized to various singing styles and pronunciation effects. The voice of the a cappella solo singing is a speaking style that clearly demonstrates the singer's multiple features [9]. There are significant differences between the spoken language and a cappella solo of the same speaker. Apart from the perceived differences in pitch, intensity, and timbre, there are also differences in the physiological formation of sung speech [10], [11]. Different singing styles and languages further enrich the acoustic differences between spoken and sung sounds, bringing some challenges to the speaker recognition system [12]. Due to intentional voice modulation, singing voice increase intra-speaker variance and decrease inter-speaker variance, resulting in a broader acoustic spectrum, which is one of the main challenges in identifying a singer from a singing voice [13]. In addition, the presence of background music and choruses in existing music datasets increases the uncertainty of the task [8]. Thus, the ability of a singer recognition system to correctly evaluate whether it belongs to the same person in multiple songs can be used to assess its robustness.

Although many audio datasets exist, speech-singing modes audio datasets, especially those containing a cappella solo singing, still need to be improved. Therefore, a sizeable dataset containing a cappella solo singing and speech is necessary. In this study, we collected a new audio dataset, Vocal92, to study singer identification from speech and a cappella singing voice. We also explore the influence of taking singing data into training and testing on the generalization ability and robustness of the singer recognition model.

The structure of this paper is as follows. First, we compare Vocal92 with existing datasets, including existing work on singing sound analysis and application. Then, we record the collection and collation of Vocal92 and describe the structure of the dataset in detail. Finally, we construct a singer recognition baseline system to prove a broader range and richer feature information of the a cappella solo and achieve better performance, which also shows the practicability of this dataset.

II. RELATED WORK

Some early literature classified singing as a speech style and used speaker clustering algorithms to cluster it [5], [9], [16]. In another paper [14], the author used the singing voice for speaker recognition. However, cross-modal experiments were not committed in which models were trained on the speaking data and tested on the singing voice (and vice versa).

Those works were extended in [15] and [16] to evaluate cross-modal speaker recognition; moreover, the results of it needed to be more satisfactory. The JukeBox dataset expanded cross-modal experiments and facilitated speaker recognition research on singing voice data.

TABLE 1. A list of music datasets compared to Vocal92.

Dataset	Number of Samples	Number of Artists	Label	Raw Audio	Clean
Artist20[19]	1413	20	artists /group	Yes	NO
Vocalset[20]	3560	20	singer	Yes	Yes
Singing Voice Dataset[21]	Unknown	28	singer	Yes	Yes
JukeBox[22]	7000	936	singer	Yes	NO
Vocal92	4456	92	singer	Yes	Yes

A key reason for the lack of research on singer recognition is the need for adequate developmental and evaluation data [17], [18]. Although there have been some singing voice datasets in the field of singer recognition, they have yet to be able to evaluate the robust performance of singer recognition systems across modals.

The Artist20 dataset [19] contains 1413 songs from different albums by 20 European and American pop music artists or groups. The labels of artists/groups refer to associated musical groups or bands rather than individual singers so the features may vary considerably.

Vocalset [20] is a dataset of clean singing by nine female and eleven male professional singers. The dataset consists of 3560 wave files with a total of 10.1 hours of recorded audio ranging from 1s to 1 minute. Vocalset recorded vowels and various vocal techniques, such as scales, arpeggios, and long notes.

The singing voice dataset [21] contains over 70 significant recordings of Chinese opera performed by 28 professional and amateur singers. It is mainly opera, with no multilingual popular music, and the dataset can only be used as a test set since it is not large enough.

The JukeBox [22] dataset contains 467 hours of 16 kHz sampled singing audio data downloaded from the Internet Archive (IA). With a total of 936 different singers, 533 of whom are male. The singing voice dataset is annotated with singer, gender, and language labels for developing and evaluating speaker recognition methods. However, most of the singing audio data downloaded from the Internet has background accompaniment, which affects the accuracy of singer identification.

In the era of data-driven deep learning technology development, the lack of high-quality datasets with a cappella solo singing data has limited the progress of singer recognition research and applications [23].

In this paper, we propose a large vocal dataset of a cappella solo singing and speech that is annotated with labels such as singer, gender, age, and language. Table 1 shows a list of music datasets in comparison to Vocal92. It also illustrates the usefulness of Vocal92 by implementing a baseline system based on Vocal92 training, which can be used in areas such as singer recognition and song conversion. In the following

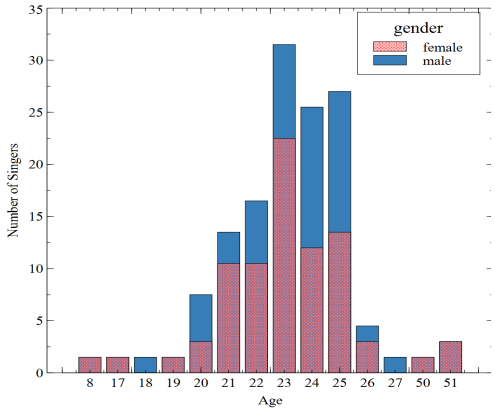

FIGURE 1. The distribution of singers' gender and age in the Vocal92 dataset.

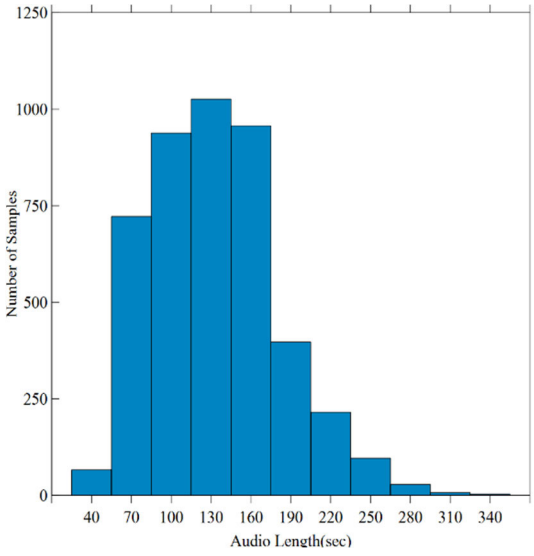

FIGURE 2. The distribution of audio length in the Vocal92 dataset.

few sections, we will describe this dataset in detail, the data collection process, several experimental scenarios, and analyze the performance of state-of-the-art singer recognition methods on this dataset.

III. DATA COLLECTION

A. SINGER RECRUITMENT

A call for participants for vocal recordings was posted online when offline activities were suspended because of the COVID-19 pandemic. We have the following requirements for volunteers:

1) Passion for music and a certain level of singing ability.
2) Recording clear audio through a smartphone or computer microphone in a quiet environment.
3) Choose at least 10 songs that you are good at singing, and each song should be no less than 2 minutes long.

We recruited 92 amateur singers to record voice data. The majority of participants were university graduates, along with some of their family and friends. The data set collected

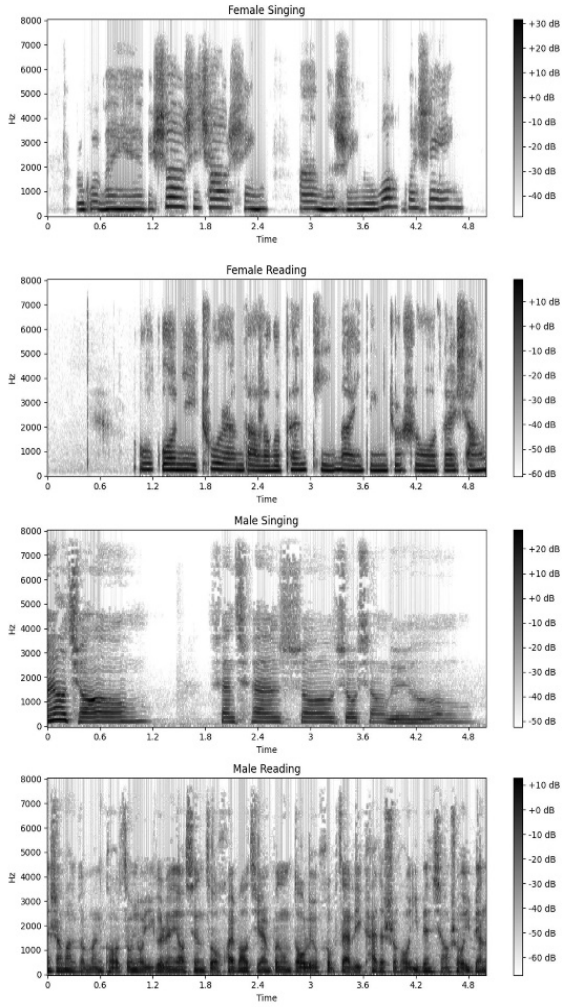

FIGURE 3. The narrowband spectrograms of four audio files.

consisted of popular music sung in Mandarin Chinese, Cantonese and English, in the same language as the song chosen by each volunteer. The recordings were collected over a period of 10 to 50 days. The gender and age distribution of the singers is shown in Figure 1.

B. RECORDING SETTINGS

Singing individuals recorded a cappella songs and read lyrics in a quiet setting using either a mobile phone or computer microphone. A minimum of ten songs were recorded by each performer, with each vocalist utilizing a separate audio file in various formats such as WAV, MP3, and M4A. Sampling rates for these recordings generally ranged from 48kHz and 44.1kHz.

The audio files were recorded in a 2-channel stereo format and subsequently converted to a single-channel, 16kHz sampling rate wav format after undergoing preprocessing.

The distribution of audio length is shown in Figure 2.

Dataset Organization:

The Vocal92 dataset includes 4453 a cappella solo recordings and lyric readings, totaling 146.73 hours of voice data.

TABLE 2. Dataset statistics of the Vocal92 dataset.

Dataset	Train	Test	All
# of Subjects	83	9	92
# of Male Subjects	32	4	36
# of Speech Samples	2139	90	2229
# of Singing Samples	2134	90	2224
# of Speech Hours	49.2989	2.2003	51.4992
# of Singing Hours	91.3434	3.8137	95.1568
Max # of Hours/Singer	2.9585	1.1586	2.9585
Min # of Hours/Singer	0.5128	0.2901	0.2901
Avg # of Hours/singer	1.6953	0.6682	1.5948

It consists of both singing and speech by 92 singers (36 male and 56 female) speaking in various languages. The data is organized in folders, with subfolders for each artist and song title, and includes audio files. To facilitate use in research, the dataset has been divided into train and test subsets. Additionally, the set of 92 singers in the dataset has been split into two subsets, as shown in Table 2.

We also selected audio from two volunteers singing and speaking for 5 seconds each and plotted the narrowband spectrograms, shown in Figure 3.

- Training set: Some singers with at least ten audio samples constitute the training set (83subjects). The training set consists of the speech training set, the singing training set, and the overall training set. This set is reserved for training singer recognition models.
- Test set: Some singers with at least two audio samples constitute the test set (9 subjects). The test set is separated into a speech test set, a singing test set, and a speech plus singing test set. This test set is reserved for evaluating trained singer recognition models on speech and singing voice data.

IV. METHODS

In this section, the proposed methodology will be discussed with the workflow that will be incorporated for identifying the singers. First, we discuss the architecture of our singer recognition baseline system.[1]

A. SINGER RECOGNITION SYSTEM

The singer recognition model consists of a training stage and a testing stage. During the training phase, three advanced neural networks are used to create embedding models for individual singers. In the testing phase, the similarity between the enroll audio and the test audio is measured using probabilistic linear discriminant analysis (PLDA) scoring and cosine similarity scoring to calculate the embedding similarity. The general architecture of our baseline is depicted in Figure 4. The feature extraction component converts the input audio into spectrogram features using the Speech Brain toolkit [24].

[1]The baseline system has provided at https://github.com/dengzhuo97/Vocal92-dataset

TABLE 3. Parameters for X-vector and ECAPA-TDNN model architecture.

Model	Layer	Layer context	Total context	Input x output
X-vector	Frame1	[-2, +2]	5	120 x 512
	Frame2	{-2, 0, +2}	9	1536 x 512
	Frame3	{-3, 0, +3}	15	1536 x 512
	Frame4	{0}	15	512 x 512
	Frame5	{0}	15	512 x 1500
	Stats pooling	[0,T]	T	1500T x 3000
	Segment6	{0}	T	3000 x 512
	Segment7	{0}	T	512 x 512
	sofmax	{0}	T	512 x N
ECAPA-TDNN	TDNN-ReLU1	[-2, +2]	5	C x 512
	SE-Res2Block1	{-2, 0, +2}	9	1536 x 512
	SE-Res2Block2	{-3, 0, +3}	15	1536 x 512
	SE-Res2Block3	{-4, 0, +4}	24	1536 x 512
	TDNN-ReLU2	{0}	24	1536 x 1536
	ASP	[0, T]	T	1536T x 3072
	FC	{0}	T	3072 x 192
	AAM-sofmax	{0}	T	192 x N

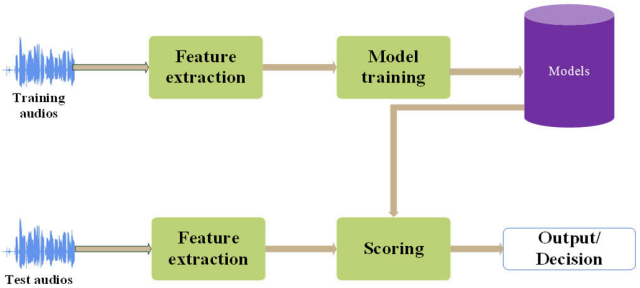

FIGURE 4. The general architecture of our baseline.

Our baseline system has been designed to facilitate the recognition of singers through the integration of these advanced algorithms and the neural network model.

B. MODEL ARCHITECTURE

In this study, we utilize three state-of-the-art systems to extract speaker embeddings: X-vector [25], emphasized channel attention, propagation and aggregation in time delay neural network (ECAPA-TDNN) [26], and ResNet50 for the singer identification task. Table 3 lists the detailed parameters for X-vector and ECAPA-TDNN architectures.

1) X-VECTOR-PLDA

The x-vector model [25] is a time delay neural network (TDNN) that aggregates variable length inputs across time

to create fixed-length representations capable of capturing speaker characteristics. Speaker embeddings are extracted from a bottleneck layer before the output layer. The method follows an end-to-end system that uses time-delayed DNNs to generate embeddings combined with similarity measures. It compares them through an independently trained classifier such as PLDA. Firstly, the time delay is used to extract short-time frame-level context. The statistical pooling layer aggregates over the input segments and calculates the mean and standard deviation. Finally, the singer is classified by DNN. The resulting segment-level singer embeddings are called x vectors. Zhang et al. [27] selected the X-vector model in the training phase of the singer recognition system and used PLDA to calculate the verification score in the testing phase.

In our approach, the x-vector of the training set is used to train the PLDA model [28], which is subsequently utilized for scoring. The parameters ϕ and Σ of the PLDA model are estimated from the training data. The method to estimate these two parameters is the classical EM algorithm iterative solution. In the test phase, we calculate whether two audio sounds are generated in the same speaker space regardless of intra-class spatial differences.

We use the log-likelihood ratio to calculate the score presented in Equation (1).

$$scor = log \frac{p\left(\eta_1, \eta_2 \mid H_s\right)}{p\left(\eta_1 \mid H_d\right) p\left(\eta_2 \mid H_d\right)}, \tag{1}$$

$\eta_1$and η_2 are the x-vectors of two sounds, respectively. The hypothesis that these two sounds come from the same space is Hs, and the hypothesis that they come from different Spaces is Hd. The $p(\eta_1, \eta_2|H_s)$ for two voices come from the same space likelihood function, $p(\eta_1|H_d)p(\eta_2|H_d)$ respectively to different space likelihood function. Calculating the log-likelihood ratio, we can measure how similar the two sounds are. The higher the score, the more likely it is that the two voices belong to the same speaker.

2) ECAPA-TDNN

Desplanques et al. [26] propose several improvements to the x-vector architecture. Specifically, they introduce the ECAPA-TDNN model, which includes 1-dimensional Res2Net modules with skip connections and squeeze excitation (SE) blocks to capture channel interdependencies and a channel-dependent self-attention mechanism that uses global context at the frame-level layers and the statistics pooling layer. Additionally, the ECAPA-TDNN model aggregates and propagates features across multiple layers.

To measure the similarity of two audio segments using the ECAPA-TDNN model, we employ the cosine similarity measure. Cosine similarity is a measure of similarity between two non-zero vectors in a multi-dimensional space, calculated as the cosine of the angle between the vectors. The functions can be mathematically presented in Equation (2):

$$Cos\left(\theta\right) = \frac{A \cdot B}{\|A\| \, \|B\|}, \tag{2}$$

Here, A, B are two non-zero vectors and$Cos\left(\theta\right)$ refers to the cosine similarity.

3) RESNET50

ResNet50, a 50-layer convolutional neural network architecture, was first introduced by Microsoft Research in 2015. The design of this network was motivated by the vanishing gradient problem, which affects the effectiveness of very deep neural networks in image recognition tasks. ResNet50 addresses this issue by incorporating residual connections that allow the network to bypass certain layers, thereby mitigating the vanishing gradient problem.

Residual connections in ResNet50 allow link connections to skip one or more layers and add their output directly to the output of the link, facilitating effective training of very deep networks. The architecture includes 1×1, 3×3 and 5×5 convolutional layers, as well as maximum pooling and average pooling layers. It also incorporates batch normalization and ReLU activation functions to further enhance performance. These design elements contribute to ResNet50's high performance in various image recognition tasks, including object detection, image classification, and semantic segmentation.

We introduce ResNet50 as the third neural network model for the baseline system and use the same cosine similarity as the ECAPA TDNN model for scoring.

C. METRICS

The equal error rate (EER) is a commonly used metric for evaluating the performance of singer recognition systems. It is defined as the point at which the false acceptance and rejection rates are equal. In addition to EER, the minimum detection cost function (minDCF) is also used as a secondary metric for comparing the confirmation thresholds of speaker recognition systems. This is represented by Equation (3). The minDCF is computed at a prior probability of 0.01 for the specified target speaker (P_{target}) with the cost of missed detection (C_{MISS}) and the cost of wrong detection ($C_{FalseAlarm}$) of 1.0.

$$\begin{aligned} C_{DCF} = {} & C_{MISS} \cdot P_{target} \cdot FRR \\ & + C_{FalseAlarm} \cdot \left(1 - P_{target}\right) \cdot FAR, \end{aligned} \tag{3}$$

V. EXPERIMENT AND RESULTS

A. EXPERIMENTAL SETUP

1) DATASET PARTITIONING

In each experiment described in this work, the entire dataset is randomly divided into a training set and an evaluation set with a 9:1 ratio. The evaluation set consists of an enrollment set and a test set, with each audio file from each singer in the evaluation set becoming enrollment data in succession. The remaining items will be reviewed.

The training set consists of the speaking training set, the singing training set, and the overall training set. Similarly, the evaluation data is separated into a speech evaluation set, a singing evaluation set, and a speech plus singing evaluation set.

2) TRAINING SETUP

During the training phase, we implement a random sampling strategy in which 3-second segments are randomly chosen from audio files and their starting times are selected on the fly. The ECAPA TDNN models in this study are trained using the Additive Angular Margin (AAM) loss [29] and the x-vector models are trained using the Negative Log Likelihood (NLL) [30] loss.

The input features for the x-vector models consist of 24-dimensional filterbanks with a frame length of 25ms, which are mean-normalized over a sliding window of up to 3 seconds. The input features for the ResNet50 model and the ECAPA-TDNN model are 80-dimensional filterbanks from a 25ms window.

The training set is divided into 90% and 10% for training and validation purposes. The validation set is randomly chosen from the training set. As in the test set, it is possible for different performers to sing the same song. To optimize the ECAPA-TDNN models, we utilized the Adam optimizer [31] with a learning rate of 0.0001 and a weight decay of 0.000002. If the validation loss does not change for two epochs, the learning rate is reduced by a factor of 0.3.

3) DATA AUGMENTATION

In this study, we also investigated suitable Data Augmentation (DA) strategies, i.e., the creation of moderately changed new data obtained from the original. As a result of DA, neural networks can learn new parameters and improve performance without overfitting. In addition to the training and evaluation datasets mentioned above, we use the MUSAN and RIRs datasets for noise augmentation. The former contains three types of noise, and the latter contains reverberation data in several different conditions. We used Speech Brain's augment model to add room impulse responses (RIRs) and noises and resample the audio at a slightly different rate to alter its speed. Models trained without DA required 100 epochs of training, while models trained with DA required 150.

4) ADAPTIVE SCORE NORMALIZATION

Adaptive score normalization means that the mean and variance are calculated for selecting voices from the impersonated speech set. Through adaptive normalization, each validation pair may use different impersonated speech sets. Adaptive score normalization selects the impersonated speech set according to specific rules, often using the top speech with the highest score of the registered speech or test speech. We performed adaptive normalization of the test scores of the ECAPA-TDNN model, which provided some optimization for the results of singer identification.

5) SINGLE OR MULTIPLE SPEAKING STYLES EXPERIMENTS

Both human and machine recognition performance degrades when the audio being evaluated is in a modality unfamiliar to the evaluator. Most of the previous speaker recognition systems use homologous data for experiments.

TABLE 4. Experimental results of a single speaking styles. (Divide each audio of the training set into 3 seconds, and input all the audio of the enrolled set and the test set).

Train	Enroll	Test	Models	EER/%	minDCF
Speech	Speech	Speech	X-vector-PLDA	0.8938	0.0540
			ECAPA-TDNN	0.9006	0.0471
			ResNet50	0.8732	0.0538
Singing	Singing	Singing	X-vector-PLDA	0.4718	0.0355
			ECAPA-TDNN	0.0000	0.0000
			ResNet50	0.0587	0.0308

We experimentally investigate the effect of cross-modal training and test data on speaker recognition systems' performance and generalization ability. In this paper, experiments were conducted on both unimodal and multiple speaking styles data to verify the effect of adding modality on speaker recognition. We found that the a cappella solo singing data performs better in the cross-modal experiments and generalizes better to the singer recognition system because of its more comprehensive range and variable timbre.

B. RESULTS

1) EXPERIMENTS ON SINGLE AUDIO MODALITIES OF EXPRESSION

The experimental results of X-vector-PLDA, ECAPA-TDNN, and ResNet50 models are shown in Table 4 when homogenous data are used for training and evaluating.

It is observed that a cappella solo singing data presents a more comprehensive representation of the singer due to its wider vocal range and more diverse features. When trained with a cappella solo singing data, the robustness of the models was significantly improved compared to using speech data, with the equal error rate of the x-vector-PLDA model decreasing to 0.4718%. Additionally, the recognition performance of the ECAPA-TDNN model was also satisfactory, correctly identifying the singing evaluation set. Upon comparison of the three models, it appears that the ECAPA-TDNN model exhibits a greater advantage on this dataset.

2) EXPERIMENTS ON MULTIPLE AUDIO MODALITIES OF EXPRESSION

The experimental results obtained when using different audio modalities of expression data for training and testing are presented in Table 5. These results demonstrate that the ECAPA-TDNN model trained on a cappella singing data exhibits a stronger migration ability and robustness among the different speaking styles of test data, achieving an equal error rate of 1.4723%. The use of all available training data led to a noteworthy enhancement in the performance of the cross-modal test set, as evidenced by the equal error rate (EER) of 1.1659% for the X-vector-PLDA model and an EER of 1.0496% for the ECAPA-TDNN model. The ResNet50

TABLE 5. Results of multiple speaking styles experiments (the training set audio in the experiments was divided into 3s, and the whole audio was input for both registration and test audio).

Train	Enroll	Test	Models	EER/%	minDCF
Speech	Speech	Speech	X-vector-PLDA	0.8938	0.0540
			ECAPA-TDNN	0.9006	0.0471
			ResNet50	0.8732	0.0538
Speech	Singing	Singing	X-vector-PLDA	8.5373	0.7423
			ECAPA-TDNN	8.7879	0.6659
			ResNet50	5.4996	0.2826
Speech	Singing	Speech	X-vector-PLDA	14.2857	0.8057
			ECAPA-TDNN	13.2361	0.9034
			ResNet50	8.5059	0.5120
Singing	Singing	Singing	X-vector-PLDA	0.4718	0.0355
			ECAPA-TDNN	0.0000	0.0000
			ResNet50	0.0587	0.0308
Singing	Speech	Speech	X-vector-PLDA	2.3418	0.1300
			ECAPA-TDNN	1.1360	0.0832
			ResNet50	0.4435	0.0540
Singing	Singing	Speech	X-vector-PLDA	2.4055	0.4897
			ECAPA-TDNN	1.4723	0.1588
			ResNet50	1.1588	0.1499
All	Speech	Speech	X-vector-PLDA	0.2217	0.0179
			ECAPA-TDNN	0.6789	0.0451
			ResNet50	0.4435	0.0181
All	Singing	Singing	X-vector-PLDA	0.4572	0.0166
			ECAPA-TDNN	0.0294	0.0142
			ResNet50	0.0440	0.0047
All	Speech	Singing	X-vector-PLDA	1.1659	0.0666
			ECAPA-TDNN	1.0496	0.0659
			ResNet50	0.9262	0.0764

model shows good performance when the training, enrollment, and test data are not in the same audio expression, especially when the training and test data are from different sources.

The results of the multiple speaking styles experiments indicate that the models trained with a cappella solo singing data exhibit superior generalization ability. According to our findings, models trained on clear singing song data exhibit superior generalization performance when evaluated on speech data. The X-Vector-PLDA model, ECAPA-TDNN model, and ResNet50 model have equal error rates of 2.3418%, 1.1360%, and 0.4435%, respectively. In contrast, the models trained on speech data do not perform well when evaluated on singing data. The ResNet50 model performs comparably to ECAPA-TDNN in terms of total performance, and both exhibit strong performance.

TABLE 6. Experimental results of data augmentation.

DA	Train	Enroll	Test	Models	EER/%	minDCF
YES	All	Speech	Speech	X-vector-PLDA	0.2354	0.0135
				ECAPA-TDNN	0.9006	0.0314
				ResNet50	0.7200	0.0426
	All	Singing	Singing	X-vector-PLDA	0.1297	0.0046
				ECAPA-TDNN	0.2359	0.0385
				ResNet50	0.2800	0.0095
	All	Singing	Speech	X-vector-PLDA	0.7281	0.0301
				ECAPA-TDNN	2.6381	0.3812
				ResNet50	1.8878	0.2863
NO	All	Speech	Speech	X-vector-PLDA	0.2217	0.0179
				ECAPA-TDNN	0.6789	0.0451
				ResNet50	0.4435	0.0181
	All	Singing	Singing	X-vector-PLDA	0.4572	0.0166
				ECAPA-TDNN	0.0294	0.0142
				ResNet50	0.0440	0.0047
	All	Speech	Singing	X-vector-PLDA	1.1659	0.0666
				ECAPA-TDNN	1.0496	0.0659
				ResNet50	0.9262	0.0764

TABLE 7. Experiment results of the X-Vector model using different back-ends for scoring.

Train	Enroll	Test	X-vector-PLDA		X-vector-Cos	
			EER/%	minDCF	EER/%	minDCF
All	Speech	Speech	0.2217	0.0179	1.6630	0.0785
All	Singing	Singing	0.4572	0.0166	1.6588	0.1446
All	Singing	Speech	1.1659	0.0666	3.4693	0.2844

3) DATA AUGMENT EXPERIMENTS

The results of our experiments revealed the effectiveness of using data augmentation in the X-vector-PLDA model. In the singer verification experiments reported in Table 6, we observed that the X-vector-PLDA models performed better with data augmentation. However, the use of data augmentation did not improve the performance of the ECAPA-TDNN and ResNet50model.

4) X-VECTOR MODEL WITH PLDA OR COSINE SIMILARITY FOR SCORING

Multiple scoring methods were utilized for the X-Vector model, as demonstrated in Table 7. The experimental results show that scoring after training a PLDA model has better robustness than directly calculating cosine similarity.

5) THE EFFECT OF AUDIO LENGTH

We also explored the effect of different audio lengths on the experimental results which are shown in Table 8. During the evaluation of 3-second audio segments from our

TABLE 8. The experimental results of different audio lengths in the models trained without data augment. (All training audios are divided into 3-second segments).

Train	Enroll	Test	X-vector-PLDA		ECAPA-TDNN	
			EER/%	minDCF	EER/%	minDCF
All	Speech/3s	Speech/3s	7.6483	0.5850	9.8382	0.5852
All	Singing/3s	Singing/3s	13.9782	0.8532	12.1202	0.6540
All	Singing/3s	Speech/3s	12.5000	0.8341	13.7467	0.7930
All	Speech/5s	Speech/5s	4.4893	0.2556	3.8105	0.3251
All	Singing/5s	Singing/5s	7.7115	0.5581	6.1634	0.4142
All	Singing/5s	Speech/5s	7.4492	0.7348	6.8294	0.5491
All	Speech/10s	Speech/10s	1.1223	0.0673	1.5658	0.1126
All	Singing/10s	Singing/10s	2.7131	0.3509	1.2383	0.2051
All	Singing/10s	Speech/10s	2.7190	0.3891	2.7332	0.2367
All	Speech/30s	Speech/30s	0.2217	0.0314	0.4435	0.0247
All	Singing/30s	Singing/30s	1.6661	0.1452	0.4718	0.0308
All	Singing/30s	Speech/30s	1.3489	0.1313	1.5956	0.0855
All	Speech/60s	Speech/60s	0.2217	0.0112	0.6789	0.0157
All	Singing/60s	Singing/60s	0.8404	0.0498	0.2506	0.0095
All	Singing/60s	Speech/60s	1.1588	0.0945	0.8383	0.4335
All	Speech/all	Speech/all	0.2217	0.0179	0.6789	0.0451
All	Singing/all	Singing/all	0.4572	0.0166	0.0294	0.0142
All	Singing/all	Speech/all	1.1659	0.0666	1.0496	0.0659
All	Speech/all	Speech/3s	4.1225	0.2439	4.7945	0.3188
All	Singing/all	Singing/3s	6.9891	0.7062	5.9274	0.3402
All	Singing/all	Speech/3s	5.2479	0.5458	5.9841	0.5829
All	Speech/all	Speech/5s	2.3555	0.1235	1.9052	0.1278
All	Singing/all	Singing/5s	4.1433	0.5004	2.9637	0.1707
All	Singing/all	Speech/5s	3.5927	0.2658	2.9587	0.2729
All	Speech/all	Speech/10s	1.4549	0.0628	1.2332	0.0797
All	Singing/all	Singing/10s	2.0127	0.2655	0.7078	0.0536
All	Singing/all	Speech/10s	1.4793	0.1324	1.5744	0.1566
All	Speech/all	Speech/30s	0.2354	0.0224	0.4503	0.0157
All	Singing/all	Singing/30s	1.0690	0.1096	0.3539	0.0213
All	Singing/all	Speech/30s	1.1588	0.0683	0.8666	0.0634
All	Speech/all	Speech/60s	0.2217	0.0135	0.6720	0.0135
All	Singing/all	Singing/60s	0.4865	0.0355	0.2359	0.0107
All	Singing/all	Speech/60s	1.1801	0.0599	0.9475	0.0571

test set, the x-vector-PLDA model and the ECAPA-TDNN model demonstrated equal error rates (EER) of 7.6483% and 9.8382%, respectively, on the speech evaluation set. When the audio segment length was increased to 5 seconds, the EERs of the two models on the speech evaluation set were 4.4893% and 3.8105%, respectively. The results indicated that longer audio segments resulted in better performance. When the audio length reached 10 seconds, both models exhibited significant improvements in all test results. On the other hand, the song evaluation set did not perform as well as the speech evaluation set when the test audio length was shorter. However, as the test audio length increased, the singing data experimental results improved. In comparison, the x-vector-PLDA model outperformed the ECAPA-TDNN model on the speech evaluation set.

These findings suggest that longer audio samples tend to have more rich and varied audio features, leading to improved experimental results in singer recognition tasks.

VI. CONCLUSION

We present the first audio dataset specifically focusing on a cappella solo singing and speech. Vocal92 consists of both singing and speech by 92 singers and represents a significant advancement in the field, filling a gap in the availability of multiple speaking styles audio datasets for singer recognition. The experimental results demonstrate the singer recognition models trained on singing data exhibit a more vital migration ability and robustness among the cross-modal test data. These findings suggest that singing data may contain a more exhaustive range and features, such as timbre and pitch, which contribute to better model performance. The Vocal92 dataset will also be a valuable resource for music information retrieval, singer recognition, and speaker recognition.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the volunteers who generously contributed to the Vocal92 dataset, without whom this audio dataset would not have been possible. They would also like to thank the anonymous reviewers for their valuable comments and suggestions that helped improve this article. Their participation and support are greatly appreciated.

REFERENCES

[1] E. A. Zveglic, "Speech and singing," in *Recognizing and Treating Breathing Disorders*, 2013, pp. 203–214.

[2] F. Simonetta, S. Ntalampiras, and F. Avanzini, "Multimodal music information processing and retrieval: Survey and future challenges," in *Proc. Int. Workshop Multilayer Music Represent. Process. (MMRP)*, Jan. 2019, pp. 10–18.

[3] E. J. Humphrey, S. Reddy, P. Seetharaman, A. Kumar, R. M. Bittner, A. Demetriou, S. Gulati, A. Jansson, T. Jehan, B. Lehner, A. Krupse, and L. Yang, "An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music," *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 82–94, Jan. 2019.

[4] B. Medeiros and J. Cabral, "Acoustic distinctions between speech and singing: Is singing acoustically more stable than speech?" in *Proc. 9th Int. Conf. Speech Prosody*, Jun. 2018, pp. 542–546.

[5] S. R. Livingstone, K. Peck, and F. A. Russo, "Acoustic differences in the speaking and singing voice," in *Proc. Meetings Acoust.*, 2013, pp. 1–6, doi: 10.1121/1.4799460.

[6] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, 2010.
[7] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Netw.*, vol. 140, pp. 65–99, Aug. 2021.
[8] B. Sharma, R. K. Das, and H. Li, "On the importance of audio-source separation for singer identification in polyphonic music," in *Proc. Interspeech*, Sep. 2019, pp. 2020–2024.
[9] E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. S. Kajarekar, H. Jameel, C. Richey, and F. Goodman, "Effects of vocal effort and speaking style on text-independent speaker verification," in *Proc. Interspeech*, Sep. 2008, pp. 609–612.
[10] X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Singer identification for metaverse with timbral and middle-level perceptual features," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–7.
[11] A. Chowdhury and A. Ross, "Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1616–1629, 2019, doi: 10.1109/TIFS.2019.2941773.
[12] J. H. L. Hansen, M. Bokshi, and S. Khorram, "Speech variability: A cross-language study on acoustic variations of speaking versus untrained singing," *J. Acoust. Soc. Amer.*, vol. 148, no. 2, pp. 829–844, Aug. 2020.
[13] M. Mehrabani and J. H. L. Hansen, "Singing speaker clustering based on subspace learning in the GMM mean supervector space," *Speech Commun.*, vol. 55, no. 5, pp. 653–666, Jun. 2013, doi: 10.1016/j.specom.2012.11.001.
[14] H. A. Patil, M. C. Madhavi, and N. H. Chhayani, "Person recognition using humming, singing and speech," in *Proc. Int. Conf. Asian Lang. Process.*, 2012, pp. 149–152, doi: 10.1109/IALP.2012.58.
[15] N. H. Chhayani and H. A. Patil, "Development of corpora for person recognition using humming, singing and speech," in *Proc. Int. Conf. Oriental COCOSDA Conf. Asian Spoken Lang. Res. Eval.*, Nov. 2013, pp. 1–6, doi: 10.1109/ICSDA.2013.6709863.
[16] S. Biswas and S. S. Solanki, "Speaker recognition: An enhanced approach to identify singer voice using neural network," *Int. J. Speech Technol.*, vol. 24, pp. 1–13, Mar. 2020.
[17] A. Chowdhury, A. Cozzo, and A. Ross, "Domain adaptation for speaker recognition in singing and spoken voice," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7192–7196, doi: 10.1109/ICASSP43922.2022.9746111.
[18] S. Hu, B. Liang, Z. Chen, X. Lu, E. Zhao, and S. Lui, "Large-scale singer recognition using deep metric learning: An experimental study," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–6, doi: 10.1109/IJCNN52387.2021.9533911.
[19] D. P. Ellis, "Classifying music audio with timbral and chroma features," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, vol. 7, 2007, pp. 339–340.
[20] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, "VocalSet: A singing voice dataset," in *Proc. ISMIR*, 2018, pp. 468–474.
[21] D. A. Black, M. Li, and M. Tian, "Automatic identification of emotional cues in Chinese opera singing," in *Proc. Int. Conf. Music Perception Cognition Conf. Asian-Pacific Soc. Cogn. Sci. Music*, 2014, pp. 250–255.
[22] A. Chowdhury, A. Cozzo, and A. Ross, "JukeBox: A multilingual singer recognition dataset," in *Proc. Interspeech*, 2020, pp. 2267–2271.
[23] E. Gómez, M. Blaauw, J. Bonada, P. Chandna, and H. Cuesta, "Deep learning for singing processing: Achievements, challenges and impact on singers and listeners," 2018, *arXiv:1807.03046*.
[24] M. Ravanelli et al., "SpeechBrain: A general-purpose speech toolkit," 2021, *arXiv:2106.04624*.
[25] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5329–5333, doi: 10.1109/ICASSP.2018.8461375.
[26] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, Oct. 2020, pp. 3830–3834.
[27] Y. Zhang, Y. Xiao, W. Q. Zhang, X. Tan, L. Lei, and S. Wang, "Mixing or extracting? Further exploring necessity of music separation for singer identification," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2021, pp. 1127–1132.
[28] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, Jul. 2010.
[29] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.
[30] H. Yao, D. L. Zhu, B. Jiang, and P. Yu, "Negative log likelihood ratio loss for deep neural network classification," in *Proc. Future Tech. Conf. (FTC)*, 2019, pp. 276–282.
[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

ZHUO DENG is currently pursuing the master's degree in control science and engineering with Beijing University of Civil Engineering and Architecture. His research interests include music information retrieval and singer recognition.

RUOHUA ZHOU received the bachelor's degree in microelectronics from Beijing Institute of Technology, Beijing, China, in 1994, the master's degree in microelectronics from the Institute of Microelectronics, Chinese Academy of Sciences, in 1997, and the Ph.D. degree in signal processing from Swiss Federal Institute of Technology Lausanne, in 2006.

He was a Professor with the Institute of Acoustics, Chinese Academy of Sciences, and an academic leader in the fields of speaker recognition, language recognition, and music signal processing. He is currently a Professor with the School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture. He was selected to join the "Hundred Talents Program" of Chinese Academy of Sciences. He has presided more than 20 projects, including the Strategic Science and Technology Pilot Project of the Chinese Academy of Sciences, the National Natural Science Foundation, and the National Defense Science and Technology Innovation Fund. He has published more than 50 academic articles.