KunquDB: An Attempt for Speaker Verification in the Chinese Opera Scenario

Huali Zhou[1,2], Yuke Lin[1,2], Dong Liu[2], and Ming Li[1,2]*

[1] School of Computer Science, Wuhan University, Wuhan, China
[2] Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Data Science Research Center, Duke Kunshan University, Kunshan, China
hualizhou@whu.edu.cn, {yuke.lin,dong.liu,ming.li369}@dukekunshan.edu.cn

Abstract. This work aims to promote Chinese opera research in both musical and speech domains, with a primary focus on overcoming the data limitations. We introduce KunquDB, https://hualizhou167.github.io/KunquDB, a relatively large-scale, well-annotated audio-visual dataset comprising 339 speakers and 128 hours of content. Originating from the Kunqu Opera Art Canon (*Kunqu yishu dadian*), KunquDB is meticulously structured by dialogue lines, providing explicit annotations including character names, speaker names, gender information, vocal manner classifications, and accompanied by preliminary text transcriptions. KunquDB provides a versatile foundation for role-centric acoustic studies and advancements in speech-related research, including Automatic Speaker Verification (ASV). Beyond enriching opera research, this dataset bridges the gap between artistic expression and technological innovation. Pioneering the exploration of ASV in Chinese opera, we construct four test trials considering two distinct vocal manners in opera voices: stage speech (***ST***) and singing (***S***). Implementing domain adaptation methods effectively mitigates domain mismatches induced by these vocal manner variations while there is still room for further improvement as a benchmark.

Keywords: Kunqu Opera · Dataset · Multi-modal · Speaker verification · Cross-domain.

1 Introduction

Chinese opera, or *Xiqu*, is a distinguishable and traditional art form that has gained worldwide recognition. Kunqu Opera, Beijing Opera, and Cantonese Opera have been proclaimed World Intangible Cultural Heritage, highlighting their exceptional artistic contributions and rich cultural heritage. Chinese opera is a confluence of song, speech, mime, dance, and acrobatics, bound together by theatrical conventions that differ significantly from Western opera [20].

As a distinctive form of performing arts, Chinese opera diverges from conventional speech and typical singing. In the realm of speech research, opera provides a distinctive experimental ground, given its intricate fusion of speech, music, and theatrical elements. The multifaceted acoustic expressions within opera voices create an exceptional context

* Corresponding author. E-mail: ming.li369@duke.edu

for in-depth exploration in speech research. Regardless, previous research on Chinese opera has predominantly stemmed from musical and literary perspectives, relying on traditional methodologies rather than integrating state-of-the-art technical tools. The absence of automated deep-learning tools has led to a heavy reliance on manual data pipelines for collecting and annotating Chinese opera datasets. Consequently, existing opera datasets [5,16,12,6] face limitations in terms of scale and annotation richness, typically covering only a few hours [16,6] and providing genre information exclusively [6]. In contrast to the comprehensive annotations provided in speech and singing datasets, which include speaker labels, text transcriptions, phoneme-level durations, and pitch information, existing Chinese opera datasets lack comparable richness.

The scarcity of detailed annotations poses a significant obstacle for numerous research tasks on opera data. This obstacle is particularly pronounced for tasks requiring comprehensive annotations, including automatic speaker recognition for speaker label prediction, Automatic Speech Recognition (ASR) for text transcription retrieval, speaker diarization for role detection, as well as speech and singing voice synthesis. Meanwhile, speech-related research predominantly focuses on conventional speech. Existing open-source models designed for various speech tasks, such as speaker diarization, exhibit inadequate robustness when applied to opera data. The complex acoustic characteristics in opera voices provide a diverse testing ground for evaluating the robustness of speech models. The absence of automated tools further obstructs large-scale data collection and cleaning, restricting access to diverse and abundant datasets. This dilemma creates a cycle that impedes progress in data availability, hindering the development of advanced tools for digitizing opera research.

In response to the challenges posed by insufficient data and limitations of existing models in the field of Chinese opera, our primary objective is to create a symbiotic relationship between data and models. To achieve this, we present a comprehensive and publicly accessible audio-visual dataset characterized by its richness and scale. This resource is designed to lay the groundwork for developing specialized automated tools applicable to Chinese opera, thereby facilitating advancements in the study of this art form. Narrowing down from the landscape of Chinese traditional opera, we focus on one exquisite domain, Kunqu Opera. Reputed as the mother of Chinese operas, Kunqu Opera boasts a history spanning over 600 years [10], giving rise to numerous operas, including Beijing Opera. In alignment with [5], we selectively choose classic and authoritative audio-visual materials sourced from the Kunqu Opera Art Canon (*Kunqu yishu dadian*)[3] [39] to ensure both quantity and quality. The source video undergoes sentence-level segmentation, generating preliminary text transcriptions. Subsequently, we proceed with speaker annotation and explicitly categorize each utterance as either stage speech (***ST***) or singing (***S***) based on vocal manner. Ultimately, KunquDB[4], the curated audio-visual dataset comprises 339 performers, totals approximately 128 hours, with stage speech and singing voices each constituting about half of the dataset. As an

[3] **Note**: After purchasing the book, we negotiated with the publisher and secured their authorization for its utilization in Kunqu Opera research. The publisher explicitly stated that the book's digital resource can be employed solely for scholarly or research endeavors upon the approval of the publisher. It may not be illegally disseminated or used for commercial purposes.

[4] https://hualizhou167.github.io/KunquDB

audio-visual dataset, it is applicable in various scenarios, including ASV, ASR, speaker diarization, singing voice synthesis, person re-identification and multi-modal understanding.

Building on KunquDB, we investigate automatic speaker verification within the Kunqu Opera context. We aim to provide insights that enhance subsequent synthesis efforts, accommodating variations in role types and vocal manners. Speaker verification in Kunqu Opera bears similarities to the task in [2], involving speech from interviews (typically calm and quiet) and speech from movies (with varying emotion and background noise) across different domains. This yields a cross-domain speaker verification challenge induced by vocal manner, an internal factor of the speaker [23]. To tackle cross-domain issues, we implement domain adversarial training, leveraging domain prediction to obtain speaker-discriminative and domain-invariant representations. Furthermore, we employ the batchwise Siamese training strategy to maintain consistency across different vocal manners for the same speaker. Experimental results validate the efficacy of the domain adaptation methods.

Our main contributions are summarized as follows:

– We curate KunquDB[4], a comprehensive audio-visual dataset specifically tailored for Kunqu Opera. Its large scale effectively mitigates data shortages and fosters a positive feedback loop between data and tool models.

– To the best of our knowledge, we are the first to explore ASV within Chinese opera, addressing mismatches across stage speech and singing. The implementation of domain adaptation methods sets a benchmark for future research.

2 Vocal Distinctions in Chinese Opera Versus Speech

The aural aspect of Chinese traditional opera significantly differs from ordinary spoken and contemporary singing, including textual structure, pronunciation, intonation, vocal manner, and overall expressive forms. (1) The textual dimension of Chinese traditional opera involves two types: song lyrics (*changci*) for expressing emotions and stage speech (*nianbai*) for advancing the narrative [40]. Within the text, two linguistic levels emerge: classical Chinese (*wenyan wen*), an archaic written language, and vernacular (*baihua*), which includes standard spoken Mandarin or regional dialects with distinct phonetic variations. (2) From a melodic perspective, Chinese opera draw its musical compositions from a pre-existing repertoire of tunes. Unlike Western opera, where a designated "composer" is assigned, in Chinese opera, the scriptwriter selects tunes deemed suitable for the dramatic context from the repertoire and crafts the accompanying text. Musical notation is absent; instead, the script specifies tunes by name, with the text intended to be sung accordingly [47]. Notably, stage speech and singing exhibit considerably higher Equivalent Sound Levels (Leq) compared to regular speech [10]. (3) From a vocalization standpoint, Chinese opera utilizes two vocal techniques: "false-voice" (*jiasangzi*), executed in falsetto, and "true-voice" (*zhensangzi*), produced by vocal cord vibration. Falsetto serves various purposes. Firstly, male actors, exemplified by renowned figures like Mei Lanfang, portray female characters in Chinese opera, employing falsetto to imitate the female voice [17]. Secondly, falsetto is believed to ideally produce essential, extended sounds pronounced with a nearly closed mouth [40].

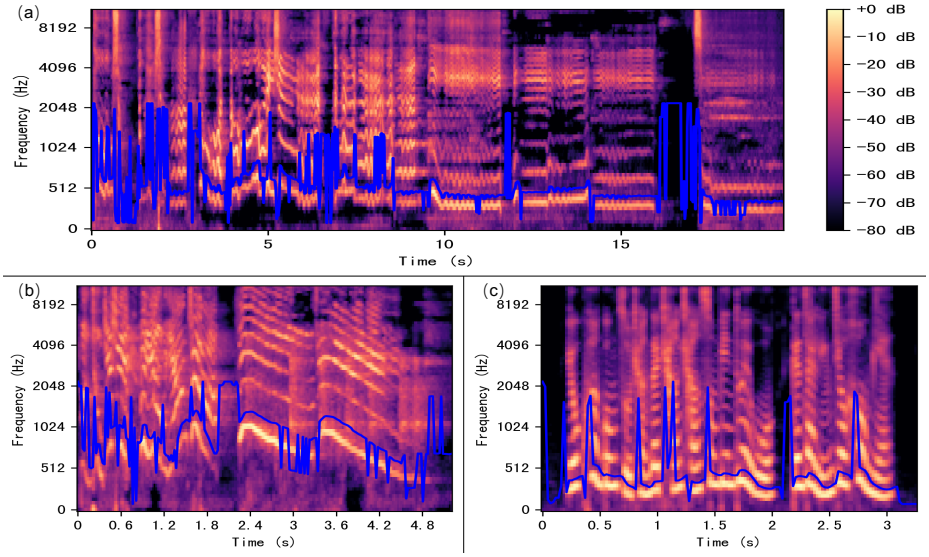

Fig. 1. Mel spectrograms with overlaid pitch contours for singing (**a**), stage speech (**b**), and regular speech (**c**).

Fig. 1 displays Mel spectrograms with overlaid pitch contours for randomly selected utterances representing singing, stage speech, and regular speech (from an external speech dataset). Singing and stage speech consistently exhibit higher frequency compared to regular speech. Moreover, singing showcases more dynamic pitch variation than stage speech, highlighting two distinct acoustic characteristics in Chinese opera.

3 Related Works

3.1 Chinese Opera

Open-source datasets for Chinese opera remain limited. While [16] and [6] propose datasets for opera genre and Cantonese singing genre classification, respectively, these datasets are not publicly accessible. Due to the lack of publicly available corpora, [48,19,45] targeting opera genre classification, collect data individually for personalized experiments. Typically, these datasets consist of individual instances structured as audio files paired with corresponding Chinese opera genre labels. On the other hand, open-access datasets have driven advancements in academic research. For example, the CompMusic Beijing Opera corpus proposed in [32] aids [35] in acquiring Beijing Opera percussion patterns for transcription and recognition. Similarly, the unaccompanied singing data released by [1] provides the foundation for [44] to analyze pitch histograms and vibrato statistics in Beijing Opera singing.

Due to the lack of automatic tools for data collection and cleaning, existing opera datasets [5,16,12,6] are limited in scale and annotation richness. Typically spanning only a few hours [16,6] and offering plain labels [6], they are insufficient for downstream recognition tasks like ASV and singing speech recognition in the field of Chinese opera. In the intersection of opera and speech-related research, most efforts are

focused on synthesis, with reliance on the only publicly accessible, well-annotated yet small-scale dataset, "Jingju a cappella singing" [12]. It serves as the basis for subsequent neural network-based opera synthesis by [41,42,49,25]. While [41] and [42] pioneer neural network-based synthesis using the DurIAN [46] framework, [49] introduces OperaSinger, based on the FastSpeech2 [30] framework, exploring novel data augmentations within this small-scale dataset [12]. In a related vein, [25] attempts to transfer popular singers' timbre to Chinese opera using the VITS [18] model with the same dataset [12].

3.2 Automatic Speaker Verification

Automatic Speaker Verification (ASV) aims to verify whether a given utterance (test utterance) matches the claimed identity by comparing it with the speaker's known utterance (enrollment utterance). The rise of DNNs in recent years has triggered the evolution of ASV systems from traditional probabilistic models [31,7] to deep embedding models [34,9]. A typical DNN-based ASV architecture consists of key components, including: (i) **neural network backbone** [37,9,13] (encoder), (ii) **pooling layer** [38,33,24] for temporal aggregation, (iii) **loss function** [36,8] for training optimization, (iv) **scoring strategy** [26] for assessing similarity between embeddings.

The neural network backbone, as the encoder, extracts frame-level features from the input utterance. This backbone has evolved from architectures like 2D Convolutional Neural Networks (CNNs) [37], Time Delay Neural Networks (TDNNs) [9], and Transformers [13]. Currently, 2D CNNs with ResNet [14] are the most widely adopted. The pooling layer aggregates frame-level features into a fixed-length, utterance-level representation, which is then projected linearly to generate the speaker embedding. Common temporal aggregation techniques include average pooling [38], statistical pooling [33] and attentive pooling [24]. The loss function is the optimized objective during training, such as the Additive Margin Softmax (AM-Softmax) [36] and ArcFace [8]. The scoring strategy, or the back-end model, measures the similarity between enrollment and test utterance embeddings for verification. Typically, cosine similarity or Probabilistic Linear Discriminant Analysis (PLDA) [26] are utilized.

Despite significant progress in ASV, speaker embeddings' robustness falters with domain shifts, facing challenges from real-world variations [23], resulting in performance degradation. For extrinsic factors, [3] and [27] target noise and far-field conditions for more robust voiceprint representation. Addressing internal factors, [29] and [2] investigate cross-age and diverse emotional scenes, respectively, to further enhance the robustness.

4 KunquDB Dataset[4]

To obtain authentic singing data for Kunqu Opera and ensure an ample dataset, we leverage audio-visual materials from the authoritative Kunqu Opera Art Canon (*Kunqu yishu dadian*)[3] [39] as reliable sources. The source videos in this collection [39] contain credits, dialogue lines, and information about vocal manner categories (***ST*** or ***S***), all of which are hard-coded directly or indirectly.

4.1 Overall: QAs about KunquDB

What is KunquDB? KunquDB is a Kunqu Opera audio-visual dataset derived from videos featuring manual annotations for opera character names, speaker identity (ID) labels, gender information, singing/stage speech category labels, and preliminary text transcriptions.

Why is Manual Labeling Required? Due to the nature of Kun Opera performances, where the entire stage is often captured rather than close-ups of characters, human faces occupy limited space in the frame. Moreover, performers' heavy makeup and theatrical costumes further obscure facial features, particularly the waist-length beards (*rankou*) [40] worn by male characters typically completely conceal their mouths. Consequently, conventional pipelines, as used in [23,21], involving face detection, tracking, verification, and audio-video synchronization for mouth movement and speech, are unsuitable for these opera videos.

How to Get KunquDB? The book [39] purchase grants access to the digital source video data in a supplementary disc. It is the user's responsibility to get the approval from the publisher to conduct research for non-commercial purposes. We provide annotated data, including segment start and end timestamps, along with associated information, such as character names, speaker names, and preliminary text transcriptions. The open-source annotations and processing scripts can be accessed and downloaded online[4].

4.2 Data Collection Pipeline

Step 1: Video Segmentation We utilize VideoSubFinder[5], in conjunction with PaddleOCR[6] to extract hardcoded subtitles from source videos, yielding timestamps for each dialogue line and corresponding text transcriptions. Using ffmpeg[7], we then segment the videos into clips based on the acquired timestamps, resulting in individual video clips for each dialogue line.

Step 2: Manual Labeling The manual annotation process includes categorizing vocal manner and active speaker annotations. Vocal manner annotation is straightforward, with stage speech and singing categorized based on the font style in the original video subtitles. Active speaker annotation is detailed below and is divided into (i) discriminative speaker tag, (ii) tag-character annotation, and (iii) character-performer mapping based on each play. Eventually, the dataset is structured per dialogue line, encompassing all lines delivered by each performer across different plays.

i We recruit 20 graduate students to assign active speaker tags, each annotating an average of 8.5 hours of source videos. Participants use XnView MP[8] software to tag active speakers for each line while watching the complete source video. They adhere to a naming format like *spk_01* to ensure consistency and avoid repetition within each play. Overlapping speech segments are instructed to be discarded.

[5] https://sourceforge.net/projects/videosubfinder
[6] https://github.com/PaddlePaddle/PaddleOCR
[7] https://ffmpeg.org
[8] https://www.xnview.com/en/xnviewmp

Table 1. Dataset statistics for KunquDB

Types of Utterances	Stage Speech	Singing
# of speakers	288 + 50	288 + 1
# of videos	339 + 5	339 + 2
# of utterances	60066	17902
# of hours	67.46	60.88
Avg # of videos per speaker	3	3
Avg # of utterances per speaker	178	62
Avg length of utterances(s)	4.04	12.24

Table 2. Training and test data split

		#Speakers	#Utterances
Training	Stage Speech	200	55889
	Singing	200	16941
Test	Stage Speech	88 + 50	4177
	Singing	88 + 1	961

ii Match the active speaker tags akin to *spk_01* obtained in i with the corresponding characters in each play.
iii Extract character-performer mapping by digitizing the embedded credits in source videos.

Step 3: Extract Audio from Video Initially, we use ffmpeg[7] to extract 48kHz stereo audio from video segments, then Spleeter [15] isolates background music, and finally ffmpeg[7] downsamples the audio to mono-channel at 16kHz.

Step 4: Assessment and Recheck We extract speaker embeddings for individual utterances, using WeSpeaker's [37] ResNet34-based model pretrained on Cn-Celeb [11]. Then, we compute average embeddings for each speaker in each category and assess the cosine similarity between each utterance's embedding and the corresponding average. Utterances with a similarity score below the threshold of 0.4 undergo manual review.

4.3 Dataset Statistics

Table 1 summarizes key statistics for the KunquDB dataset, differentiating stage speech (***ST***) and singing (***S***) categories. The dataset contains 60, 066 ***ST*** utterances and 17, 902 ***S*** utterances, contributed by 288 speakers for both ***ST*** and ***S*** data, 50 exclusively providing ***ST*** data, and 1 exclusively offering ***S*** data. Additionally, there are 339 videos featuring both ***ST*** and ***S***, along with 5 exclusively for ***ST*** and 2 for ***S***. Fig. 2 visually represents the distribution of utterance lengths and speakers enacting role types.

4.4 Split: Training and Test

We divide speakers based on their total number of utterances, with the initial 200 individuals allocated to the training set and the remaining 139 to the test set. See Table 2 for details.

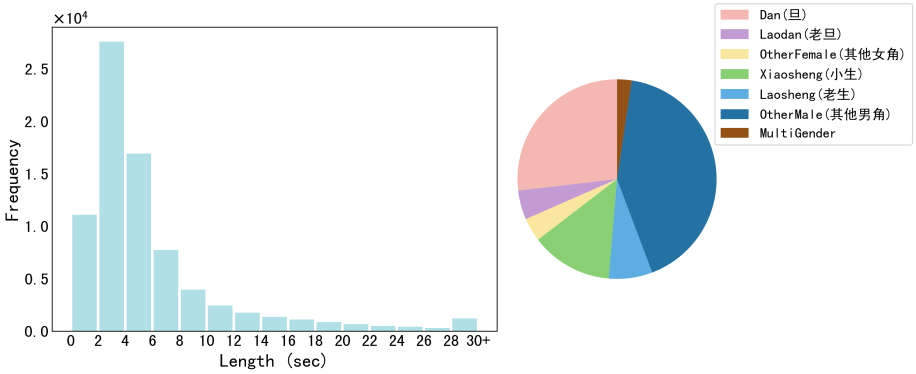

Fig. 2. Left: Histogram of utterance lengths in the dataset. **Right:** Distribution of speaker role type information. The legend indicates the role type performed by speakers throughout the dataset. ***Dan*** for young female characters, ***LaoDan*** for old female characters, ***OtherFemale*** for additional female characters; ***XiaoSheng*** for young male characters, ***LaoSheng*** for old male characters, ***OtherMale*** for additional male characters; and ***MultiGender*** means speakers portraying characters of both genders.

4.5 Trial Construction

When generating test trials for speaker verification experiments, we adopt a consistent procedure for each utterance, randomly selecting five positive and five negative samples. Investigating four trial scenarios considering two vocal manners (stage speech and singing), we have:

– Undifferentiated Trial: No distinction between enrollment and test utterance regarding vocal manner categories; samples are randomly chosen from either stage speech or singing.

– Stage Speech Domain Trial: Both enrollment and test utterances are from the stage speech category.

– Singing Domain Trial: Both enrollment and test utterances are from the singing category.

– Cross-domain Trial: Enrollment is from singing, while test utterances are from the stage speech category.

5 Learning Domain-invariant Speaker Embeddings

5.1 Domain Discrepancy Adversarial Learning

As discussed in Section 3.2, the speaker ID embedding extractor comprises a feature encoder, pooling, and linear layer. Traditionally, it is assumed that this extractor, depicted by the pink dashed box in Fig. 3, exclusively captures acoustic features defining speaker identity, denoted by the equation $\mathbf{f} = \mathbf{f}_{\mathbf{id}}$. However, it may inadvertently conflate identity-specific traits with variations from intrinsic factors like vocal mannerisms, formalized as Equation 1, where $\mathbf{f}$ denotes the extracted features, $\mathbf{f}_{\mathbf{id}}$ refers to the identity-specific features, and $\mathbf{f}_{\mathbf{domain}}$ represents features associated with vocal manners.

$$\mathbf{f} = \mathbf{f}_{\mathbf{id}} + \mathbf{f}_{\mathbf{domain}} \tag{1}$$

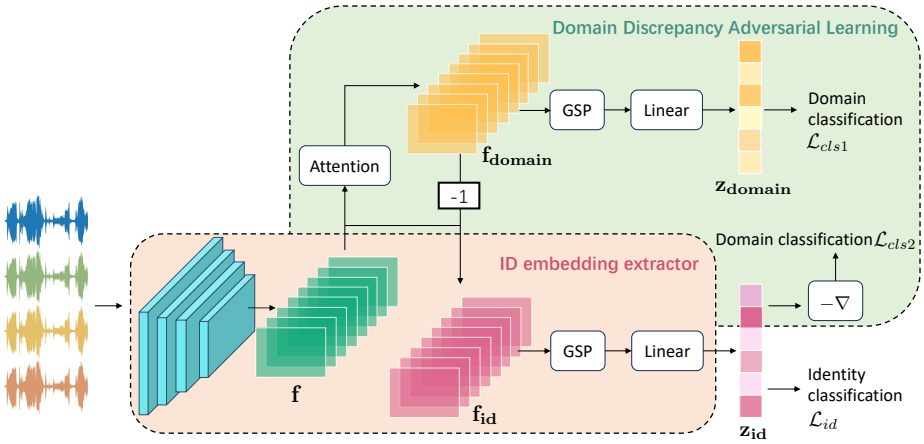

Fig. 3. Schematic of the DDAL framework. The pink dashed box outlines the identity embedding extractor; the green dashed box highlights the core components of the DDAL mechanism.

Borrowing insights from [29], we implement an optimized multi-task paradigm called Domain Discrepancy Adversarial Learning (DDAL), as illustrated in Fig. 3, to isolate domain-specific variables from speaker embeddings. This framework integrates speaker identity verification, domain classification, and domain adversarial training. Diverging from [29], we disentangle domain characteristics at the feature map layer instead of the abstract embedding space. This early disentanglement capitalizes on the richer domain-specific details in the feature map layer, facilitating a cleaner separation and enhancing verification precision across domains.

We leverage an attention mechanism to disentangle domain-related features $\mathbf{f_{domain}}$ induced by different vocal manners from the feature map $\mathbf{f}$ extracted by the backbone model. Next, we refine speaker-specific features, $\mathbf{f_{id}}$, by filtering out $\mathbf{f_{domain}}$. Following this, both $\mathbf{f_{domain}}$ and $\mathbf{f_{id}}$ undergo pooling and fully connected layers, producing the domain embedding $\mathbf{z_{domain}}$ for domain classification and speaker ID embedding $\mathbf{z_{id}}$ for speaker classification. Further, we employ a gradient reversal layer (GRL) before an auxiliary domain classifier to eliminate domain influence from $\mathbf{z_{id}}$ through adversarial learning.

Equation 2 defines the composite loss function, comprising the standard identity loss $\mathcal{L}_{id}$ and the weighted sum of domain classifier losses $\mathcal{L}_{cls1}$ and $\mathcal{L}_{cls2}$. The weight λ_{ddal} acts as a tuning hyperparameter to balance these components:

$$\mathcal{L}_{DDAL} = \mathcal{L}_{id} + \lambda_{ddal}(\mathcal{L}_{cls1} + \mathcal{L}_{cls2}) \tag{2}$$

5.2 Batchwise Contrastive Siamese Training

To effectively utilize utterances from the same speakers, we adopt a Batchwise Contrastive Siamese Training (BCST) strategy, inspired by [22], to refine speaker embeddings across different domains into a unified, domain-independent representation. As depicted in Fig. 4, the model receives paired utterances from the same speaker but in different vocal manners.

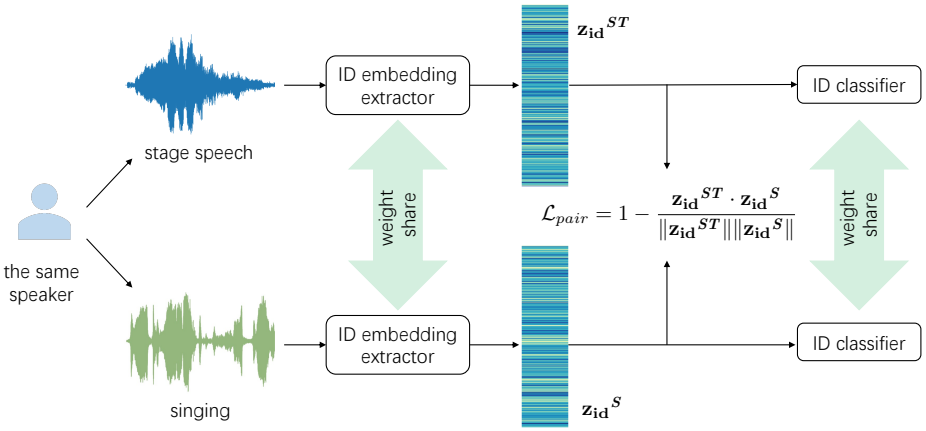

Fig. 4. Overview of the BCST structure

The optimization process focuses on the $\mathcal{L}_{BCST}$, a combined loss comprising the individual utterance losses, $\mathcal{L}_{uttS}$ and $\mathcal{L}_{uttST}$, as well as the pair loss $\mathcal{L}_{pair}$ scaled by a factor λ_{bcst}. The pair loss quantifies the cosine distance between speaker embeddings, $\mathbf{z_{id}}^{ST}$ and $\mathbf{z_{id}}^{S}$, extracted from paired utterances. By leveraging both singular utterance traits and relational information from utterance pairs, the model is encouraged to enhance its ability to distinguish between speakers and maintain feature consistency for the same speaker, even when their vocal manner varies.

$$\mathcal{L}_{BCST} = \mathcal{L}_{uttS} + \mathcal{L}_{uttST} + \lambda_{bcst}\mathcal{L}_{pair} \tag{3}$$

$$\mathcal{L}_{pair} = 1 - \frac{\mathbf{z_{id}}^{ST} \cdot \mathbf{z_{id}}^{S}}{\|\mathbf{z_{id}}^{ST}\|\|\mathbf{z_{id}}^{S}\|} \tag{4}$$

6 Experiments

6.1 Experimental Setup

Dataset We pretrain the model on VoxBlink2 [21] with over 16,000 hours of audio data from 110k speakers. Thereupon, we fine-tune the model using KunquDB's training set. Evaluation is performed on the KunquDB test set.

Network In our baseline (detailed in Table 3), we use ResNet34 [14] as the feature extractor, followed by a Global Statistic Pooling (GSP) layer to condense the length-variable frame-level feature map into a fixed-length representation. This representation is then input to a fully connected layer with 256 dimensions. For speaker identification, we employ the ArcFace classifier [8] (m=0.2, s=32). Binary domain classifier involves stacking Linear-ReLU-Linear structures on $\mathbf{z_{domain}}$ and $\mathbf{z_{id}}$ for domain classification and adversarial learning, respectively. In the attention mechanisms that decouple domain-related features $\mathbf{f_{domain}}$ from global features $\mathbf{f}$, we employ two approaches: a neural network-based method known as Attentive Statistics Pooling (ASP) [24] and a Simple, Parameter-free Attention Module (SimAM) [43].

We initialize the baseline model by pre-training on the VoxBlink2 dataset and experiment with various fine-tuning strategies using the KunquDB training set, as detailed

Table 3. The architecture of our ResNet34 backbone network. The residual building blocks are shown in $[\cdot]$, with the numbers of blocks stacked. Downsampling is performed by Layer2_1, Layer3_1, Layer4_1 with a stride of 2.

Layer	**Structure**	**Output Size**
Conv1	$3 \times 3, 64$	$64 \times 80 \times T$
Layer1	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$64 \times 80 \times T$
Layer2	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 4$	$128 \times 40 \times \frac{T}{2}$
Layer3	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 6$	$256 \times 20 \times \frac{T}{4}$
Layer4	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$	$512 \times 10 \times \frac{T}{8}$
Encoding	Global Statistics Pooling	1024
ID Embedding	Linear	256
Domain Embedding	Linear	256

Table 4. Models varied in architectures, training data, and strategies. **KunquDB fine-tuning** indicates whether to utilize the KunquDB training set for fine-tuning. **DDAL** denotes Domain Discrepancy Adversarial Learning as described in Section 5.1; **BCST** refers to Batchwise Contrastive Siamese Training as detailed in Section 5.2.

ID	**Model**	**Size**	**KunquDB fine-tuning**	**BCST**
M0	ResNet34-GSP	20.54M	×	×
M1			✓	×
M2			✓	✓
M3	+ SimAM-based DDAL	20.79M	✓	×
M4			✓	✓
M5	+ ASP-based DDAL	27.35M	✓	×
M6			✓	✓

in Table 4. **M0** serves as the standard and starting point for all subsequent fine-tuning experiments; it is pre-trained but not fine-tuned. **M1** and **M2** undergo fine-tuning using the standard ResNet34-GSP architecture, aligning with **M0**. In contrast, **M3** and **M4** are built on the SimAM-based DDAL framework; likewise, **M5** and **M6** adopt the ASP-based DDAL approach. **M2**, **M4**, and **M6** incorporate the BCST strategy, further building upon **M1**, **M3**, and **M5**, respectively.

Training Details During pre-training, we apply on-the-fly data augmentation [4] and follow a training setting similar to [28]. For fine-tuning, we utilize a multi-step learning rate (LR) scheduler starting with an initial LR of 10^{-3} to modulate the SGD optimizer, gradually updating the model parameters until convergence. The hyperparameters λ_{ddal} and λ_{bcst} are assigned with a value of 0.5 when used independently within the model (**M1, M2, M3, M5**). However, when both are employed (**M4, M6**), λ_{ddal} is set to 1, while λ_{bcst} is adjusted to 1.5. Input utterances are truncated to 2 seconds and converted to 80-dimensional log Mel-filterbank energies.

Table 5. The performance comparison of different speaker verification systems in terms of Equal Error Rate (EER) across four distinct test sets, as outlined in Section 4.5.

ID	Undifferentiated		*ST*-Domain		*S*-Domain		Cross-Domain	
	EER[%]	mDCF	EER[%]	mDCF	EER[%]	mDCF	EER[%]	mDCF
M0	21.48	0.99	18.81	0.97	23.06	0.97	28.52	1.00
M1	7.95	0.66	7.53	0.61	7.29	0.77	9.84	0.84
M2	7.79	0.67	7.67	0.65	6.47	0.70	9.37	0.79
M3	7.79	0.71	7.57	0.64	7.20	0.87	9.40	0.88
M4	**7.36**	0.71	**7.12**	0.70	**6.21**	0.72	8.37	0.84
M5	7.64	0.71	7.56	0.63	6.41	0.78	8.79	0.88
M6	7.39	0.69	7.41	0.63	6.32	0.71	**8.25**	0.78

Evaluation Metrics Cosine similarity is used for trial scoring. The verification performances are measured by the Equal Error Rate (EER) and the minimum normalized detection cost function (mDCF) with $P_{target} = 0.01$.

Experimental Results Table 5 reports the performance of models on different test sets, with several key observations discussed below.

(1) Model **M0** shows weak robustness on Kunqu data, performing best in the ***ST***-domain due to its exclusive pretraining on speech data. Nevertheless, its performance is still markedly inferior to its excellent performance on regular speech test sets, often below 1% EER.

(2) Models generally perform best when enrollment and test utterances share the same vocal manner, whether in the ***S*** or ***ST*** category. However, their performance notably declines in cross-domain scenarios, indicating the difficulty in extracting domain-agnostic speaker embeddings.

(3) DDAL or BCST individually improves model performance on Kunqu datasets. Deploying both approaches concurrently (**M4** and **M6**) substantially augments this enhancement, delivering superior outcomes.

(4) Regarding the two implementations of attention within the DDAL strategy, the ASP-based implementation (**M5**) outperforms the SimAM-based counterpart (**M3**) across all test sets without BCST. However, with BCST integration, the SimAM-based approach (**M4**) yields better results than the ASP-based method (**M6**) in three out of four test sets, except for the cross-domain scenario.

We randomly select eleven individuals from the test data and visualize their speaker embeddings using the t-distributed stochastic neighbor embedding (t-SNE) algorithm in Fig. 5. Each subfigure corresponds to a specific model (**M0**∼**M6**), providing a visual representation of the distribution patterns learned under various domain adaptation approaches. Notably, the **M0** subfigure reveals a lack of convergence in the distributions of utterances from the same speaker across different domains. In contrast, coherent distributions are observed among similar utterance types, with ***S*** utterances predominantly in the left upper quadrant and ***ST*** utterances in the right lower quadrant. These t-SNE visualizations consistently mirror the objective performance metrics presented in Table 5, confirming the effectiveness of the domain adaptation methods.

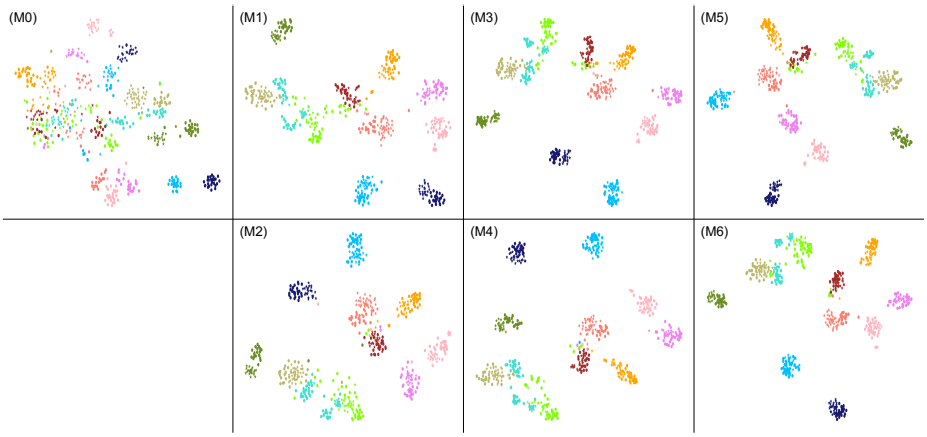

Fig. 5. t-SNE visualization of speaker embedding extracted by seven models (**M0**∼**M6**). Unique colors signify individual distinctions, with circular markers (●) representing stage speech utterances and pentagonal stars (★) denoting singing utterances.

7 Conclusion

This paper introduces KunquDB, a relatively large-scale, publicly accessible audio-visual dataset designed to address research gaps in Chinese opera studies. With detailed annotations, KunquDB aims to serve as a valuable resource for opera and speech-related research endeavors. Leveraging domain discrepancy adversarial learning and batchwise contrastive Siamese training, we establish benchmarks for ASV on Chinese opera data, offering unique insights distinct from conventional speech datasets.

8 Acknowledgements

This research is funded by the Kunshan Municipal Government Research Funding under the project "Deep Learning based Singing Voice Synthesis for Kun Opera". We want to thank the publisher for allowing us to conduct research on their data and DKU library staff members for their coordination. Special thanks to Xiaoyi Qin for his assistance.

References

1. Black, D.A., Li, M., Tian, M.: Automatic identification of emotional cues in chinese opera singing. ICMPC, Seoul, South Korea (2014)
2. Brown, A., Huh, J., Nagrani, A., Chung, J.S., Zisserman, A.: Playing a part: Speaker verification at the movies. In: Proc. ICASSP. pp. 6174–6178 (2021)
3. Cai, D., Cai, W., Li, M.: Within-sample variability-invariant loss for robust speaker recognition under noisy environments. In: Proc. ICASSP. pp. 6469–6473 (2020)
4. Cai, W., Chen, J., Zhang, J., Li, M.: On-the-fly data loader and utterance-level aggregation for speaker and language recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 1038–1051 (2020)
5. Caro Repetto, R., Serra, X.: Creating a corpus of jingju (beijing opera) music and possibilities for melodic analysis. In: Proc. ISMIR (2014)

6. Chen, Q., Zhao, W., Wang, Q., Zhao, Y.: The sustainable development of intangible cultural heritage with ai: Cantonese opera singing genre classification based on cogcnet model in china. Sustainability **14**(5), 2923 (2022)
7. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing **19**(4), 788–798 (2010)
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proc. CVPR. pp. 4690–4699 (2019)
9. Desplanques, B., Thienpondt, J., Demuynck, K.: Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In: Proc. Interspeech. pp. 3830–3834 (2020)
10. Dong, L., Sundberg, J., Kong, J.: Loudness and pitch of kunqu opera. Journal of Voice **28**(1), 14–19 (2014)
11. Fan, Y., Kang, J., Li, L., Li, K., Chen, H., Cheng, S., Zhang, P., Zhou, Z., Cai, Y., Wang, D.: Cn-celeb: a challenging chinese speaker recognition dataset. In: Proc. ICASSP. pp. 7604–7608 (2020)
12. Gong, R., Caro, R., Zhu, T.: Jingju a cappella recordings collection (2019), https://doi.org/10.5281/zenodo.3251761
13. Han, B., Chen, Z., Qian, Y.: Local information modeling with self-attention for speaker verification. In: Proc. ICASSP. pp. 6727–6731 (2022)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. CVPR. pp. 770–778 (2016)
15. Hennequin, R., Khlif, A., Voituret, F., Moussallam, M.: Spleeter: a fast and efficient music source separation tool with pre-trained models. Journal of Open Source Software **5**(50), 2154 (2020)
16. Islam, R., Xu, M., Fan, Y.: Chinese traditional opera database for music genre recognition. In: Proc. O-COCOSDA/CASLRE. pp. 38–41 (2015)
17. Jinpei, H.: Xipi and erhuang of beijing and guangdong operas. Asian Music **20**(2), 152–195 (1989)
18. Kim, J., Kong, J., Son, J.: Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: Proc. ICML. pp. 5530–5540 (2021)
19. Li, Q., Hu, B.: Joint time and frequency transformer for chinese opera classification. In: Proc. Interspeech (2023)
20. Lin, L.: Modernising Cantonese opera through contemporary sound production design. Ph.D. thesis, Middlesex University (2022)
21. Lin, Y., Cheng, M., Zhang, F., Gao, Y., Zhang, S., Li, M.: Voxblink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark. arXiv preprint arXiv:2407.11510 (2024)
22. Lin, Y., Qin, X., Jiang, N., Zhao, G., Li, M.: Haha-pod: An attempt for laughter-based non-verbal speaker verification. In: Proc. ASRU. pp. 1–7 (2023)
23. Nagrani, A., Chung, J., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. In: Proc. Interspeech (2017)
24. Okabe, K., Koshinaka, T., Shinoda, K.: Attentive statistics pooling for deep speaker embedding. arXiv preprint arXiv:1803.10963 (2018)
25. Peng, Z., Wu, J., Li, Y.: Singing voice conversion between popular music and chinese opera based on vits. In: Proc. DASC/PiCom/CBDCom/CyberSciTech. pp. 0999–1003 (2023)
26. Prince, S.J., Elder, J.H.: Probabilistic linear discriminant analysis for inferences about identity. In: Proc. ICCV. pp. 1–8 (2007)
27. Qin, X., Cai, D., Li, M.: Robust multi-channel far-field speaker verification under different in-domain data availability scenarios. IEEE/ACM Transactions on Audio, Speech, and Language Processing **31**, 71–85 (2022)

28. Qin, X., Li, N., Lin, Y., Ding, Y., Weng, C., Su, D., Li, M.: The dku-tencent system for the voxceleb speaker recognition challenge 2022. arXiv preprint arXiv:2210.05092 (2022)
29. Qin, X., Li, N., Weng, C., Su, D., Li, M.: Cross-age speaker verification: Learning age-invariant speaker embeddings. In: Proc. Interspeech (2022)
30. Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.Y.: Fastspeech 2: Fast and high-quality end-to-end text to speech. In: Proc. ICLR (2020)
31. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digital signal processing **10**(1-3), 19–41 (2000)
32. Serra, X.: Creating research corpora for the computational study of music: the case of the compmusic project. In: Audio engineering society conference: 53rd international conference: Semantic audio (2014)
33. Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S.: Deep neural network embeddings for text-independent speaker verification. In: Proc. Interspeech. vol. 2017, pp. 999–1003 (2017)
34. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: Robust dnn embeddings for speaker recognition. In: Proc. ICASSP. pp. 5329–5333 (2018)
35. Srinivasamurthy, A., Caro Repetto, R., Sundar, H., Serra, X.: Transcription and recognition of syllable based percussion patterns: The case of beijing opera. In: Proc. ISMIR. pp. 431–436 (2014)
36. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Processing Letters **25**(7), 926–930 (2018)
37. Wang, H., Liang, C., Wang, S., Chen, Z., Zhang, B., Xiang, X., Deng, Y., Qian, Y.: Wespeaker: A research and production oriented speaker embedding learning toolkit. In: Proc. ICASSP. pp. 1–5 (2023)
38. Wang, S., Yang, Y., Qian, Y., Yu, K.: Revisiting the statistics pooling layer in deep speaker embedding learning. In: Proc. ISCSLP. pp. 1–5 (2021)
39. Wang, W.: Kunqu yishu dadian (昆曲艺术大典). Anhui Literature and Art Publishing House, Anhui (2016), http://www.awpub.com/front/book/10-858
40. Wichmann, E.: Listening to theatre: the aural dimension of Beijing opera. University of Hawaii Press (1991)
41. Wu, Y., Li, S., Yu, C., Lu, H., Weng, C., Zhang, L., Yu, D.: Synthesising expressiveness in peking opera via duration informed attention network. arXiv preprint arXiv:1912.12010 (2019)
42. Wu, Y., Li, S., Yu, C., Lu, H., Weng, C., Zhang, L., Yu, D.: Peking opera synthesis via duration informed attention network. In: Proc. Interspeech (2020)
43. Yang, L., Zhang, R.Y., Li, L., Xie, X.: Simam: A simple, parameter-free attention module for convolutional neural networks. In: Proc. ICML. pp. 11863–11874 (2021)
44. Yang, L., Tian, M., Chew, E., et al.: Vibrato characteristics and frequency histogram envelopes in beijing opera singing (2015)
45. Yao, M., Liu, J.: The analysis of chinese and japanese traditional opera tunes with artificial intelligence technology based on deep learning. IEEE Access (2024)
46. Yu, C., Lu, H., Hu, N., Yu, M., Weng, C., Xu, K., Liu, P., Tuo, D., Kang, S., Lei, G., et al.: Durian: Duration informed attention network for multimodal synthesis. arXiv preprint arXiv:1909.01700 (2019)
47. Yung, B.: Creative process in cantonese opera iii: the role of padding syllables. Ethnomusicology **27**(3), 439–456 (1983)
48. Zhang, H., Jiang, Y., Zhao, W., Jiang, T., Hu, P., Entertainment, T.M.: Chinese opera genre investigation by convolutional neural network. Proc. ISMIR (2021)
49. Zhou, X., Sun, W., Shi, X.: A high-quality melody-aware peking opera synthesizer using data augmentation. In: Proc. ICME. pp. 1092–1097 (2023)