Disentangled Adversarial Domain Adaptation for Phonation Mode Detection in Singing and Speech

Yixin Wang, *Student Member, IEEE,* Wei Wei, Xiangming Gu,
Xiaohong Guan, *Life Fellow, IEEE,* Ye Wang, *Member, IEEE,*

***Abstract*—Phonation mode detection predicts phonation modes and their temporal boundaries in singing and speech, holding promise for characterizing voice quality and vocal health. However, it is very challenging due to the domain disparities between training data and unannotated real-world recordings. To tackle this problem, we develop a disentangled adversarial domain adaptation network, which adapts the phonation mode detection model with the structure of the convolutional recurrent neural network pre-trained on the source domain to the target domain without phonation mode labels. Based on our curated sung and spoken dataset for phonation mode detection, we demonstrate that the subject and the singing-speech mismatches cause performance decline. By disentangling domain-invariant phonation mode and domain-specific embeddings, our method greatly enhances the effectiveness and explainability of unsupervised adversarial domain adaptation. Experiments show that the performance drop caused by the subject mismatch is alleviated via adaptation, resulting in 44.7% and 6.8% improvement of the F-score for singing and speech, respectively. The singing and speech domain adaptation experiment indicates that a model trained on singing data can be adapted to speech, yielding an F-score of 0.56, commensurate with the F-score of 0.59 achieved using a model exclusively trained on speech data. By further investigating the disentangled embeddings, we find that the phonation mode feature shared by singing and speech is invariant to pitch. These results inspire reliable and versatile applications in voice quality evaluation and paralinguistic information retrieval.**

***Index Terms*—Phonation mode detection, voice quality, unsupervised domain adaptation, domain adversarial training.**

I. Introduction

PHONATION modes or types, present in both singing and speech, can serve as valuable indicators of sung/spoken voice quality and overall vocal health. The research problem addressed in this paper is *phonation mode detection* (PMD) for multi-phonation audio in various application scenarios. The three common *phonation modes* (PM) in singing and speech are analyzed: *breathy*, *neutral*, and *pressed* [1], [2]. A PMD system is expected to predict the onset time, offset time, and label for each detected phonation mode. However, lack of annotations and domain disparities in actual audio can pose significant challenges for PMD. Phonation modes provide insights into various types of vocalizations with a practical PMD system, which can be used for evaluating the performance and style of a singer in the music industry [3]–[5], diagnosing potential voice disorders in clinical settings [6], and retrieving paralinguistic information [7], [8].

To create a generalized PMD system, we address the research gaps from three perspectives. First, existing PM datasets [3], [4], [9]–[11] only contain segmented vowels, thus most studies consider PMD a classification task, where the input must be a steady vowel and the output is a PM label, and is not applicable for evaluating real singing or speech. This task is denoted as phonation mode classification (PMC) in this paper. In addition, although extensive studies have been conducted to find a representation for phonation modes, such as glottal source excitation, acoustic parameters, and deep learning embeddings [4], [12], [13], these features extracted from the steady part of vowels can only be used for PMC that neglects temporal dynamics. Therefore, we create a dataset for the PMD task, and develop a model from the temporal-spectral features to estimate phonation mode label sequences and pinpoint the onset and offset timestamps for each phonation.

Second, in previous studies on PMC [3], [4], [9]–[13], it is problematic to train and test a model using data collected from a single individual, as this can lead to overfitting of the dataset and poor performance on real-world data. The discrepancy between the ideal testing data and the real test set may be due to the lack of annotation and domain mismatch. The manual annotation of phonation modes is a costly and labor-intensive process requiring expert knowledge, making it impractical to annotate each audio input to fine-tune the model with labels. Additionally, in real-world application scenarios, it is challenging to employ a PMD model across various unseen domains, which may differ regarding user traits. On one hand, the PMD model sees a severe performance drop when tested on unseen data, as depicted in Fig. 1. On the other hand, fine-tuning a pre-trained PMD model to a new domain requires annotations of phonation modes labeled by experts. This problem can be solved by unsupervised domain adaptation (UDA) [14], which aims to provide an accurate prediction on the target domain using only labeled data from the source domain and unlabeled data from the target domain. Most UDA methods are proposed for computer vision tasks, and

Manuscript received xx xx, xxxx; revised xx xx, xxxx. *(Corresponding author: Ye Wang)*

Yixin Wang is a Ph.D. candidate at Faculty of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China, and a visiting Ph.D. student at the School of Computing, National University of Singapore, 117543, Singapore, funded by China Council Scholarship. (E-mail: yxwang@sei.xjtu.edu.cn).

Wei Wei, Xiangming Gu and Ye Wang are with the School of Computing, National University of Singapore, Singapore. (E-mail: weiwei, xiangming@u.nus.edu, wangye@comp.nus.edu.sg)

Xiaohong Guan is with the MOE KLINNS Laboratory, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China, and also with the Center for Intelligent and Networked Systems, Tsinghua University, Beijing 100084, China (e-mail: xhguan@xjtu.edu.cn).

This work is supported in part by Ministry of Education of Singapore (R-252-000-A56-114) and the National Natural Science Foundation of China (T2341003).

some recent studies present domain mismatches for speaker verification [15], [16], and sound event detection [17].

Lastly, although phonation modes in singing and speech share similar glottal vibration patterns [2], most research has primarily focused on either singing technique identification or paralinguistic information recognition in speech. Few studies consider phonation modes in both singing and speech together [12], as the existing datasets for sung [3], [4], [9] and spoken [10], [11] PMC have been collected separately. This motivates us to investigate the relationship between singing and speech by proposing a paired sung and spoken phonation mode dataset and performing singing and speech domain adaptation.

To address the above issue, we develop a disentangled adversarial domain adaptation network (DADAN) for PMD based on unsupervised domain adaptation (UDA) applicable in various practical scenarios. The unsupervised domain adaptation model consists of two stages. First, we build a PMD system with a phonation mode encoder (PMEncoder) and a phonation mode decoder (PMDecoder) based on the structure of a convolutional recurrent neural network (CRNN) to detect phonation modes and their boundaries as shown in Fig. 1. Based on the temporal-spectral features, the PMD system can evaluate not only static audio but also predict the phonation modes presented in the transitional states of audio. The CRNN predicts frame level output, which is then smoothed to provide phonation level output, including phonation mode labels and their onset/offset timestamps. In the second stage, a domain classifier (DClassifier) is utilized to identify the source and target domain. The pre-trained PMD system is adapted to the unseen domain by minimizing the distance between the source and target domain embeddings based adversarial training strategy of UDA [18]. Unlike typical adversarial DA models, DADAN is composed of a PMEncoder and a domain encoder (DEncoder) to disentangle the input features and explore the explainability of phonation mode representation. The DADAN is applicable to various domain mismatches since it is a general domain adaptation model. This paper focuses on two typical domain adaptation scenarios: subject domain adaptation (SDA) and content domain adaptation (CDA). We use the method of controlling variables to study these two domain adaptation issues by keeping the recording environment and equipment constant. SDA is attributed to the unique voice quality of different subjects. In this paper, CDA indicates the singing and speech domain adaptation based on the inherent difference between singing and speech voice rather than linguistics. Our system greatly improves the PMD performance and outperforms other popular unsupervised domain adaptation models in SDA experiments. In the CDA experiment, DADAN manages to detect phonation modes in speech using only annotated singing data. By further examining the disentangled domain and phonation mode embedding, we demonstrate that the common phonation mode embedding in singing and speech is pitch-invariant.

Our main contributions are:

- We introduce PMD in singing and speech to estimate the phonation modes and their boundaries and create the first sung and spoken phonation mode dataset for PMD.
- We demonstrate domain mismatches for PMD in real application scenarios and develop an unsupervised PMD system, DADAN, applicable for new domains without requiring annotated data.
- Experiments report that DADAN outperforms other DA models on singing and speech data with domain mismatch. Additionally, the PMD model pre-trained on singing data exhibits adaptability to speech. Further experiments reveal the PMEncoder's capacity to learn a pitch-invariant representation shared in singing and speech.

The rest of the paper is organized as follows: in Section II, we review related work on phonation mode analysis in singing and speech and unsupervised domain adaptation methods. In Section III, we formulate the PMD problem and introduce the domain adaptation scenarios of PMD. A description of the Sung and Spoken Dataset for Phonation Mode Detection (SSD4PMD) is illustrated in Section IV. The spectro-temporal feature extraction and the proposed DADAN are presented in Section V. Section VI provides the experimental setup and Section VII discusses the results of the glottal source feature comparison and the domain adaptation experiments. Finally, conclusions are summarized in Section VIII.

II. Related work

A. Phonation mode definition in singing and speech

Most previous studies analyzed the phonation modes in speech and singing separately. Early research inaugurated the phonetic description of spoken voice quality, defined as the quasi-permanent quality of a speaker's voice [19]. A standardized system of phonetic description was first proposed by Ladefoged, who introduced a continuum phonation type model for speech based on the glottal stricture varying from the most closed to the most open state in speech [1]. The near-closed vocal folds vibration results in a creaky or tense phonation type, and the open state of vocal folds leads to a breathy or whispery phonation type, where the modal (neutral) type is the optimal pronunciation state between the tight and loose adduction [20]–[22]. Using a set of representative phonetic labels, voice quality was described in terms of phonation types such as *breathy*, *normal (neutral)*, and *pressed* [11], [23], [24]. In singing, voice qualities are described as phonation modes attributed to vocal folds vibration, defined by the ratio between transglottal airflow and subglottal pressure to identify four essential laryngeal characteristics in a 2-D space: *breathy*, *neutral*, *pressed*, and *flow* [2], [25].

Phonation types in speech and phonation modes in singing reflect the perceived laryngeal voice quality [26]. Voice quality is usually used to describe the perceived timbre resulting from articulation activity. However, the voice quality label set varies across different tasks. This work considers the two terms to be equivalent under the constraint that only the three phonetic labels shared by singing and speech are analyzed, which are *breathy*, *neutral*, and *pressed*. These phonation modes are similarly defined in both singing and speech. In the *breathy* mode, loose vocal folds result in a more open state [1] and a larger transglottal flow versus glottal pressure ratio than the

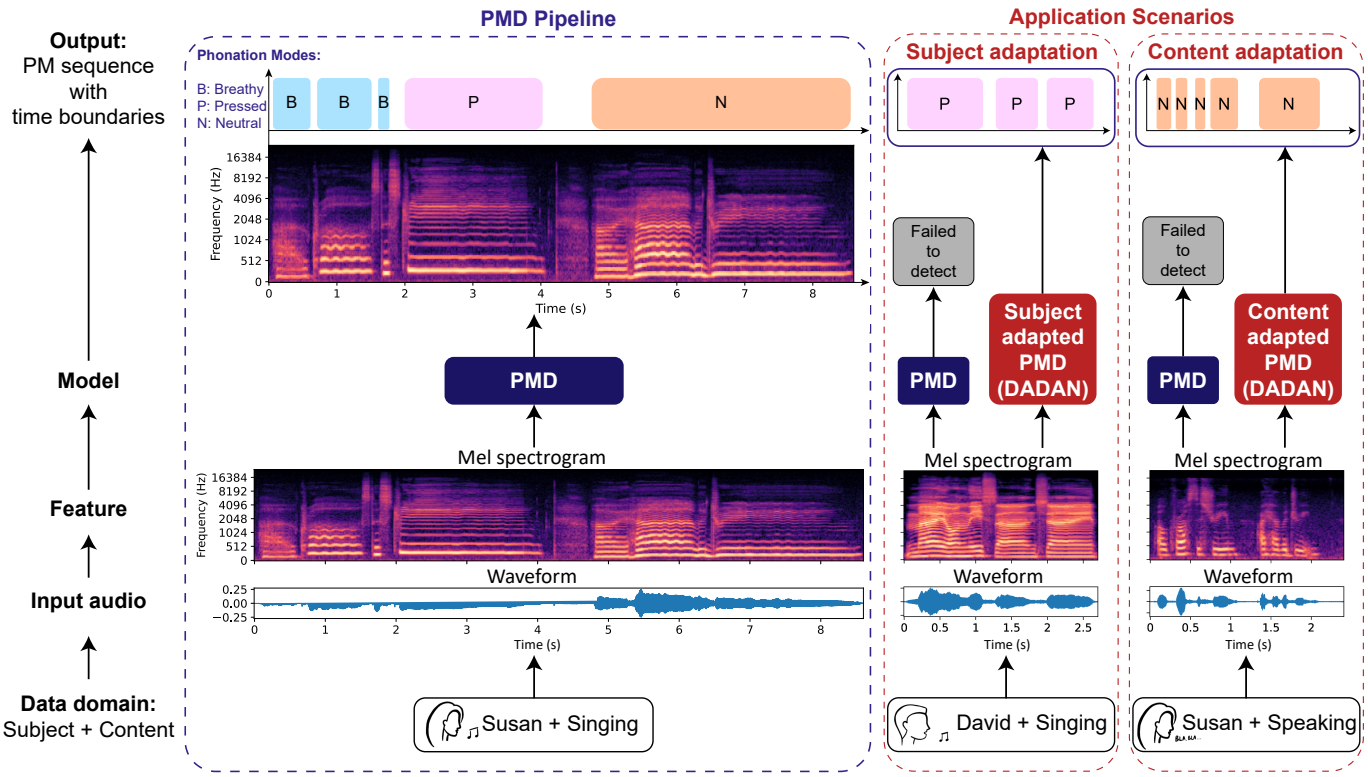

Fig. 1. Phonation mode detection pipeline and two types of domain mismatches (subject and content domain) in real application scenarios.

neutral mode [25]. In contrast, for the *pressed* mode, tight vocal folds result in a less open state and a lower ratio between transglottal flow and glottal pressure. These phonation modes reflect the voice quality attributed to vocal folds vibration. The *flow* phonation mode defined in singing [25] is not considered because it is seldom used in speech and involves both laryngeal (vocal folds vibrations) and supralaryngeal (articulators and resonators) systems [3], [12], [25]. Additionally, *flow* and *neutral* cannot be distinguished using only the ratio between glottal flow and subglottal pressure as demonstrated in [25]. Thus, this work focuses on the three common phonation modes attributed to vocal tension, providing valuable information on voice production [21].

B. The development of research on phonation modes

The development of phonation mode research has two main phases. In the earlier phase, researchers used either clinical or acoustic measurements to parameterize different phonation modes. [27], [28] used clinical equipment to measure the glottogram, open-closed quotient, and laryngeal resistance. However, these methods require precise instruments and qualified experts to perform invasive examinations. Then signal processing was then mainly used for parameterization, and statistical tests were often used to analyze the relation between phonation modes and voice production.

In more recent research, automatic classification of phonation modes from singing voice and speech was studied based on signal processing and machine learning. Most studies on PMC [3], [4], [9]–[13], employed a classical pipeline system including two steps: feature extraction and classification.

The features used for PMC can be categorized into three types: glottal source, acoustic, and general audio features. Phonation modes are believed to be strongly related to glottal source airflow based on the voice production mechanism. Most works aimed to extract features from an estimated glottal source signal using glottal inverse filtering (GIF) since the features could not be measured directly [10]. Some temporal features have been proposed to describe glottal adduction, such as the open quotient (OQ), the closing quotient (ClQ), and the amplitude quotient (AQ) [3], [10], [11], [29], [30], normalized amplitude quotient (NAQ) [27], [31], and maxima dispersion quotient (MDQ) [32]. Commonly used spectral features include the difference of harmonics, the harmonics richness factor (HRF), and the peak slope [3], [10], [11], [33]. However, it is not always reliable to use the source-filter model to decouple the impact of the glottal source and vocal tract, especially for high-pitched voice and time-varying voice production [34], [35]. For sung PMC, [4] compared the glottal source features with the acoustics features and reported that the acoustic feature set gave a higher classification accuracy of 81.62% for soprano and 88.51% for baritone singers. The acoustic features, also known as voice quality features in speech evaluation, mainly contain harmonics, formants, cepstral peak prominence (CPP), and harmonic-to-noise ratio (HNR). Further studies [9], [36] reported that the overall accuracy for PMC in singing and speech could be improved by combining two types of features using the COVAREP features

set [37], showing a PMC accuracy of 90.26% for singing and 79.97% for speech. Some of the latest PMC studies [12], [38]–[40] addressed the problem of GIF-based methods for glottal source extraction. Some excitation features were proposed for singing and speech PMC based on the single frequency filtering (SFF) [38], the zero-time windowing (ZTW) [39], and the zero frequency filtering (ZFF) [40]. [12] compared these three features and found that fusion of all feature sets achieved the best classification accuracy of 85.24% in singing and 78.71% in speech.

A simple classifier, such as the support vector machine [3], [12], [32], [38], [39], is usually used in the classical pipeline system. Recently, [13] reported that deep learning methods could significantly improve PMC performance compared to previous methods which used hand-crafted features and basic classifiers. By adopting a residual attention neural network [41], they achieved the highest classification accuracy of 94.58% for singing. A limitation of the PMC studies is that they only address the single-phonation classification task, where each audio file contains only one phonation mode. A recent work [5] looked into singing technique detection with Mel spectrogram as input. Our previous work [42] studied PMD for singing and discovered domain gaps between singers. This study follows the classical pipeline system, but in contrast to previous studies, this paper systematically examines the SDA and CDA problems and evaluates the glottal source features and UDA methods for PMD, which have not been studied before.

The development of the automatic PMD system has yielded practical applications in potential vocal and neurological disorder evaluation [30], [40]. A PMD system can assist in the diagnosis of vocal disorders by detecting abnormal or unhealthy phonation modes like *pressed* and *breathy*, which may cause vocal fold nodules and polyps [6]. In automatic singing evaluation, a PMD system can serve as a basis for helping identify the vocal characteristics and singing skills [3]–[5]. As the perception of fear, anger, and neutral emotions primarily depends on specific phonation types in speech [8], a PMD system would play a crucial role in retrieving paralinguistic information [43], [44].

C. Unsupervised domain adaptation (UDA)

Since manual annotation for audio data is laborious and expensive, UDA is a promising approach to generalize a model to an unseen domain without labeled data. The unseen domain without labeled data is named as the target domain, and the labeled data belong to the source domain. UDA assumes that there is a gap between the data distribution of the source and target domain. UDA mainly comprises two types of methods. One type of UDA method aligns the source and target domain distribution by minimizing the domain distance, such as maximum mean discrepancy (MMD) [45]. Another popular type of UDA method utilizes a domain adversarial training strategy, which forces the encoder to learn a domain-invariant representation. The adversarial learning method is more widely used because it does not require identical source and target domain data and can be easily applied to different scenarios [18], [46], [47].

Recent research addressed the domain mismatch problem for audio-related tasks, such as sound event detection [17], speaker verification [15], [16], [48], and audio steganalysis [49]. The domain mismatch for audio-related tasks can be ascribed to subject traits (speaker/singer's age, accent, voice quality, etc.), acoustic environment, and recording conditions.

III. Problem formulation

A. Phonation mode detection (PMD)

In this work, we consider input audio consisting of M phonation modes and aim to detect the phonation mode labels and pinpoint their boundaries. Since we analyze the phonation modes in singing and speech in tandem, three shared phonation modes, namely, *breathy*, *neutral*, and *pressed*, as well as the intervals between phonation modes, are to be detected.

The pipeline of PMD is shown in the left part of Fig. 1. The input feature is denoted as $\mathcal{X} \in \mathbb{R}^{T\times K}$, where T is the number of frames in an utterance, and K is the dimension of the feature. The output phonation label set $\mathcal{C} = \{c_1, c_2, c_3, c_4\}$ contains the three phonation modes and a *rest* class representing the silent pauses in an utterance. A PMD system predicts a phonation mode sequence and the corresponding onset and offset of each phonation. The predicted phonation mode sequence is denoted as $\hat{\mathcal{P}} = \{\hat{p}_1, ..., \hat{p}_M\}$, where $p_m \in \mathcal{C}$ is the m-th phonation mode label and M is the number of phonation modes in an utterance. The boundaries sequence is $\hat{\mathcal{S}} = \{(\hat{o}_1, \hat{e}_1), .., (\hat{o}_M, \hat{e}_M)\}$, where $\hat{o}_m$ and $\hat{e}_m$ are the onset and offset timestamps calculated as the frame number of the m-th phonation mode. $\hat{o}_m, \hat{e}_m \in [1, T]$ and $\hat{o}_m < \hat{e}_m$.

B. Domain adaptation scenarios of PMD

In real application scenarios, the mismatch between the training and testing data may result in a severe performance decay for a pre-trained system. We address domain mismatches of PMD in singing and speech, as depicted in Fig. 1, which is not investigated in previous phonation mode analysis studies. It is demonstrated that a PMD model trained on one domain fails to predict an unseen domain and thus necessitates the following domain adaptation tasks:

1) Subject domain adaptation (SDA): to adapt a pre-trained PMD model using label data of one subject (singer or speaker) to an unseen subject without additional annotations. Given the importance of generalization to various subjects, we address the issue of SDA for the development of a practical and usable PMD system that can be applied to a broad user base.

2) Content domain adaptation (CDA): is to generalize a PMD model pre-trained with singing data to speech data without phonation mode annotation, and vice versa. Given that the sung and spoken phonation modes share similar voice production mechanisms [25], this experiment aims to develop a PMD system that can be used for both singing and speech and to explore the underlying similarities and differences between phonation mode representations in singing and speech, which will contribute to a better understanding of voice quality.

TABLE I
INFORMATION ABOUT THE SUNG AND SPOKEN DATASET FOR PHONATION MODE DETECTION (SSD4PMD). "F" AND "M" REPRESENT FEMALE AND MALE SUBJECTS, RESPECTIVELY. THE NUMBERS IN THE LAST TWO COLUMNS ARE PRESENTED AS: MIN ∼ MAX (AVERAGE).

Domain	Subject ID (gender)	Duration (hours:minutes:seconds)	# of songs	# of utterances	# of phonation modes in each utterance	Duration of each phonation mode (s)
Singing	S (F)	0:11:32	7	112	1 ∼ 9 (5)	0.05 ∼ 4.72 (1.02)
	V (F)	0:27:10	12	360	1 ∼ 12 (5)	0.02 ∼ 4.06 (0.71)
	D (M)	0:38:27	16	470	1 ∼ 11 (4)	0.01 ∼ 6.89 (0.86)
	M (M)	0:13:26	7	148	1 ∼ 14 (5)	0.02 ∼ 4.67 (0.71)
	J (M)	0:30:55	11	357	2 ∼ 13 (7)	0.05 ∼ 3.19 (0.42)
	Total	**2:01:30**	**53**	**1347**	**1 ∼ 14 (5.5)**	**0.01 ∼ 6.89 (0.69)**
Speech	S (F)	0:08:29	7	136	2 ∼ 11 (6)	0.02 ∼ 3.03 (0.46)
	V (F)	0:15:27	12	341	1 ∼ 10 (5)	0.02 ∼ 1.9 (0.35)
	D (M)	0:21:48	16	470	1 ∼ 19 (6)	0.02 ∼ 2.79 (0.3)
	M (M)	0:08:53	7	178	1 ∼ 9 (5)	0.02 ∼ 2.38 (0.38)
	J (M)	0:16:13	11	358	2 ∼ 12 (7)	0.03 ∼ 0.51 (0.17)
	Total	**1:10:51**	**53**	**1483**	**1 ∼ 19 (6)**	**0.02 ∼ 3.03 (0.31)**

IV. DATA COLLECTION

A. Background and motivation

Previous phonation mode datasets for singing [3], [4], [9] and speech [10], [11] are unsuitable for PMD, since each utterance in the dataset contains only one vowel with one phonation mode. For the PMD task, each audio should contain more than one phonation mode with a label and boundary timestamps. We are thus motivated to collect the first multi-phonation dataset.

In addition, none of the existing voice quality datasets address the domain adaptation problem. Since the subject traits impact the voice quality evaluation, we collect this phonation mode dataset for the SDA experiments.

Furthermore, despite extant studies [3], [4], [9]–[12], [27], [30], [35], [38]–[40] on phonation modes using similar glottal feature sets for singing and speech, only [12] investigates their commonalities and differences. In consistent with the recent research on sung and spoken phonation modes [12], our dataset includes the three phonation modes shared by singing and speech, which can contribute to more practical applications for amateur singers since the *flow* mode is only defined in the context of singing used by professional Bel Canto singers [25]. With this new corpus, a quantitative comparison of the singing voice and speech phonation modes can be conducted.

B. Dataset curation

To explore the difference between singing and speech regarding voice quality, we adopt the sung and spoken lyrics in [50] for a parallel phonation mode recording collection. Institutional Review Board (IRB) approval was granted for our data collection (NUS IRB Reference No. SOC-21-08). Five subjects (three males and two females) with formal vocal training backgrounds are recruited from a university choir. Participants select the number of songs from a provided list [50] in their preferred key, and they have the autonomy to determine the sequence of phonation modes within each utterance. For the sake of label balance, subjects are asked to use three phonation modes iteratively in each song. Participants first note down the phonation mode sequence on the lyrics sheet placed on a music stand. Then, both singing and speech audio using the same lyrics are recorded consecutively.

The data is collected in a sound-proof recording studio (STC 50+) using an Audio-Technica 4050 condenser microphone with a pop filter. The audio files are all saved in WAV format with 48 kHz sampling rate and 32-bit depth. The annotation files are generated by Adobe Audition from singers' annotation and saved in CSV format. A phonation mode label and its onset and offset timestamps are marked for each vowel.

C. Dataset description

Our sung and spoken dataset for PMD comprises 2.03 hours sung and 1.18 hours spoken utterances from the same collection of songs. Due to the disparate punctuation in singing and speech, the number of utterances is slightly different. The sung vowels usually have a longer duration than spoken vowels. Hence there are more spoken phonation modes in each utterance than sung phonation modes for the same lyrics. The ratio of phonation mode classes are *breathy* : *neutral* : *pressed* $= 53 : 56 : 40$. Detailed information on the SSD4PMD is shown in Table I.

V. METHODOLOGY

In this section, we first introduce spectro-temporal feature extraction. Then we present the disentangled adversarial domain adaptation network (DADAN) for PMD. The model consists of two stages: pre-training of the phonation mode detection system, as shown in Fig. 2 (a), and adversarial domain adaptation, as shown in Fig. 2 (b). As [12] demonstrated good PMC performance using glottal source features such as SFF and ZTW, we adopt these features as spectro-temporal input features for the PMD task in comparison with Mel-spectrogram.

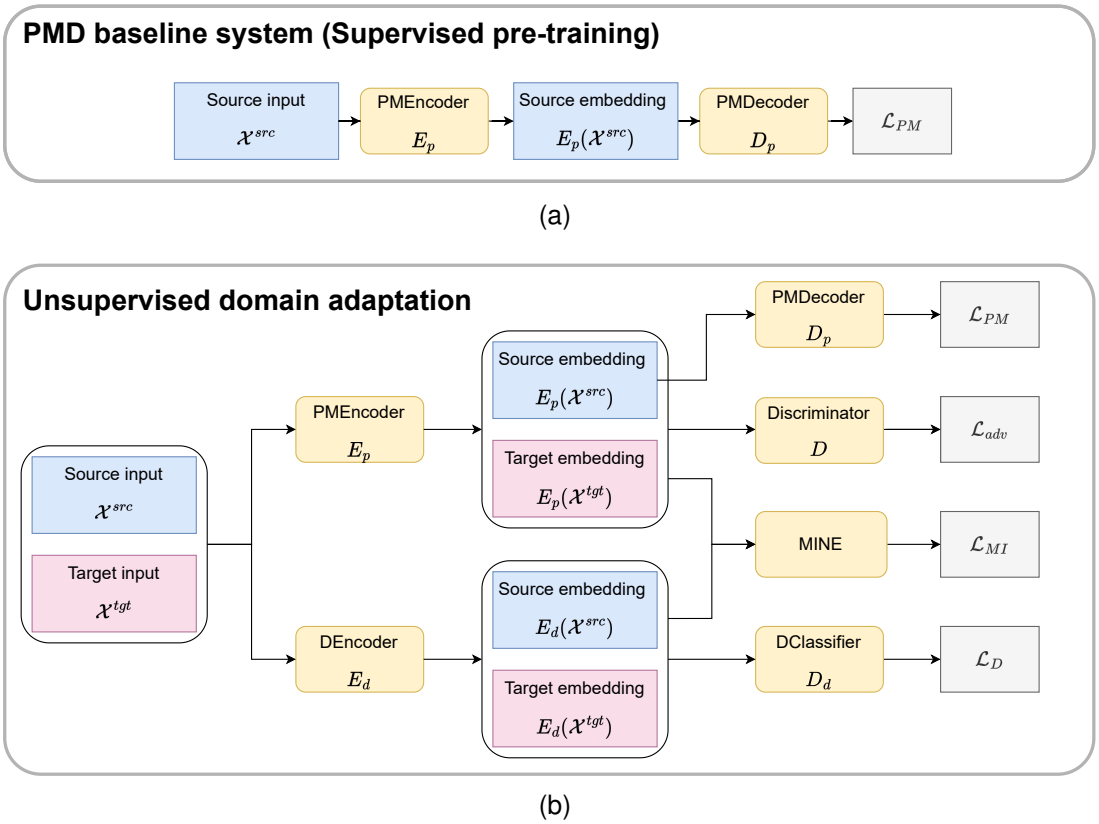

Fig. 2. Overview of the Disentangled Adversarial Domain Adaptation Network (DADAN), which comprises two stages: supervised pre-training and unsupervised domain adaptation. The target input for adaptation is from a new subject or domain with different singing or speech content.

A. Spectro-temporal feature extraction

Previous GIF-based features capture averaged spectral characteristics on the steady part of the voice, which is only suitable for classification tasks rather than detection. The time-varying characteristics play a pivotal role in voice production. Therefore we utilize the glottal source and Mel spectrogram features for PMD.

1) Single frequency filtering (SFF) spectrogram [51]: is proposed to discriminate between speech and nonspeech. SFF represents glottal excitation to depict harmonics in the spectrum with high temporal and spectral resolution without using the inverse filtering method and prior knowledge of fundamental frequency. This method assumes that the glottal impulse signal has a flat spectrum and extracts the amplitude envelope at each frequency step.

The steps involved in calculating an SFF spectrogram are as follows. The input audio is pre-emphasized and framed into T segments $s[n]$. Then $s[n]$ is modulated to a frequency f_k, a sampled frequency in a range of the whole spectrum with step Δf, where $f_k = k\Delta f, k = 1, 2, ..., K$. $\Delta f = 10$ Hz in this study. The modulated signal $u[n, k]$ is given by

$$u[n, k] = s[n, k] \exp(-j2\pi n \bar{f}_k) \tag{1}$$

where $\bar{f}_k$ is the selected frequency (i.e. $\bar{f}_k = f_s/2 - f_k$) and f_s is the sampling frequency. The signal is passed through a single-pole filter with a transfer function of $H(z) = \frac{1}{1+rz^{-1}}$. We choose $r = 0.995$, the same value in [12]. The amplitude envelope of the filtered signal $v[k]$ at frequency f_k is obtained

$$x[n, k] = \sqrt{(v_{Re}[n, k])^2 + (v_{Im}[n, k])^2} \tag{2}$$

where $v_{Re}[n, k]$ and $v_{Im}[n, k]$ are the real and imaginary components of $v[n, k]$. The SFF spectrogram $x[n, k] \in \mathbb{R}^{T\times K}$ is obtained for each frame.

2) Zero time windowing (ZTW) spectrogram [35]: is proposed to extract instantaneous spectral vocal excitation characteristics with high spectro-temporal resolution. The essence of ZTW is to use a decaying window to emphasize the values at the beginning of the window, which is regarded as "zero-time"; hence, the windowed signal renders an impulse-like vocal excitation maintaining a good temporal resolution. Meanwhile, the group-delay function highlights the formants showing vocal resonance.

The input audio is first framed into T segments $s[n]$ of length L point, where $n = 0, 1, 2, ..., L-1$ and multiplies a decaying window $\omega_1^2[n]\omega_2[n]$, which are

$$\omega_1[n] = \begin{cases} 0, & \text{if } n = 0, \\ \frac{1}{4sin^2(\pi n/2L)}, & \text{if } n = 1, ..., L-1 \end{cases} \tag{3}$$

$$\omega_2[n] = 4cos^2(\pi n/2L), n = 1, ..., L-1 \tag{4}$$

The spectrum of each windowed segment (i.e., of $u[n] = \omega_1^2[n]\omega_2[n]s[n]$) is estimated as the numerator of group delay (NGD) function

$$n_{gd}[k] = U_{Re}[k]V_{Re}[k] + U_{Im}[k]V_{Im}[k], k = 0, 1, ..., K-1 \tag{5}$$

where $U[k]$ and $V[k]$ are the K-point discrete-time Fourier transform (DTFT) of $u[n]$ and $v[n]$ ($v[n]$=n$u[n]$), and $U_{Re}[k]$, $V_{Re}[k]$ and $U_{Im}[k]$, $V_{Im}[k]$ correspond to the real and imaginary part of $U[k]$ and $V[k]$, separately. In this work, we use $K = 1024$. The ZTW spectrogram $x[n,k] \in \mathbb{R}^{T\times K/2}$ is calculated from the NGD of each segment.

B. Phonation mode detection system pre-training

The PMD system is pre-trained on each source domain with phonation mode label sequence $\mathcal{P} = \{p_1, ..., p_M\}$ and the corresponding boundaries $\mathcal{S} = \{(o_1, e_1), .., (o_M, e_M)\}$. The backbone network consists of a PMEncoder E_p and a PMDecoder D_p, as shown in Fig. 2 (a). The PMEncoder is a convolutional recurrent neural network (CRNN), which is commonly used in sound event detection [17], [52] as convolutional layers are often used to capture local phonation mode features, and recurrent layers can deal with the long-term temporal connection of audio with varied lengths. CRNN is also the state-of-the-art of singing technique detection [5]. Supposing the PMEncoder can learn an effective phonation mode representation, the PMDecoder only contains fully-connected layers to map the embedding to frame-level phonation mode sequence $\hat{\mathcal{P}}' = \{\hat{p'}_1, ..., \hat{p'}_T\}$, where T is the number of frames in an utterance. Finally, the frame level output is grouped and smoothed to provide phonation-level labels $\hat{\mathcal{P}} = \{\hat{p}_1, ..., \hat{p}_M\}$ as well as their onset and offset timestamps in frame $\hat{\mathcal{S}} = \{(\hat{o}_1, \hat{e}_1), .., (\hat{o}_M, \hat{e}_M)\}$. Specifically, in the output smoothing procedure, after eliminating the segments labeled as *rest*, phonation mode segments that are shorter than five frames are disregarded. Then, segments with the same label that are adjacent to each other are merged if the interval between them is less than two frames. This is done to achieve a more precise and reliable phonation-level output sequence. The PMD system is trained to predict the phonation mode label for each frame so that the time boundaries will also be aligned. Thus, we define the PM loss $\mathcal{L}_{\text{PM}}$ as the cross-entropy loss between $\hat{\mathcal{P}}$ and the frame-level ground-truth.

$$\begin{aligned}\mathcal{L}_{\text{PM}} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{4}\Big(&\mathbb{1}_{p'_t=c_i}\log P(\hat{p}'_t = c_i) \\ &+ (1-\mathbb{1}_{p'_t=c_i})\log P(\hat{p}'_t \neq c_i)\Big),\end{aligned} \tag{6}$$

where p'_t and $c_i \in \mathcal{C}$ are the predicted phonation mode label and ground-truth at frame t, respectively.

C. Adversarial domain adaptation

The PMD model pre-trained on one domain usually does not fit data from an unseen domain, and fine-tuning the model on a new domain also requires the annotation of the target domain. To tackle this problem, we introduce an unsupervised domain adaptation method that can effectively generalize a model to the unseen domain without additional annotation.

The scheme of the proposed DADAN is shown in Fig. 2. The input utterances used to pre-train the PMD model are denoted as *source domain input* $\mathcal{X}^{src} = \{x_1, ..., x_{T^{src}}\}$, while *the target domain input* $\mathcal{X}^{tgt} = \{x_1, ..., x_{T^{tgt}}\}$ is used to predict PM without the ground-truth label, where $x_{T^{src}}$ and $x_{T^{tgt}}$ are the number of frames in the source and target utterances, respectively.

1) Adversarial PM embedding alignment: is implemented to further ensure the PMEncoder would learn a domain-invariant feature from $\mathcal{X}^{src}$ and $\mathcal{X}^{tgt}$. An adversarial discriminator (Discriminator) D takes the PM embedding $\mathcal{Z}^{src}$ and $\mathcal{Z}^{tgt}$ from the source and target domain as input and predicts the possibility of each frame being from the source domain. The adversarial loss is written as:

$$\begin{aligned}\mathcal{L}_{\text{adv}} = &-\frac{1}{T^{src}}\sum_{t=1}^{T^{src}}\log D(E_p(\mathcal{X}_t^{src})) \\ &-\frac{1}{T^{tgt}}\sum_{t=1}^{T^{tgt}}\log(1 - D(E_p(\mathcal{X}_t^{tgt})),\end{aligned} \tag{7}$$

Note that when D is trained to distinguish the domain given the PM embedding, the PM system is learned to minimize phonation mode loss simultaneously to encourage E_p to learn a domain-invariant representation. This min-max process is achieved by gradient reversal layers (GRLs) [18], which do not contain trainable parameters and just inverts the sign of gradient during back-propagation.

2) Mutual Information Minimization: is exploited to enhance the feature disentanglement performance, by integrating the mutual information minimization constraint into the training process of PMEncoder E_p and DEncoder E_d. Mutual information is defined as quantifying the information shared by two random variables (X, Y) and calculated by the Kullback–Leibler divergence (i.e., $I(X,Y) = D_{KL}(P_{(X,Y)}||P_X \otimes P_Y)$), where $P_{(X,Y)}$ is the joint distribution, and $P_X \otimes P_Y$ is the product of marginal distributions. Recent work [53] leverages deep learning to propose a mutual information neural estimator (MINE) for high-dimensional continuous random variables, which is used for learning latent representation in domain adaptation [54], [55]. Therefore, the mutual information can be estimated by maximizing:

$$I(X;Y) = \sup_{\theta\in\Theta}\mathbb{E}_{P_{(X,Y)}}[\text{MI}_\theta] - \log(\mathbb{E}_{P_X\otimes P_Y}[e^{\text{MI}_\theta}]) \tag{8}$$

where MI_θ is a deep neural network parameterized with $\theta \in \Theta$. In this work, it is composed of two linear layers of 64 neurons.

To disentangle the PM embedding and domain embedding, the PMEncoder E_p and DEncoder E_d are trained to minimize the mutual information between their latent representation. Hence, the mutual information loss is written as the min-max optimization function:

$$\begin{aligned}&\mathcal{L}_{\text{MI}} \\ &= \min_{E_p, E_d}\max_{\text{MI}}\sum_{i=1}^{\max(T_t^{src}, T_t^{tgt})} I(E_p(\mathcal{X}_t^{src}, \mathcal{X}_t^{tgt}); E_d(\mathcal{X}_t^{src}, \mathcal{X}_t^{tgt}))\end{aligned} \tag{9}$$

3) Domain disentanglement: is proposed to disentangle the domain-related information from the input feature and encourage the PMEncoder to learn a domain-invariant representation. A CRNN-based DEncoder E_d is used to extract a domain-specific embedding $\mathcal{Z}$ from both the source and target domain, and the embedding passes a DClassifier D_d with two fully-connected layers for domain classification. The domain disentanglement model is optimized by a binary cross entropy (BCE) loss:

$$\begin{aligned}\mathcal{L}_{\mathrm{D}} = &- \frac{1}{T^{src}} \sum_{t=1}^{T^{src}} \log D_d\big(E_d(\mathcal{X}_t^{src})\big) \\ &- \frac{1}{T^{tgt}} \sum_{t=1}^{T^{tgt}} \log D_d\big(1 - E_d(\mathcal{X}_t^{tgt})\big),\end{aligned} \tag{10}$$

VI. Experimental setup

A. Metrics

In our work, each audio sample contains multiple phonation modes, and the number of detected phonation modes may differ from the number of phonation modes in the ground truth. Additionally, the evaluation metrics must consider not only the accuracy of the predicted labels but also the correctness of boundaries.

Therefore, we propose to exploit a set of evaluation metrics similar to the those proposed in [56]. In the context of PMD, we define three intermediate metrics as follows:

- True positive (TP): a phonation mode in the ground truth has both the same phonation label and onset/offset boundaries, with a time difference tolerance (e.g., 20ms) allowed for the boundaries.
- False positive (FP): a phonation mode in the ground truth does not have both the same phonation label and onset/offset boundaries.
- False negative (FN): the model fails to predict a phonation mode with both the correct phonation label and onset/offset boundaries.

Afterward, the precision, recall, and F-score [57] can be calculated as usual. Using different averaging methods, macro-averaged and micro-averaged precision, recall, and F-score can also be computed [58]. Macro-averaged metrics are obtained by taking the average of the metrics (precision, recall, and F-score) for each class, whilst micro-averaged metrics are calculated by aggregating TP, FP, and FN over all test data and then computing the precision, recall, and F-score based on the aggregated numbers. Since the micro-averaged method puts equal importance on each instance, the class with more instances has great impact on the final score. While curating the dataset, we have taken the class balance into consideration by iteratively using each phonation mode. However, it is not possible to strictly ensure a balance in the distribution of classes. As a result, we present macro-averaged metrics as default, unless otherwise stated explicitly. This is done for providing an overall performance of the model across all classes, regardless of any potential imbalances in the class distribution.

Aside from the F-score, the error rate (ER) is computed based on the number of three types of errors: insertion (I), deletion (D), and substitution (S), following the convention of sound event detection [56]. The ER is defined as: $\mathrm{ER} = (\mathrm{I}+\mathrm{D}+\mathrm{S})/N$, where N is the number of phonation modes in the ground truth. Insertion is defined as $\mathrm{I} = \max(0,\ \mathrm{FP}-\mathrm{FN})$, which evaluates the number of phonation modes in the predicted output but does not appear in the ground truth. Deletion, i.e., $\mathrm{D} = \max(0,\ \mathrm{FN}-\mathrm{FP})$, measures the number of phonation modes in the ground truth that does not appear in the predicted output. Substitution $\mathrm{S} = \min(\mathrm{FN},\ \mathrm{FP})$ measures the number of phonation modes in the ground truth that are detected as something else.

To evaluate the boundary detection performance, the *seg F-score* and *seg ER* are calculated adhering to a similar approach, where a TP is counted when the system successfully detects the onset and offset of a segment, regardless of the label assigned to that segment.

B. Ablation study

- **Source Only** is a model trained only on the source data, presenting the model without domain adaptation. The Source Only result is obtained by directly testing the pre-trained model on the target domain.
- **Domain Adversarial Neural Network (DANN) [18]** is a typical unsupervised adversarial domain adaptation model, in which a discriminator is trained to distinguish target from source features. A PMEncoder is trained to generate domain-invariant features to fool this discriminator. The GRLs are first introduced to train the discriminator and encoder simultaneously. It is an ablated model of DADAN without the DEncoder-DClassifier and MINE modules.
- **DADAN without MINE module (DADAN$-$)** is an ablated DADAN model without the MINE module.

C. UDA Baselines

- **Maximum Mean Discrepancy (MMD) [45]** is a non-adversarial unsupervised domain adaptation method based on domain alignment. By minimizing the MMD distance of embeddings from the source and target domain, the distribution of the source and target feature space is aligned so that the model achieves better performance by diminishing the domain gap.
- **Conditional Domain Adversarial Networks (CDAN) [46]** involves using two techniques to enhance the performance of unsupervised adversarial domain adaptation for classification. These techniques include incorporating the classifiers' predictions and feature embeddings to improve discriminability and quantifying the output prediction's uncertainty through entropy.
- **Minimum Class Confusion (MCC) [47]** is an unsupervised adversarial domain adaptation model designed to improve classification accuracy. By leveraging the class confusion matrix of the source data, an MCC loss function is proposed to minimize classification error. The MCC loss function works by converging the binary class confusion values of the predictions, resulting in a matrix in which the values

TABLE II
COMPUTATIONAL COMPLEXITY OF DADAN AND OTHER UDA MODELS.

Model	#Parameters (M)	#MACs (G)
DANN [18]	5.44	21.86
MMD [45]	5.44	21.86
CDAN [46]	5.71	22.64
MCC [47]	5.44	21.86
DADAN	10.89	18.46

on the diagonal represent the confusion of each class with itself. This allows the model to make more stable predictions on unlabeled data. The MCC method is versatile and has been shown to improve performance in a variety of domain adaptation scenarios significantly.

D. Implementation details

We calculate three feature maps for the input audio signal: SFF, ZTW, and Mel spectrogram. For SFF and ZTW features, we downsample the input audio to 16 kHz to speed up the calculation. The Mel spectrogram is calculated from the 128-dimension Mel filterbank, as well as its derivatives and second derivatives, to keep the balance of the input feature dimension. Each frame of Mel, ZTW, and SFF has 384, 512, and 400 spectral dimensions separately. All the features are extracted using a 25-ms window and a 10-ms hop length from the audio. For the SDA experiments, we divide different utterances into training, validation, and test sets. To evaluate the generalization ability of the PMD system and the CDA performance, we split the PM dataset into the three sets and ensure each set contains different subjects.

The pre-trained PMEncoder consists of three convolutional blocks of 3×3 kernel, two recurrent layers of 256 dimensions, and a fully-connected layer of 256 dimensions. The PMDecoder only contains two fully-connected layers with inner dimensions 64 to map each frame to 4-dimensional output. The DEncoder has the same architecture as the PMEncoder, while the DClassifier and the Discriminator have two fully-connected layers of inner dimension 64 and an activation layer for the binary outputs.

For adversarial training, we use the weight of the adversarial loss $\lambda_{adv} = 0.5$, domain loss $\lambda_D = 0.2$, and the mutual information loss $\lambda_{\text{MI}} = 0.2$. The weight of MCC loss is 0.2, and the temperature scaling is 2.5. The Adam optimizer [59] is applied with a learning rate of 1e-4, together with a Newbob scheduler using an initial value of 1e-4 and an annealing factor of 0.8. The model is built on Pytorch library [60] with SpeechBrain toolkit [61] and trained on an RTX2080Ti GPU for 30 epochs with a batch size of 4. Table II reports the number of parameters and multiply-accumulate operations (MACs), which are commonly used metrics for evaluating the spatial and time complexity of deep neural networks. Although DADAN requires more storage than other models, it reduces the time complexity because of fewer input neurons in the dense layer of the classifier. The epoch with the best PMD performance on the validation set is selected to evaluate the model. Our implementation is available online[1].

VII. RESULTS AND DISCUSSION

A. Performance of PMD system with different features

In this experiment, the effectiveness of the proposed PMD system is evaluated with regard to the overall detection performance and the segmentation boundaries based on the metrics mentioned in Section VI-A. We divide the subjects such that three are in the training set, one is in the validation set, and one is in the test set. The PMD pipeline without domain adaptation, depicted in the left block of Fig. 1, is used in this experiment. As shown in Table III, the proposed system manages to detect phonation modes in both singing and speech with an F-score of 0.54 and 0.53, respectively. We compare three input features used for PMC to investigate whether they can keep the temporal dynamic characteristic of phonation modes. The result shows that the Mel spectrogram outperforms other features for all metrics, and it specifically achieves a lower segmentation error rate on SSD4PMD compared with the ZTW and SFF features. Although the two features perform better on PMC, they are not sensitive to the onset and offset boundaries. This may be explained by that the annotators may refer to the Mel spectrogram for phonation mode annotation. Notably, although the segmentation result of the SFF spectrogram with the *segment F-score* of 0.70 and 0.62 in singing and speaking, separately, is not as good as the Mel spectrogram with the *segment F-score* of 0.79 and 0.64 in singing and speaking, the macro F-scores of the former are fairly higher. The results indicate that the SFF spectrogram can capture useful static phonation modes features, hence demonstrating the best performance on PMC [12]. Due to the overall superiority of Mel spectrogram, we extract it as the input feature for the following SDA experiments.

Fig. 3 shows the confusion matrices of the proposed PMD tested on the SSD4PMD with three types of input features. The proposed system demonstrates strong class-wise performance for both singing and speech. By looking into the three columns, it is shown that the three features perform well in classifying *breathy* and *neutral*. Although the *pressed* mode is prone to be classified as *neutral*, SFF outperforms the other two methods for both singing and speech data.

As for segmentation performance, a column named *undetected* in the confusion matrix is added to present the detection errors, where a phonation segment in the ground truth is considered *undetected* if the prediction does not provide a correct timestamp pair of the onset and offset. The PMD system manages to detect voiced segments (e.g., *neutral*, *pressed*) which show strong periodicity; however, *breathy* segments are not easy to distinguish by their periodic variation in amplitudes, which may result in more segmentation errors. Furthermore, phonation boundary detection for singing demonstrates a better performance than speech. It can be explained by spoken lyrics usually having shorter vowels compared to sung lyrics, demonstrated by comparing

[1]https://github.com/aliceyixin/PMD-SingingSpeech

TABLE III
EXPERIMENT RESULTS OF PMD SYSTEM WITH DIFFERENT FEATURES

Feature	Singing					Speech				
	macro F-score ↑	micro F-score ↑	ER ↓	seg F-score ↑	seg ER ↓	macro F-score ↑	micro F-score ↑	ER ↓	seg F-score ↑	seg ER ↓
Mel	**0.54**	**0.63**	**0.56**	**0.79**	**0.40**	**0.53**	**0.60**	**0.58**	**0.64**	**0.53**
ZTW	0.44	0.53	0.71	0.79	0.47	0.39	0.571	0.74	0.54	0.78
SFF	0.48	0.57	0.71	0.70	0.48	0.57	0.59	0.70	0.62	0.70

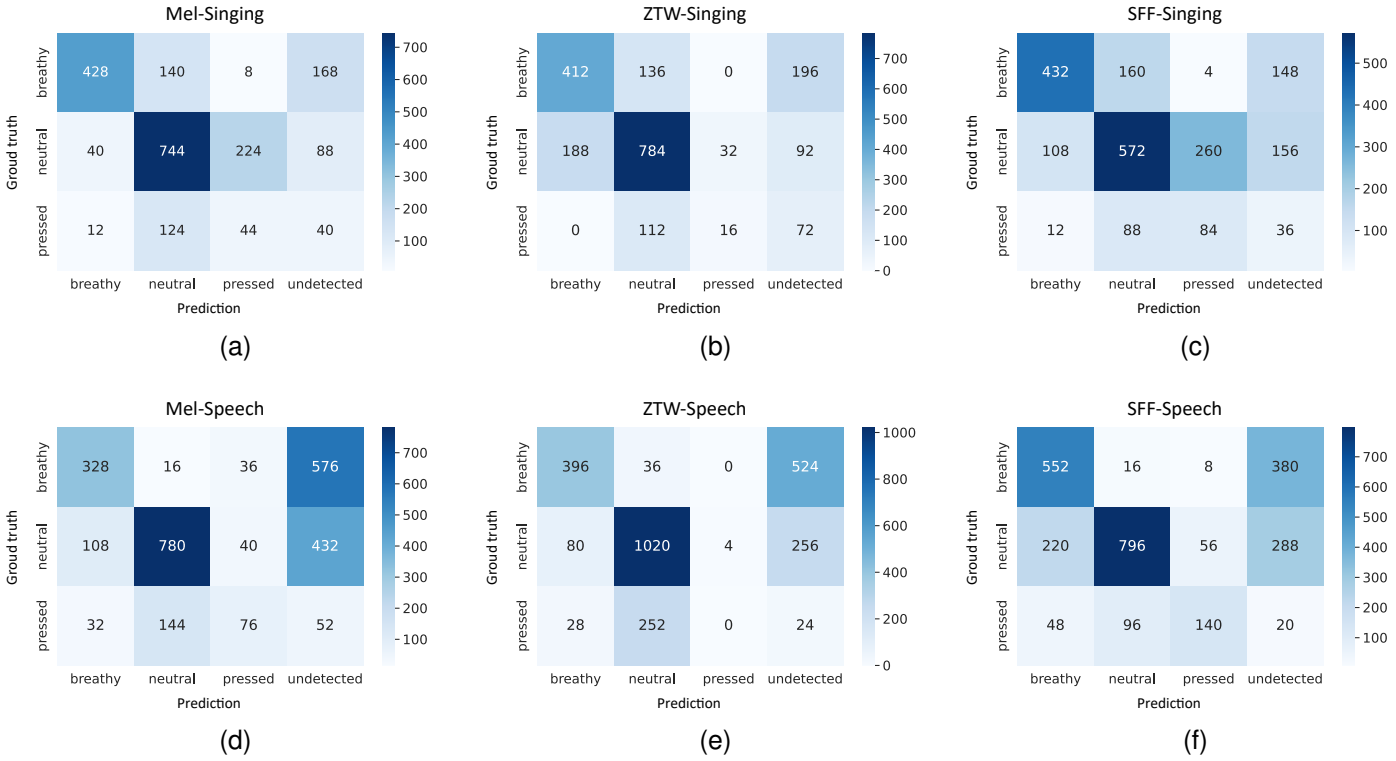

Fig. 3. The confusion matrix of our proposed PMD in singing (first row) and speech (second row). The *undetected* class stands for the phonation modes in the ground truth not being detected by the model with the onset and offset timestamps. (a) PMD in singing with Mel input. (b) PMD in singing with ZTW input. (c) PMD in singing with SFF input. (d) PMD in speech with Mel input. (e) PMD in speech with ZTW input. (f) PMD in speech with SFF input.

the average phonation mode duration between singing and speech in Table I, making it more difficult to pinpoint the transition between different phonation mode patterns.

B. Performance on SDA

The necessity of subject domain adaptation is demonstrated by a 45% and 46% drop in average F-score for singing and speech, respectively, when testing the pre-trained PMD model on a new subject. The difficulty of SDA, evidenced by varying F-scores for different subjects as shown in Table IV and Table V, can be attributed to individual differences in the distribution of *breathy*, *neutral*, and *pressed* phonation modes. For example, subjects D and M have similar phonation modes label distributions, resulting in high domain adaptation F-scores for D-M and M-D, while subjects D and J pronounce differently, resulting in a low F-score. Therefore, we employ DADAN to address this problem.

As shown in Table IV and Table V, the proposed model surpasses all the baselines, achieving an average F-scores of 0.55 for singer adaptation, and 0.47 for speaker adaptation, accordingly. Compared to the Source Only result, the proposed domain adaptation model significantly improves the F-score with 44.7%. MMD is a non-adversarial method and is less effective than the adversarial models in this task since the PMEncoder would learn a PM embedding including domain knowledge from the source and target domain. CDAN, MCC, and DADAN are all improved adversarial methods based on DANN. The performance of DANN, CDAN, and MCC are nearly identical, because CDAN and MCC exploit the inter-class discrimination property, which is suitable for tasks with more categories. The classic DANN is better than CDAN and MCC for only four classes in this task. In the ablation study, compared with DANN and DADAN−, DADAN obtains the highest performance as a result of the utilization of DEncoder-DClassifier and MINE, which effectively removes irrelevant

TABLE IV
PHONATION LEVEL MACRO F-SCORE FOR SDA IN SINGING. MODEL "D-J" DENOTES THE TEST F-SCORE WHEN ADAPTING A MODEL PRE-TRAINED ON "D" (SOURCE SUBJECT ID) TO "J" (TARGET SUBJECT ID).

Model	D-J	D-M	D-S	D-V	J-D	J-M	J-S	J-V	M-D	M-J	M-S	M-V	S-D	S-J	S-M	S-V	V-D	V-J	V-M	V-S	Avg
Source Only	0.39	0.59	0.35	0.42	0.25	0.37	0.36	0.31	0.35	0.32	0.26	0.39	0.24	0.25	0.30	0.40	0.52	0.42	0.57	0.53	0.38
DANN	0.41	0.70	0.58	0.74	0.45	0.40	0.39	0.40	0.34	0.29	0.61	0.79	0.53	0.33	0.61	0.46	0.74	0.40	0.81	0.65	0.53
DADAN−	0.45	0.71	0.54	0.72	0.44	0.45	0.39	0.41	0.33	0.28	0.59	0.77	0.46	0.52	0.36	0.61	0.70	0.44	0.83	0.68	0.53
MMD	0.49	0.50	0.26	0.45	0.40	0.39	0.35	0.43	0.36	0.34	0.34	0.45	0.20	0.36	0.21	0.54	0.51	0.39	0.63	0.56	0.41
CDAN	0.40	0.66	0.58	0.75	0.41	0.42	0.40	0.42	0.35	0.30	0.64	0.73	0.56	0.35	0.64	0.42	0.72	0.43	0.80	0.64	0.53
MCC	0.43	0.73	0.57	0.74	0.33	0.40	0.34	0.35	0.45	0.31	0.59	0.77	0.57	0.24	0.66	0.35	0.75	0.36	0.81	0.60	0.52
DADAN	0.43	0.73	0.72	0.73	0.32	0.44	0.37	0.37	0.31	0.26	0.66	0.77	0.67	0.30	0.68	0.45	0.76	0.38	0.82	0.75	**0.55**

TABLE V
PHONATION LEVEL MACRO F-SCORE FOR SDA IN SPEECH. MODEL "D-J" DENOTES THE TEST F-SCORE WHEN ADAPTING A MODEL PRE-TRAINED ON "D" (SOURCE SUBJECT ID) TO "J" (TARGET SUBJECT ID).

Model	D-J	D-M	D-S	D-V	J-D	J-M	J-S	J-V	M-D	M-J	M-S	M-V	S-D	S-J	S-M	S-V	V-D	V-J	V-M	V-S	Avg
Source Only	0.52	0.50	0.42	0.72	0.62	0.41	0.42	0.48	0.31	0.30	0.28	0.34	0.42	0.47	0.43	0.29	0.50	0.36	0.51	0.42	0.44
DANN	0.56	0.50	0.38	0.61	0.60	0.40	0.29	0.45	0.34	0.49	0.25	0.36	0.48	0.51	0.32	0.32	0.58	0.50	0.49	0.45	0.44
DADAN−	0.57	0.47	0.41	0.71	0.65	0.41	0.36	0.40	0.43	0.39	0.34	0.32	0.49	0.53	0.48	0.32	0.57	0.49	0.52	0.40	0.46
MMD	0.65	0.56	0.38	0.52	0.66	0.36	0.40	0.54	0.50	0.51	0.28	0.33	0.41	0.56	0.40	0.28	0.59	0.36	0.53	0.46	0.46
CDAN	0.54	0.49	0.46	0.76	0.60	0.34	0.28	0.37	0.41	0.40	0.27	0.28	0.36	0.48	0.42	0.42	0.55	0.49	0.47	0.22	0.43
MCC	0.52	0.72	0.36	0.28	0.62	0.43	0.35	0.67	0.48	0.34	0.31	0.33	0.40	0.21	0.23	0.24	0.49	0.34	0.43	0.41	0.41
DADAN	0.57	0.49	0.41	0.67	0.65	0.40	0.37	0.44	0.32	0.33	0.28	0.40	0.48	0.54	0.55	0.36	0.56	0.58	0.53	0.45	**0.47**

information and enhances the PMEncoder.

C. Performance on CDA

To further investigate the correlation between sung and spoken phonation modes and develop a unified PMD model, we conduct CDA experiment on SSD4PMD, and show a significant improvement utilizing domain adaptation compared to the Source Only results in singing and speech. Note that the same dataset split as in Table III is used in this CDA experiment, with results reported in Table VI.

For singing to speech domain adaptation, the proposed method with the SFF feature shows an F-score of 0.56 on speech data, which is close to the F-score of 0.57 supervised trained on the speech data, as shown in Table III. For the Mel spectrogram, the proposed method also improves the PMD results on the target domain with an F-score of 0.49 compared to 0.53 which is supervised trained. Although the ZTW feature obtains an F-score of 0.39 with supervised training, which is not as good as the other two methods, our proposed method still improves PMD performance on the target domain with an F-score of 0.34. Among all the domain adaptation baselines, the proposed DADAN achieves the best result since it addresses the domain gaps using an extra DEncoder-DClassifier branch. Compared with the non-adversarial domain adaptation model MMD, the adversarial models show a more stable performance for different input features. This experiment also indicates that a similar phonation mode pattern is shared by both singing and speech so that a singing PMD system can help to identify the phonation modes in speech.

However, for the speech to singing adaptation, we observe an evident performance decline with all domain adaptation models as shown in the right part of Table VI, with an average F-score degrade of 0.42, 0.60, and 0.45 for Mel, ZTW, and SFF inputs, respectively. We attribute this to two factors: 1) singing data exhibits a broader range of pitch levels than speech, and 2) singing typically involves a diversity of vowel lengths. Therefore, we conclude that a PMD model that has been pre-trained with singing data can be effectively applied to PMD in speech. In ablation study of DADAN for these two adaptations, we conclude that the adversarial structure and the DEncoder-DClassifier significantly contribute to the overall performance by improving the discriminability of a domain adaptation model.

D. Explainability of DADAN on CDA

In the context of singing and speech adaptation, singing encompasses a broader pitch range than speech [50], [62], as demonstrated by the pitch distribution of our phonation mode dataset shown in Fig. 4. It prompts the investigation of whether the phonation mode representation shared by singing and speech is pitch-invariant. To further explore the relationship between singing and speech based on the explainability of DADAN, we conduct a monophonic pitch tracking experiment to compare the embeddings learned by the PMEncoder and DEncoder. The architecture of our pitch tracker consists of a

TABLE VI
COMPARISION OF MACRO F-SCORE FOR DIFFERENT DOMAIN ADAPTATION METHODS BETWEEN THE SINGING AND SPEECH DATA

Model	Feature	Singing → Speech	Speech → Singing
Source Only	MEL	0.12	0.08
	ZTW	0.0	0.0
	SFF	0.08	0.04
DANN	MEL	0.39	0.30
	ZTW	0.31	0.17
	SFF	0.55	0.23
DADAN−	MEL	0.46	0.21
	ZTW	0.23	0.20
	SFF	0.49	0.23
MMD	MEL	0.44	0.27
	ZTW	0.16	0.16
	SFF	0.29	0.30
CDAN	MEL	0.47	0.31
	ZTW	0.33	0.16
	SFF	0.52	0.26
MCC	MEL	0.43	**0.36**
	ZTW	0.24	0.16
	SFF	0.29	**0.31**
DADAN	MEL	**0.49**	0.32
	ZTW	**0.34**	**0.22**
	SFF	**0.56**	0.22

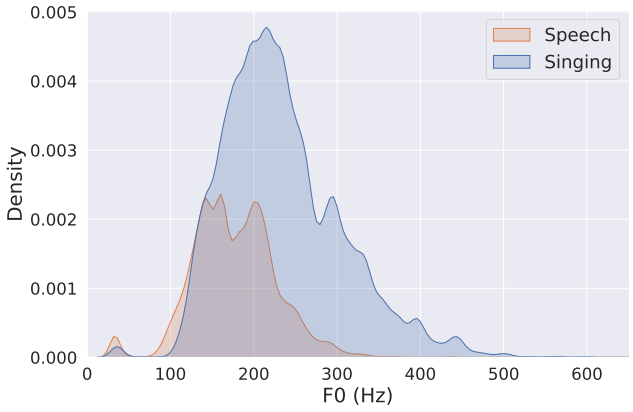

Fig. 4. Pitch distribution of the sung and spoken phonation mode dataset.

pre-trained encoder and a dense output layer of 360 nodes. The output dimension indicates the pitch level from C1 to B7 with 20-cent intervals. The CREPE [63] estimation is regarded as ground truth. The pitch tracker is trained to minimize binary cross-entropy loss, with the parameters in the encoder remaining frozen during training. Both PMEncoder and DEncoder are trained using Adam optimizer for 10 epochs with a learning rate of 1e-5. We utilize the raw pitch accuracy (RPA) with 50 cent thresholds [64], which measure the accuracy of frames

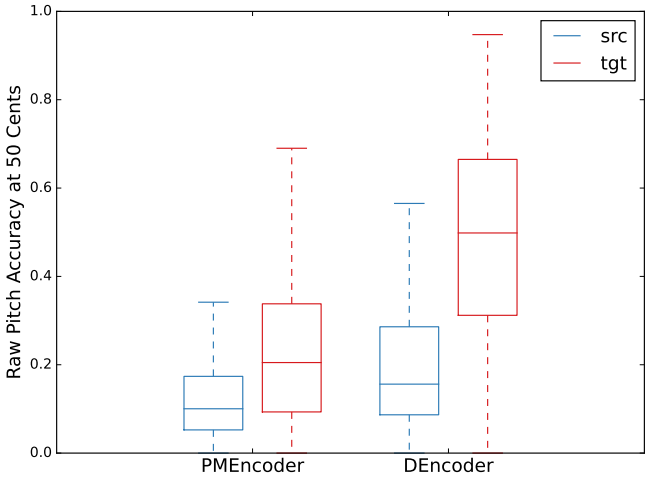

Fig. 5. Average raw pitch accuracy of pitch tracker using pre-trained PMEncoder and DEncoder.

with a quarter-tone error tolerance and are widely used to evaluate the performance of pitch tracker. We refer to the python implementation in [65].

Fig. 5 show the RPA results on source and target data using PMEncoder and DEncoder of DADAN. Here we denote the embedding learned from the PMEncoder as PMEmbedding, while the embedding learned from the DEncoder as DEmbedding. From the figure, the DEmbedding obtained better results than PMEmbedding on both source and target domains. During domain adversarial training, the DEncoder is trained to learn domain-related information, meanwhile, the PMEncoder is able to learn domain-invariant PMEmbedding. This experiment demonstrates that DEmbedding contains more pitch information than PMEmbedding. In addition, it is worth mentioning that the source domain embedding result in a lower RPA performance than the target domain. It can be attributed to that the source domain is trained to learn PM representation using ground-truth labels, whereas the target domain is trained through unsupervised methods. The presence of labeled data in the source domain allows for more accurate PM prediction resulting in a lower RPA than the unlabeled target domain, which can account for the observed difference in performance. This supports the previously reported observation that a generalized phonation mode representation in singing and speech is trained to be pitch-invariant.

VIII. CONCLUSION

Detecting phonation modes for multi-phonation sung songs and speeches is challenging due to lack of the annotated data and the various domain mismatch of recordings. To address the challenges, we create the first sung and spoken phonation mode dataset SSD4PMD and propose a two-step PMD system consisting of a basic PMD model and an unsupervised domain adaptation model, DADAN. The SSD4PMD can be a valuable resource for phonation mode analysis and domain adaptation research. The basic model with the structure of CRNN is pre-trained to detect PM labels and pinpoint their boundaries from temporal-spectral features of singing or speech data by

grouping and smoothing the frame level output, yielding an F-score of 0.54 and 0.53 for PMD in singing and speech, respectively. The proposed DADAN effectively adapts the PMD model to an unannotated target domain and outperforms other DA methods by disentangling the domain-invariant PM feature from the domain-specific feature. Experiments show that the DADAN demonstrates strong SDA performance on both sung and spoken datasets, with an average F-score of 0.55 and 0.47 for singing and speech, respectively. Finally, for CDA, we adopt a PMD model trained with singing data for PMD in speech and show that the common PM representation between singing and speech is trained to be pitch-invariant. These results demonstrate the effectiveness of the DADAN for PMD by disentangling a domain-invariant embedding and offer preliminary insights into the relation between sung and spoken voice quality. A generalized PMD system will inspire real-world applications in the emerging fields of medical, art, and AI assistants, such as voice disorder diagnosis, singing performance evaluation, and emotion recognition in speech and singing.

ACKNOWLEDGMENTS

This research is supported by Ministry of Education of Singapore (R-252-000-A56-114) and the National Natural Science Foundation of China (T2341003). The first author gratefully acknowledges the financial support from the China Scholarship Council during visiting National University of Singapore.

REFERENCES

[1] P. Ladefoged, *Preliminaries to linguistic phonetics*. University of Chicago press, 1971.

[2] J. Sundberg, "Vocal fold vibration patterns and phonatory modes," *STL-QPSR*, vol. 35(2-3), 069-080., 1994.

[3] P. Proutskova, C. Rhodes, T. Crawford, and G. Wiggins, "Breathy, resonant, pressed – automatic detection of phonation mode from audio recordings of singing," *Journal of New Music Research*, vol. 42, no. 2, pp. 171–186, 2013.

[4] J. Rouas and L. Ioannidis, "Automatic classification of phonation modes in singing voice: Towards singing style characterisation and application to ethnomusicological recordings," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*. ISCA, 2016, pp. 150–154.

[5] Y. Yamamoto, J. Nam, and H. Terasawa, "Analysis and detection of singing techniques in repertoires of j-pop solo singers," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[6] C.-T. Wang, M.-S. Lai, and T.-Y. Hsiao, "Comprehensive outcome researches of intralesional steroid injection on benign vocal fold lesions," *Journal of Voice*, vol. 29, no. 5, pp. 578–587, 2015.

[7] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[8] P. Birkholz, L. Martin, K. Willmes, B. J. Kröger, and C. Neuschaefer-Rube, "The contribution of phonation type to the perception of vocal emotions in german: An articulatory synthesis study," *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1503–1512, 2015.

[9] F. Yesiler, "Analysis and Automatic Classification of Phonation Modes in Singing," Master's thesis, Universitat Pompeu Fabra, Oct. 2018.

[10] M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[11] J. Kane and C. Gobl, "Identifying regions of non-modal phonation using features of the wavelet transform," in *Interspeech*, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:2978427

[12] S. R. Kadiri, P. Alku, and B. Yegnanarayana, "Analysis and classification of phonation types in speech and singing voice," *Speech Communication*, vol. 118, pp. 33–47, 2020.

[13] X. Sun, Y. Jiang, and W. Li, "Residual attention based network for automatic classification of phonation modes," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[14] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[15] Z. Wang and J. H. Hansen, "Multi-source domain adaptation for text-independent forensic speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 60–75, 2021.

[16] D. Zhu and N. Chen, "Multi-source domain adaptation and fusion for speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2103–2116, 2022.

[17] W. Wei, H. Zhu, E. Benetos, and Y. Wang, "A-crnn: A domain adaptation model for sound event detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 276–280.

[18] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[19] D. Abercrombie, *Elements of general phonetics*. Edinburgh University Press, 1967.

[20] J. Laver, "The phonetic description of voice quality," *Cambridge Studies in Linguistics London*, vol. 31, pp. 1–186, 1980.

[21] I. R. Titze and D. W. Martin, "Principles of voice production," *Acoustical Society of America Journal*, vol. 104, no. 3, p. 1148, 1998.

[22] J. Kuang and P. Keating, "Vocal fold vibratory patterns in tense versus lax phonation contrasts," *The Journal of the Acoustical Society of America*, vol. 136, no. 5, pp. 2784–2797, 11 2014. [Online]. Available: https://doi.org/10.1121/1.4896462

[23] P. Alku and E. Vilkman, "A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers," *Folia phoniatrica et logopaedica*, vol. 48, no. 5, pp. 240–254, 1996.

[24] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of phonetics*, vol. 29, no. 4, pp. 383–406, 2001.

[25] J. Sundberg, *The Science of the Singing Voice*. Northern Illinois University Press, 1987.

[26] J. H. Esling, S. R. Moisik, A. Benner, and L. Crevier-Buchman, *Laryngeal Voice Quality Classification*, ser. Cambridge Studies in Linguistics. Cambridge University Press, 2019, p. 37–82.

[27] J. Sundberg, M. Thalén, P. Alku, and E. Vilkman, "Estimating perceived phonatory pressedness in singing from flow glottograms," *Journal of Voice*, vol. 18, no. 1, pp. 56–62, 2004.

[28] E. U. Grillo and K. Verdolini, "Evidence for distinguishing pressed, normal, resonant, and breathy voice qualities by laryngeal resistance and vocal efficiency in vocally trained subjects," *Journal of Voice*, vol. 22, no. 5, pp. 546–552, 2008.

[29] T. v. Tarnóczy, "The opening time and opening-quotient of the vocal cords during phonation," *The Journal of the Acoustical Society of America*, vol. 23, no. 1, pp. 42–44, 1951.

[30] M. Millgård, T. Fors, and J. Sundberg, "Flow glottogram characteristics and perceived degree of phonatory pressedness," *Journal of Voice*, vol. 30, no. 3, p. 287–292, 2016.

[31] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *the Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, 2002.

[32] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1170–1179, 2013.

[33] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *the Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[34] P. Alku, "Glottal inverse filtering analysis of human voice production—a review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, 2011.

[35] Y. Bayya and D. N. Gowda, "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function," *Speech Communication*, vol. 55, no. 6, pp. 782–795, 2013.

[36] M. Borsky, D. D. Mehta, J. H. Van Stan, and J. Gudnason, "Modal and nonmodal voice quality classification using acoustic and electroglottographic features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2281–2291, 2017.

[37] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.
[38] S. R. Kadiri and B. Yegnanarayana, "Analysis and detection of phonation modes in singing voice using excitation source features and single frequency filtering cepstral coefficients (SFFCC)," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association*. ISCA, 2018, pp. 441–445.
[39] ——, "Breathy to tense voice discrimination using zero-time windowing cepstral coefficients (ztwccs)." in *Interspeech*, 2018, pp. 232–236.
[40] J. C. Vásquez-Correa, J. Fritsch, J. R. Orozco-Arroyave, E. Nöth, and M. Magimai-Doss, "On modeling glottal source information for phonation assessment in parkinson's disease." in *Interspeech*, 2021, pp. 26–30.
[41] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.
[42] Y. Wang, W. Wei, and Y. Wang, "Phonation mode detection in singing: A singer adapted model," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
[43] A. Afshan, J. Guo, S. J. Park, V. Ravi, J. Flint, and A. Alwan, "Effectiveness of voice quality features in detecting depression," *Interspeech 2018*, 2018.
[44] A. Chanclu, I. B. Amor, C. Gendrot, E. Ferragne, and J.-F. Bonastre, "Automatic Classification of Phonation Types in Spontaneous Speech: Towards a New Workflow for the Characterization of Speakers Voice Quality," in *Interspeech 2021*. Brno, Czech Republic: ISCA, Aug. 2021, pp. 1015–1018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03334492
[45] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.
[46] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," *Advances in neural information processing systems*, vol. 31, 2018.
[47] Y. Jin, X. Wang, M. Long, and J. Wang, "Minimum class confusion for versatile domain adaptation," in *European Conference on Computer Vision*. Springer, 2020, pp. 464–480.
[48] W.-w. Lin, M.-W. Mak, and J.-T. Chien, "Multisource i-vectors domain adaptation using maximum mean discrepancy based autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2412–2422, 2018.
[49] Y. Lin, R. Wang, L. Dong, D. Yan, and J. Wang, "Tackling the cover source mismatch problem in audio steganalysis with unsupervised domain adaptation," *IEEE Signal Processing Letters*, vol. 28, pp. 1475–1479, 2021.
[50] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2013, pp. 1–9.
[51] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 705–717, 2015.
[52] M. Yasuda, Y. Ohishi, and S. Saito, "Echo-aware adaptation of sound event localization and detection in unknown environments," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 226–230.
[53] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 531–540. [Online]. Available: https://proceedings.mlr.press/v80/belghazi18a.html
[54] Y. Tu, M.-W. Mak, and J.-T. Chien, "Variational domain adversarial learning with mutual information maximization for speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2013–2024, 2020.
[55] M. Sang, W. Xia, and J. H. Hansen, "Deaan: Disentangled embedding and adversarial adaptation network for robust speaker representation learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6169–6173.
[56] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.
[57] Y. Sasaki *et al.*, "The truth of the f-measure," *Teach tutor mater*, vol. 1, no. 5, pp. 1–5, 2007.
[58] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management*, vol. 45, no. 4, pp. 427–437, 2009.
[59] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2014.
[60] A. Paszke, S. Gross, F. Massa *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
[61] M. Ravanelli, T. Parcollet, P. Plantinga *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.
[62] C. Zhang, J. Yu, L. Chang, X. Tan, J. Chen, T. Qin, and K. Zhang, "Pdaugment: Data augmentation by pitch and duration adjustments for automatic lyrics transcription," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, 2022, pp. 454–461.
[63] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 161–165.
[64] J. Salamon, E. Gómez, D. P. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, pp. 118–134, 2014.
[65] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "mir_eval: A transparent implementation of common mir metrics," in *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014.

Yixin Wang received the B.S. degree in Automation Science and Engineering from Xi'an Jiaotong University, Xi'an, China in 2017. From September 2015 to August 2017, she undertook a double degree program at CentraleSupélec in France. She visited the School of Computing, National University of Singapore from February 2022 to February 2023. She is currently a Ph.D. candidate at the Faculty of Electronic and Information Engineering, Xi'an Jiaotong University. Her research interests include acoustic signal processing, music information retrieval, and applied machine learning.

Wei Wei received his B.Eng. degree in Computer Science and Technology from Beihang University, China in 2017. He recently received his Ph.D. degree in Computer Science from the National University of Singapore in December 2022. His research interests include sound event detection, automatic speech recognition, and applied machine learning.

Xiangming Gu received his B.Eng. degree from the department of Electronic Engineering, Tsinghua University in 2021. Now he is pursuing his Ph.D. degree in Computer Science at National University of Singapore, under the supervision of Prof. Ye Wang. His research interests include unsupervised domain adaptation, multimodal learning, music information retrieval, etc.

Xiaohong Guan (Life Fellow, IEEE) received the B.S. and M.S. degrees in control engineering from Tsinghua University, Beijing, China, in 1982 and 1985, respectively, and the Ph.D. degree in electrical and systems engineering from the University of Connecticut, Storrs, CT, USA, in 1993.

He visited the Division of Engineering and Applied Science, Harvard University, Cambridge, MA, USA, from 1999 to 2000. From 1985 to 1988 and since 1995, he has been with Xi'an Jiaotong University, Xi'an, China, where he has been a Cheung Kong Professor of systems engineering since 1999 and the Dean of the Faculty of electronic and information engineering since 2008. From 2003 to 2008, he served as the Head of the Department of Automation, Tsinghua University, Beijing, where he has also been with the Center for Intelligent and Networked Systems since 2001. His research interests include modeling and optimization of networked systems, including power and energy systems, manufacturing systems, and cyber-physical systems, and computational intelligence of music.

He is a member of the Chinese Academy of Sciences.

Ye Wang (Member, IEEE) received the B.Sc. degree from the South China University of Technology, China, in 1983, the M.Sc. degree from the Braunschweig University of Technology, Germany, in 1993, and the Ph.D. degree from the Tampere University of Technology, Finland, in 2002. He is an Associate Professor with the Computer Science Department, National University of Singapore (NUS) and NUS Graduate School for Integrative Sciences and Engineering (NGS). He established and directed the sound and music computing (SMC) lab (smc-nus.comp.nus.edu.sg). Before joining NUS, he was a member of the technical staff with the Nokia Research Center in Tampere, Finland, for nine years. His research interests include sound analysis and music information retrieval (MIR), mobile computing, and cloud computing, and their applications in music edutainment and e-Health, as well as determining their effectiveness via subjective and objective evaluations. His most recent projects involve the design and evaluation of systems to support therapeutic gait training using Rhythmic Auditory Stimulation (RAS), second language learning, and motivating exercise via music-based systems.