PHONATION MODE DETECTION IN SINGING: A SINGER ADAPTED MODEL

Yixin Wang[⋆†] *Wei Wei*[†] *Ye Wang*[†]

[†]School of Computing, National University of Singapore, Singapore
[⋆]MOE KLINNS Lab, Faculty of Electronics and Information Engineering,
Xi'an Jiaotong University, Xi'an 710049, China

ABSTRACT

Phonation modes play a vital role in voice quality evaluation and vocal health diagnosis. Existing studies on phonation modes cover feature analysis and classification of vowels, which does not apply to real-life scenarios. In this paper, we define the phonation mode detection (PMD) problem, which entails the prediction of phonation mode labels as well as their onset and offset timestamps. To address the PMD problem, we propose the first dataset PMSing, and an end-to-end PMD network (P-Net) to integrate phonation mode identification and boundary detection, which also prevents the over-segmentation of frame-level output. Furthermore, we introduce an adapted P-Net model (AP-Net) based on an adversarial discriminative training process using labeled data from one singer and unlabeled data from unseen singers. Experiments show that the P-Net outperforms the state-of-the-art methods with an F-score of 0.680, and the AP-Net also achieves an F-score of 0.658 for unseen singers.

***Index Terms*—** Phonation mode detection, singer adaptation, coarse-to-fine, CRNN, voice quality

1. INTRODUCTION

Phonation modes [1] are salient characteristics for vocal quality and health. They are defined by the ratio between the subglottal pressure and the glottal flow [2]. By identifying three widely used phonation modes for amateur singers, namely *breathy*, *neutral*, and *pressed*, the analysis of phonation modes can contribute to the assessment of the singing performance [3, 4] and the vocal health condition [5, 6].

There is a wide range of studies on phonation modes, one of which is the medical examination of the vocal condition. Medical examinations use clinical equipment to measure the ratio between transglottal airflow and subglottal pressure by estimating the glottogram, the open-closed quotient, and the laryngeal resistance [3, 5, 7]. However, these methods require precise instruments and qualified experts to perform invasive examinations. Therefore, most researchers attempted to build an automatic phonation mode classification (PMC) system using signal processing and machine learning methods. Some PMC research focused on designing hand-crafted features to differentiate phonation modes [3, 4, 8, 9, 10], while others performed PMC with fused features [11, 12, 13]. Recently, [14] proposed a residual attention neural network [15] for PMC and achieved the state-of-the-art (SOTA) classification accuracy. Several singing datasets have also been proposed for PMC experiments, which were recorded by professional singers [4, 12, 13].

As introduced above, existing studies only address the PMC task, where each input audio file only contains a single phonation mode. Besides, existing PMC models are all trained and tested on the same singer's data, and if we want to apply them on new singer's singing data, such models need to be fine-tuned or retrained with labeled data from that singer; otherwise, the PMC performance will drop significantly. To address such limitations, we first define the PMD task as depicted in Fig. 1, which is applicable to various real-world applications, such as automatic singing evaluation [4, 16], singing style identification [12, 17], and vocal disorder (nodules and polyps) diagnosis [6]. Each input audio file contains multiple different phonation modes, and for each detected phonation mode, the PMD system will predict an onset time, an offset time, and a label. Since there is no existing singing dataset suitable for the PMD task, a PMD dataset, PMSing, is collected in this work. Then we propose an end-to-end encoder-decoder PMD model, namely P-Net, which provides coarse-to-fine resolution outputs of the decoder to give less fragmented predictions on the phonation mode labels and the corresponding boundaries. Based on P-Net, we further propose a singer adapted PMD model (AP-Net), which does not require labeled data from unseen singers for training.

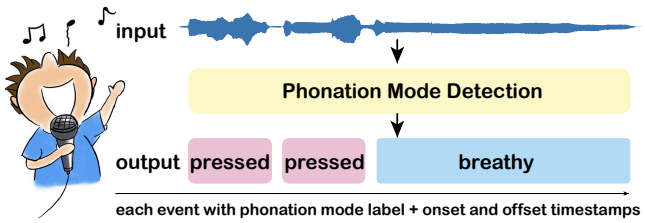


Fig. 1. PMD to predict phonation modes and their boundaries.

The main contributions of this paper include: 1) collecting the first PMD dataset [1], 2) proposing the P-Net to address the PMD problem, and 3) proposing the AP-Net to improve the performance of P-Net on unlabeled data from unseen singers [2].

2. DATASET

Existing phonation mode datasets [4, 12, 13] are only suitable for PMC but not for PMD, because their audio samples are relatively short (around one second) and each sample only contains one phonation mode. Therefore, we propose the first PMD dataset, named as PMSing.

The song list in [18] is adopted to collect PMSing for a wide range of pitches and a variety of phonemes. Two male and two female participants who both have received professional vocal training

The first author gratefully acknowledges financial support from the China Scholarship Council. This project was partially funded by a grant R-252-000-B78-114 from the Ministry of Education in Singapore.

[1] We release the dataset at https://doi.org/10.5281/zenodo.7657058

[2] The code is available at: https://github.com/aliceyixin/PMD-Singing

Singer ID	Total duration (hours:minutes:seconds)	# of songs	# of utterances	# of phonation modes in each utterance	Duration of each phonation mode (s)
DM	0:38:27	16	470	$1 \sim 11$ (4)	$0.01 \sim 6.89$ (0.86)
MM	0:13:26	7	148	$1 \sim 14$ (5)	$0.02 \sim 4.67$ (0.71)
SF	0:11:32	7	112	$1 \sim 9$ (5)	$0.05 \sim 4.72$ (1.02)
VF	0:27:10	12	360	$1 \sim 12$ (5)	$0.02 \sim 4.06$ (0.71)
Total	1:30:35	42	990	$1 \sim 14$ (5)	$0.01 \sim 6.89$ (0.83)

Table 1. Information about the PMSing dataset. The numbers in the last two columns are presented as: min $\sim$ max (average)

are selected from the choir of our university. The participants are asked to sing a song using the three phonation modes iteratively and note down these phonation modes. They are free to choose the number of songs they can sing within the list. The dataset is recorded in a sound-proof studio using an Audio-Technica 4050 condenser microphone with a pop filter and saved in WAV format with a 48kHz sampling rate and 32-bit depth. Each song is annotated using Adobe Audition and the annotations are saved in CSV format.

Detailed information of the PMSing is presented in Table 1. The total duration of PMSing is 1.51 h, containing 42 songs with an average duration of 2.16 min. Compared to existing PMC datasets, the PMSing dataset contains a longer duration, and the duration of the phonation modes varies from 0.01 to 6.89 s. Additionally, all the audio files in PMSing contain multiple phonation modes.

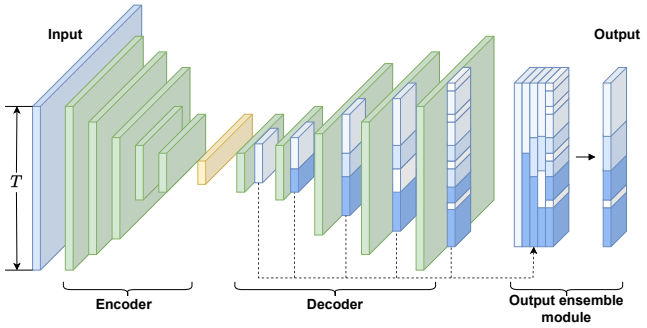


Fig. 2. Architecture of P-Net.

3. METHODS

A PMD model aims to identify the phonation mode labels in the singing voice and pinpoint their corresponding onsets and offsets. Since we only analyze phonation modes for amateur singers, three classes of phonation modes are detected in this paper, namely *breathy*, *neutral*, and *pressed*. The *flow* (*resonant*) phonation mode is usually produced by professional singers in classical singing as reported in [4]. As for the intervals between two phonation modes, we introduce a *rest* class to denote the quasi-silent parts.

The PMD problem is formulated as follows. The audio input feature is denoted as $\mathcal{X} \in \mathbb{R}^{T \times M}$, where T is the total number of frames and M is the dimension of the feature. The four phonation classes (including *rest*) are denoted as $\mathcal{C} = \{c_1, c_2, c_3, c_4\}$. Then there are two PMD outputs: 1) the predicted phonation labels $\hat{\mathcal{P}} = \{\hat{p}_1, \ldots, \hat{p}_N\}$, where $\hat{p}_n \in \mathcal{C}$, $n \in [1, N]$, and N is the number of detected phonation modes, and 2) their corresponding onsets and offsets $\hat{\mathcal{S}} = \{(\hat{o}_1, \hat{e}_1), \ldots, (\hat{o}_N, \hat{e}_N)\}$, where $\hat{o}_n, \hat{e}_n \in [1, T]$ ($\hat{o}_n < \hat{e}_n$) denote the onset and offset of the n-th detected phonation mode.

3.1. P-Net

P-Net consists of three components: an encoder, a decoder, and an output ensemble module (see Fig. 2).

Phonation mode encoder The encoder contains B sequential convolutional neural network (CNN) blocks, followed by a bottleneck embedding layer. It takes the audio feature $\mathcal{X}$ as input and outputs the embedding result $\mathcal{Z}$.

Phonation mode decoder Being symmetric to the encoder architecture, the decoder contains B sequential blocks. Each decoder block comes with a transposed convolutional module and a recurrent neural network (RNN) module. The output of the b-th decoder block is denoted as $\mathcal{Q}_b^D$.

Output ensemble module For the detection task, the model's outputs are sequential labels and their boundaries. However, directly using the last layer's output often suffers from the over-segmentation problem. Inspired by a study on video segmentation [19], a course-to-fine (C2F) output ensemble module is proposed for PMD to get a smooth and accurate prediction result. It first projects the outputs of the decoder to probabilities and then upsamples them to the input's temporal size T. Next, the upsampled probabilities are summarized by applying a weight term α_b to produce the frame-level output $\mathcal{F}$. More specifically, the ensemble output $\mathcal{F}$ is computed by:

$$\mathcal{F} = \sum_{b=1}^{B} \alpha_b * \text{Upsample}\left(\text{softmax}(\mathcal{Q}_b^D), T\right), \tag{1}$$

where α_b is the ensemble weight of each output (i.e., $\alpha_b > 0, \sum_b \alpha_b = 1$), and the upsampling function $\text{Upsample}(\cdot, T)$ maps the input to temporal size T using linear interpolation.

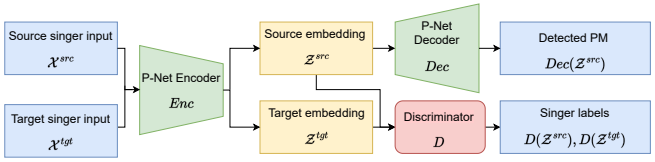
Based on the ensemble output $\mathcal{F}$, the frame-level prediction $\hat{P}' = \{\hat{p'}_1, \ldots, \hat{p'}_T\}$ can be obtained by choosing the class label yielding the maximum probability for each frame:

$$\hat{p}'_t = \underset{c_i \in C}{\text{argmax}}\, \mathcal{F}_{t,c_i}, 1 \leq t \leq T, \tag{2}$$

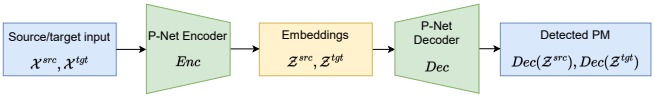
where $\mathcal{F}_{t,c}$ is the ensemble output of the t-th frame for class c_i.

By grouping frame-level outputs $\{\hat{p}'_1, ..., \hat{p}'_T\}$ and removing the *rest* segments, the predicted sequence of phonation labels $\{\hat{p}_1, ..., \hat{p}_N\}$ can be obtained, along with their onsets and offsets $\hat{\mathcal{S}} = \{(\hat{o}_1, \hat{e}_1), \ldots, (\hat{o}_N, \hat{e}_N)\}$.

Optimizing P-Net When training the P-Net model, the frame-level ground truth $P' = \{p'_1, ..., p'_T\}$ is used to compute the cross-entropy loss $\mathcal{L}_{\text{CE}}$:



(a) Adversarial training stage



(b) Test stage

Fig. 3. AP-Net: Adversarial discriminative singer adaptation training and test stages.

$$\mathcal{L}_{\text{CE}} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{4}\Big(\mathbb{1}_{p'_t=c_i}\log\mathbb{P}(\hat{p}'_t=c_i) + (1-\mathbb{1}_{p'_t=c_i})\log\mathbb{P}(\hat{p}'_t\neq c_i)\Big), \tag{3}$$

where $\mathbb{1}_{p'_t=c_i}$ is the indicator function, and it equals to one when $p'_t = c_i$ and equals to zeros otherwise.

To reduce over-segmentation errors, the smoothing loss [20] $\mathcal{L}_{\text{SM}}$ is introduced to smooth the frame-level output by minimizing the difference of log-probability within the range of τ_{max} between adjacent frames, where τ_{max} is a smoothing threshold parameter:

$$\mathcal{L}_{\text{SM}} = -\frac{1}{T}\sum_{t=1}^{T}||\min(\tau_t, \tau_{\text{max}})||^2, \qquad \tau_t = |\log\mathcal{F}_{t,c} - \log\mathcal{F}_{t-1,c}|, \tag{4}$$

The final loss to optimize P-Net is

$$\mathcal{L}_{\text{PMD}} = \mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{SM}}, \tag{5}$$

where λ controls the weight of the smoothing loss.

3.2. AP-Net

The PMD model trained on one singer usually does not fit other singers. When adapting the model to an unseen singer, we have to fine-tune the model with labeled data of the new singer.

To address this problem, inspired by [21], we propose the AP-Net based on P-Net, which can generalize well to unseen singers without using additional labels. As shown in Fig. 3, the singer who comes with labeled data is the *source singer*. The input audio feature and the phonation label sequence of the source singer are denoted as $\mathcal{X}^{src} = \{x_1, ..., x_{T^{src}}\}$ and $\mathcal{P}$. Meanwhile, the unseen singer is the *target singer*, whose audio feature is denoted as $\mathcal{X}^{tgt} = \{x_1, ..., x_{T^{tgt}}\}$, where T^{src} and T^{tgt} are the number of frames.

Singer discriminator A singer discriminator D is introduced in AP-Net to perform singer classification. The source audio feature $\mathcal{X}^{src}$ and the target audio feature $\mathcal{X}^{tgt}$ are both fed into the encoder to get the source and target embedding outputs $\mathcal{Z}^{src}$ and $\mathcal{Z}^{tgt}$. Then the discriminator takes the embedding outputs as input and predicts the possibility that each frame is from the source singer. The outputs of the discriminator are denoted by $D(\mathcal{Z}^{src})$ and $D(\mathcal{Z}^{tgt})$.

Optimizing AP-Net With a pre-trained P-Net, the AP-Net training stage aims at optimizing the singer discriminator and further tuning the P-Net encoder to fit the target singer's data while the decoder is frozen. The adversarial training strategy is adopted that the discriminator and encoder are optimized in an alternating way. For the discriminator, the following loss is minimized to obtain a better singer classification performance:

$$\mathcal{L}_{\text{D}} = -\frac{1}{T^{src}}\sum_{t=1}^{T^{src}}\log D(Enc(\mathcal{X}_t^{src})) - \frac{1}{T^{tgt}}\sum_{t=1}^{T^{tgt}}\log(1-D(Enc(\mathcal{X}_t^{tgt})), \tag{6}$$

For optimizing the encoder, the following loss is used:

$$\mathcal{L}_{\text{Enc}} = -\frac{1}{T^{tgt}}\sum_{t=1}^{T^{tgt}}\log D(Enc(\mathcal{X}_t^{tgt})) - \mathcal{L}_{\text{PMD}}. \tag{7}$$

where the first term is to maximize singer classification errors to allow the encoder to produce similar embedding outputs for both the source and target singers' data, and the second term, PMD loss of the source singer, could stabilize the training process [22, 23].

Testing AP-Net During the test stage, the optimized encoder and the P-Net decoder are used to test the AP-Net's performance on both the source and the target singer's data.

3.3. Implementation details

For the input audio signal, we calculate the 128-dimension Mel filterbank feature as well as its derivatives and second derivatives, using 25-ms window size and a 10-ms hop length from the audio. Both the encoder and the decoder contain five blocks (i.e., $B = 5$), each of which has two convolutional / transposed convolutional layers with a kernel size of 5.

For training losses, we use $\tau_{\text{max}} = 16$ and the weight of smoothing loss $\lambda = 0.15$. The Adam optimizer [24] 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 [25] with SpeechBrain toolkit [26] and trained on an RTX2080Ti GPU for 50 epochs with a batch size of 16. The epoch with the best PMD performance on the validation set is selected to evaluate the model.

4. EXPERIMENTS

4.1. Evaluation Metrics

Previous PMC studies [10, 14] only use accuracy as their evaluation metric for the reason that they do not predict time boundaries. In our PMD work, each audio sample contains more than one phonation mode, and the number of detected phonation modes may not be the same as the number of the ground truth phonation modes. Moreover, evaluation metrics need to measure not only the accuracy of the predicted label but also the correctness of the detected onset and offset boundaries. A time tolerance of 0.1 s is permitted for detecting boundaries in this work.

Therefore, we use evaluation metrics similar to those for sound event detection [27]. Under the context of PMD, three intermediate statics are defined as follows:

Model	F-score	Error rate	Training time per epoch (s)
VD-RANN	0.645	**0.37**	434
Smoothing-CRNN	0.539	0.68	14
P-Net (ours)	**0.680**	0.47	**9**

Table 2. Experiment results for P-Net.

- True positive (TP): there is a phonation mode in the ground truth with both the same phonation label and boundaries in the prediction.
- False positive (FP): there is no phonation mode in the ground truth with both the same phonation label and boundaries in the prediction.
- False negative (FN): the model fails to predict a phonation mode with both the correct phonation label and boundaries.

Afterwards, a class-based F-score is computed by taking the average of F-score on each class (F-score $= \frac{2\times\text{TP}}{2\times\text{TP}+\text{FP}+\text{FN}}$). The error rate (ER) is calculated by averaging the numbers of substitution (S), deletion (D), and insertion (I) errors. Substitution is defined as $\text{S} = \min(\text{FN},\ \text{FP})$, which measures the number of phonation modes in the ground truth that are detected as something else. Deletion, i.e., $\text{D} = \max(0,\ \text{FN} - \text{FP})$, evaluates the number of phonation modes in the ground truth that does not appear in the predicted output. Insertion $\text{I} = \max(0,\ \text{FP} - \text{FN})$ evaluates the number of phonation modes in the predicted output but does not appear in the ground truth.

4.2. Baselines

Because there is no existing work on PMD, we construct two baselines using PMC models:

VD-RANN is a combination of the vowel detection algorithm [28] and the SOTA PMC model based on the residual attention network (RANN) [14]. It first detects vowel segments from a song and then predicts a phonation mode for each segment with RANN.

Smoothing-CRNN is a convolutional recurrent network (CRNN) model with a frame-level smoothing post-processing step.

4.3. Experiment results

P-Net results We first compare the PMD results of our proposed P-Net with the baselines on the PMSing dataset. Table 2 shows that the P-Net outperforms the VD-RANN and the Smoothing-CRNN baselines with an overall F-score of 0.680.

The VD-RANN achieves relatively good results by combining the vowel detection algorithm and the SOTA PMC model. The vowel detection method works well on detecting boundaries and therefore yields a low error rate. However, it takes about quadruple more time than the end-to-end methods because of the intensive signal processing calculation. In addition, the proposed P-Net shows higher F-score than VD-RANN because the RNN layers can capture the temporal dynamics for long audio whereas the RANN only deals with short and steady audio. Thus, the RANN model is less effective for the PMD task in a real singing scenario.

We also compare the P-Net with the Smoothing-CRNN, which is an intuitive method with easy implementation. It has the same numbers of CNN and RNN layers as P-Net. It is found that without the coarse-to-fine output ensemble module, the onset and offset detection performance degrades significantly, resulting in a lower F-score and a higher error rate than those of P-Net. Furthermore, we trained a similar Smoothing-CRNN model without RNN layers, and the overall F-score decreases by 67%. It demonstrates that CNNs could capture phonation mode features of fixed length audio but cannot deal with longer and unstable audio with variable length. RNN is an essential part of PMD model to deal with temporal dynamics, we thus only consider CRNN as one of the baselines.

Model	Source singer		Target Singer	
	F-score	Error rate	F-score	Error rate
VD-RANN	0.645	**0.37**	0.523	0.49
Smoothing-CRNN	0.539	0.68	0.320	0.74
P-Net (ours)	**0.680**	0.47	0.289	0.75
AP-Net (ours)	0.668	0.45	**0.658**	**0.46**

Table 3. Experiment results for AP-Net.

Model	Breathy	Neutral	Pressed
P-Net	0.065	0.425	0.182
AP-Net	**0.625**	**0.574**	**0.601**

Table 4. Class-wise F-score for the target singer.

AP-Net results A crucial problem in the previous setting is to test on the data of the same singer as the training set. When adapting the pre-trained model to a new singer, we can see that the performance of the first three pre-trained models drops severely, as reported in Table 3. In this case, our proposed AP-Net greatly surpasses the baselines, achieving the F scores of 0.668 and 0.658 on the source and the target singers, respectively. The pre-trained VD-RANN, Smoothing-CRNN, and P-Net could fit the source singer well, but could not generalize well to an unseen singer. This indicates that the adversarial training process helps to learn a more singer-generalized embedding. Note that the VD-RANN shows better results than the Smoothing-CRNN because the vowel detection step is singer-independent.

Class-wise results Table 4 reports the F-score for each phonation mode class. The source and target data are from two different male singers. Compared to the non-adapted model (P-Net), the AP-Net can significantly improve the class-wise results on the target singer, especially for the *breathy* and *pressed* classes. As for the *neutral* class, AP-Net outperforms P-Net by a small margin, because the feature patterns of *neutral* phonation mode are fairly common among same-gender singers.

5. CONCLUSION

Previous work on PMC tries to classify the phonation modes with prior knowledge of segmentation, which only considers the ideal circumstance and neglects the application gap. In this paper, we introduce the PMD task to detect phonation modes and their boundaries for real singing data and create a dataset, PMSing, accordingly. To address this problem, we first propose the P-Net whose performance surpasses the baselines built on SOTA PMC approaches with respect to training efficiency and detection performance. Moreover, we propose the AP-Net which is trained in an unsupervised and adversarial way so that it can generalize well on unseen singers' data. The experimental results of our proposed AP-Net show superior performance on unseen singers' data compared with all baseline PMD models.

6. REFERENCES

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

[2] Johan Sundberg, "Vocal fold vibration patterns and phonatory modes," *STL-QPSR*, vol. 35, 69-80, 1994.

[3] Johan Sundberg, Margareta Thalén, Paavo Alku, and Erkki Vilkman, "Estimating perceived phonatory pressedness in singing from flow glottograms," *Journal of Voice*, vol. 18, no. 1, pp. 56–62, 2004.

[4] Polina Proutskova, Christophe Rhodes, Tim Crawford, and Geraint 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.

[5] Elizabeth U. Grillo and Katherine 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.

[6] Chi-Te Wang, Mei-Shu Lai, and Tzu-Yu 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] Moa Millgård, Tobias Fors, and Johan Sundberg, "Flow glottogram characteristics and perceived degree of phonatory pressedness," *Journal of Voice*, vol. 30, no. 3, pp. 287–292, 2016.

[8] Sudarsana Reddy Kadiri and Bayya 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*. 2018, pp. 441–445, ISCA.

[9] S. R. Kadiri and P. Alku, "Mel-frequency cepstral coefficients derived using the zero-time windowing spectrum for classification of phonation types in singing," *J Acoust Soc Am*, vol. 146, no. 5, pp. EL418, 2019.

[10] Sudarsana Reddy Kadiri, Paavo Alku, and B. Yegnanarayana, "Analysis and classification of phonation types in speech and singing voice," *Speech Communication*, vol. 118, pp. 33–47, 2020.

[11] Daniel Stoller and Simon Dixon, "Analysis and classification of phonation modes in singing," in *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR*, 2016, pp. 80–86.

[12] Jean-Luc Rouas and Leonidas 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*. 2016, pp. 150–154, ISCA.

[13] Furkan Yesiler, "Analysis and Automatic Classification of Phonation Modes in Singing," M.S. thesis, Universitat Pompeu Fabra, Oct. 2018.

[14] Xiaoheng Sun, Yiliang Jiang, and Wei 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.

[15] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[16] Ning Zhang, Tao Jiang, Feng Deng, and Yan Li, "Automatic singing evaluation without reference melody using bidense neural network," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 466–470.

[17] Margareta Thalen and Johan Sundberg, "Describing different styles of singing: A comparison of a female singer's voice source in "classical", "pop", "jazz" and "blues"," *Logopedics Phoniatrics Vocology*, vol. 26, no. 2, pp. 82–93, 2001.

[18] Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim, and Ye 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*. IEEE, 2013, pp. 1–9.

[19] Dipika Singhania, Rahul Rahaman, and Angela Yao, "Coarse to fine multi-resolution temporal convolutional network," *arXiv preprint arXiv:2105.10859*, 2021.

[20] Yazan Abu Farha and Jurgen Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3575–3584.

[21] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[22] Shayan Gharib, Konstantinos Drossos, Emre Cakir, Dmitriy Serdyuk, and Tuomas Virtanen, "Unsupervised adversarial domain adaptation for acoustic scene classification," *arXiv preprint arXiv:1808.05777*, 2018.

[23] Wei Wei, Hongning Zhu, Emmanouil Benetos, and Ye 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)*. IEEE, 2020, pp. 276–280.

[24] Diederik P. Kingma and Jimmy Lei Ba, "Adam: A Method for Stochastic Optimization," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 12 2014.

[25] Adam Paszke, Sam Gross, Francisco Massa, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[26] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, et al., "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[27] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, pp. 162, 2016.

[28] Sarmila Garnaik, Avinash Kumar, Gayadhar Pradhan, and Kabiraj Sethi, "An efficient approach for detecting vowel onset and offset points in speech signal," *International Journal of Speech Technology*, vol. 23, no. 3, pp. 643–651, 2020.