Phonation Mode Detection in Singing: A Singer Adapted Model

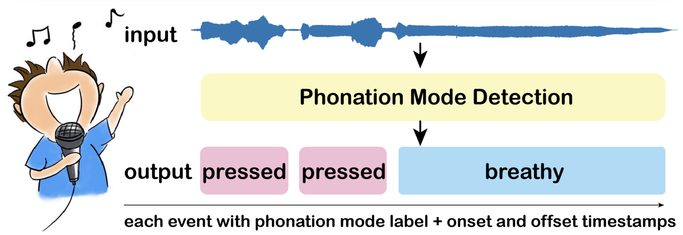
Yixin Wang[*†], *Wei Wei*[†], *Ye Wang*[†]

[*] Xi'an Jiaotong University
[†] National University of Singapore (NUS)
NUS Sound & Music Computing Lab

† work performed during visiting at NUS

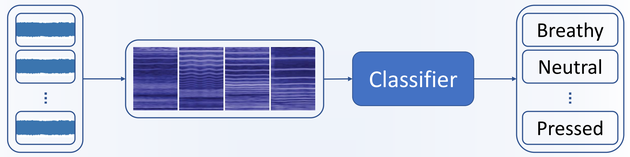
ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

Proposed Task



- **Phonation mode detection (PMD)** is the task of identifying phonation modes and their respective temporal start and end time in a recording.

Prior Work



- **Phonation mode classification (PMC)** is the task of classifying a recording into a phonation mode.

Require:
- Fixed-length audio input with only one vowel.
- Supervised training data.
- Fine tuning with labels when applying to new singers.

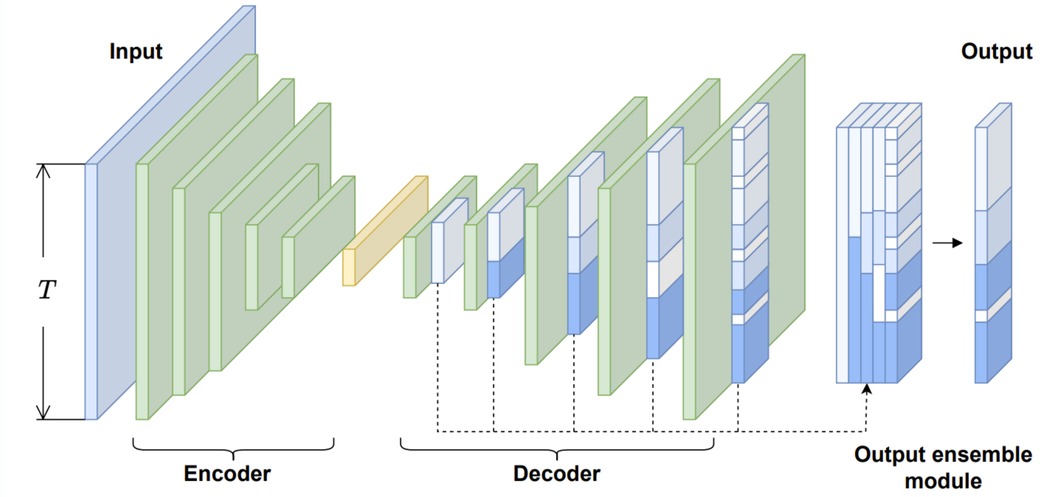
Contributions

1. We introduce a novel **PMD** problem and create a **multi-phonation singing dataset** for this task.
2. We present an Encoder-Decoder model **P-Net** to predict phonation mode labels and the boundary timestamps.
3. We propose the **AP-Net**, an improved version of P-Net, to perform PMD on unseen singers without phonation mode labels.

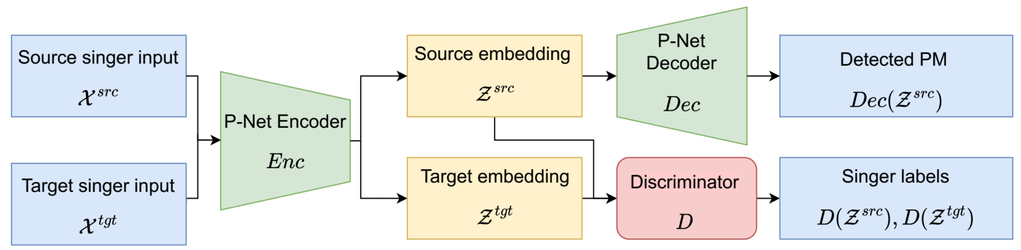
We release our PMD dataset and implementation:

Data Code

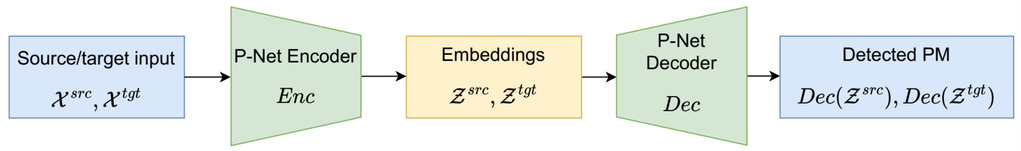
PMD Model: P-Net



Singer Adapted PMD Model: AP-Net



(a) Adversarial training stage



(b) Test stage

Phonation Modes Dataset

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)

Existing phonation mode datasets are only suitable for PMC but not for PMD. The proposed dataset contains a longer duration and multiple phonation modes.

Results

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

- P-Net outperforms the baselines with improved performance and efficiency.

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

- AP-Net surpasses the non-adapted model on target singer without label.

Class name	P-Net		AP-Net	
	Source	Target	Source	Target
breathy	67.91	6.45	54.19	62.46
neutral	65.52	42.50	74.49	57.43
pressed	86.57	18.18	71.64	60.08

Table 4. Class-wise F-score on the source and target singer

- AP-Net improves the class-wise results on both the source and target singer.

References

[1] P. Proutskova et al., Breathy, resonant, pressed – automatic detection of phonation mode from audio recordings of singing, Journal of New Music Research.
[2] D. Stoller et al., Analysis and classification of phonation modes in singing, ISMIR 2016.
[3] J. Rouas et al., Automatic classification of phonation modes in singing voice: Towards singing style characterisation and application to ethnomusicological recordings, Interspeech 2016.
[4] X. Sun et al., Residual attention based network for automatic classification of phonation modes, ICME 2020.