

VocalNET:

A solution for real-time singing voice vocal register classification

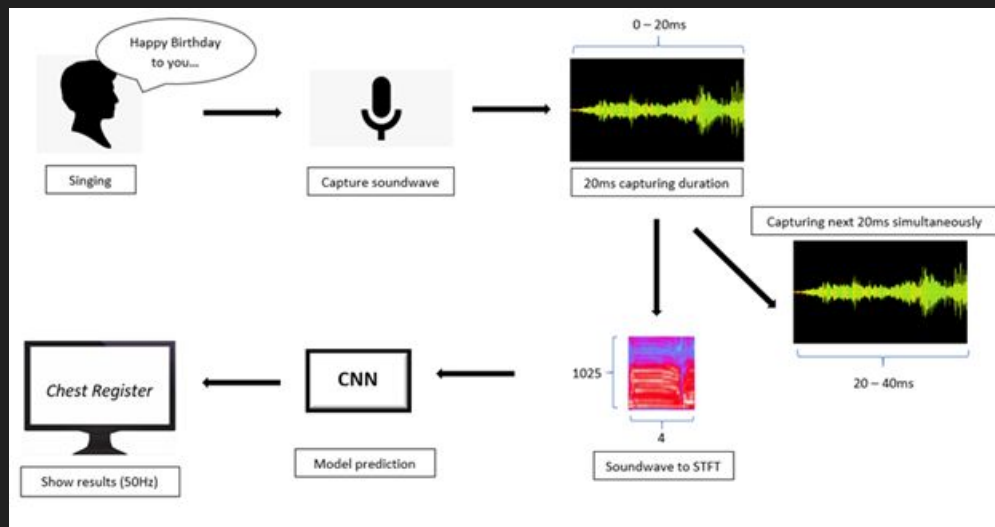
SPCC AI Project Group

Steven Luo

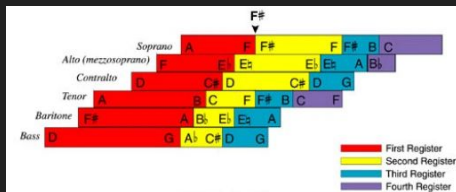
Angel Au

Justin Lam

Samuel Yau



Factors of timbre



Vocal Register

Sound quality

Mouth shape

Pitch

Gender

Age

Amplitude

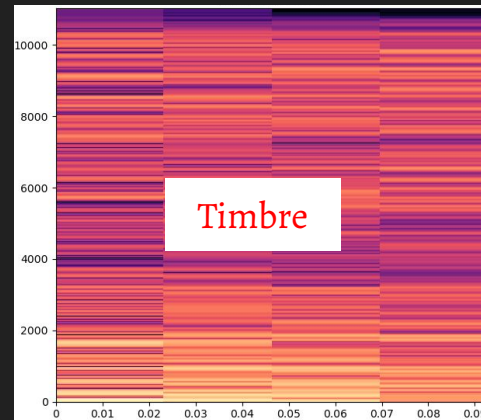
Vocal range

Amateurs


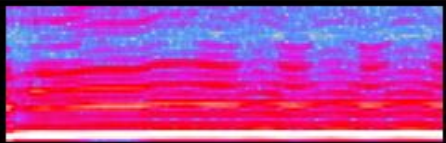
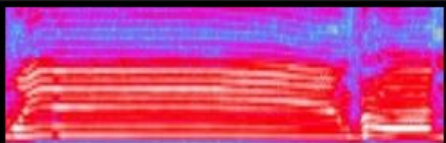
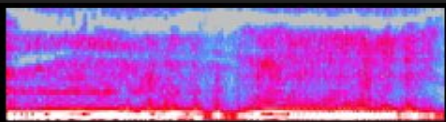


Model
/
Professional

Timbre



Vocal Registers - STFT

Register	Timbre	STFT	Descriptions
Whistle	Similar to the sound of a whistle		<ol style="list-style-type: none"> 1. Clear fundamental frequency 2. Clear and regular intervals
Head	Lighter and thinner		<ol style="list-style-type: none"> 1. Clear and low fundamental 2. Strong “noises” above fundamental
Mixed	Different ratios of chest voice to head voice would result in different timbres		<ol style="list-style-type: none"> 1. No fundamental observed 2. Harmonics are collectively strong
Chest	Heavier and fuller		<ol style="list-style-type: none"> 1. “Noisy” fundamental 2. Irregular intervals, with noise

Research Procedure

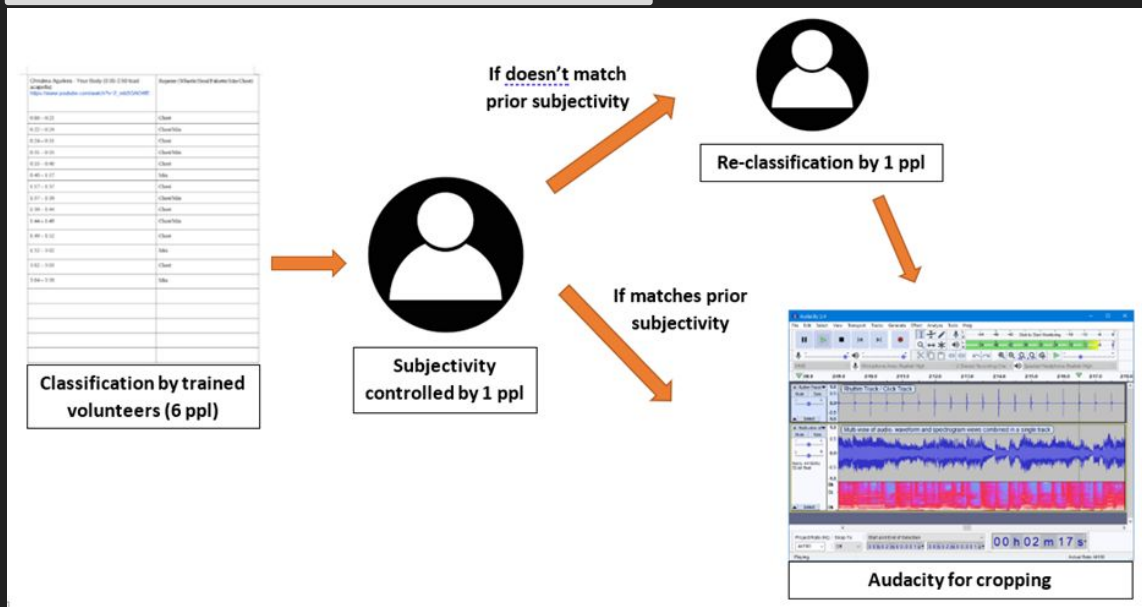
The flow

1. Data collection
2. Data processing
3. Model selection
4. Real-time prediction

Data Collection

Female Singers (min)		
YouTube Pop (for training)	Opera (for test)	Volunteer Choir (for test)
40	15	10

Data Collection pipeline



Data Processing

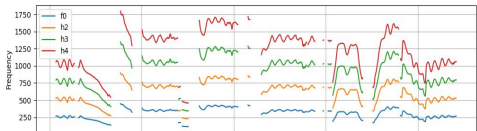
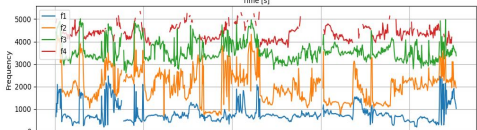
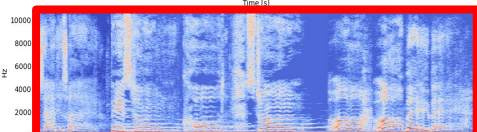
Harmonics

VS

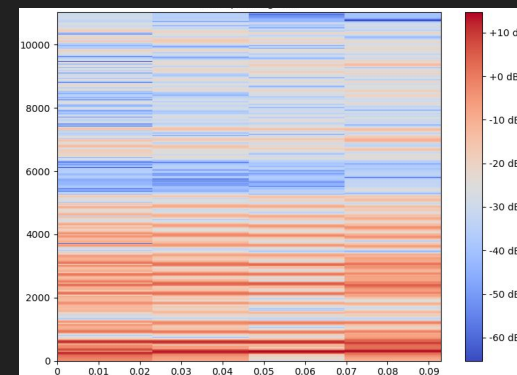
Formants

VS

STFT

MIR technique	Network	Accuracies
	FC network	32%
	FC network	37%
	CNN network	90+%

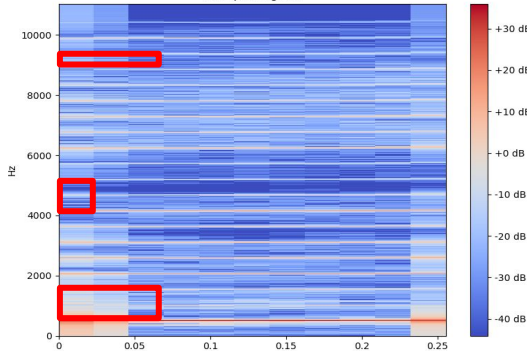
Frequency (1025)



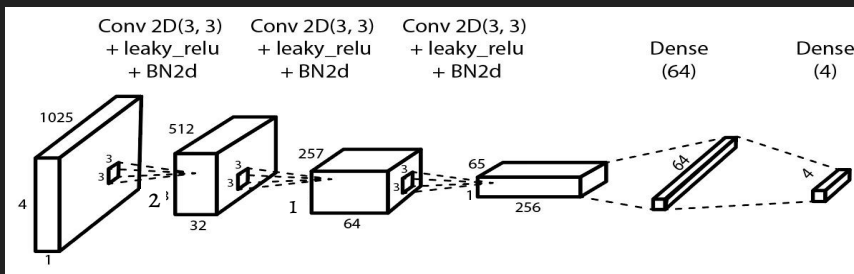
Time (4)

- $N_{fft} = 2048$
- $Hop_length = 128$ (default: 512)

Model selection -- 1D vs 2D

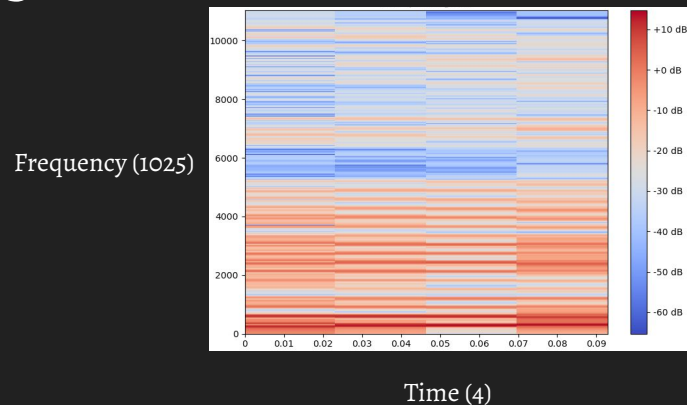
Model	Training Acc	Validation Acc	Kernels
1D-CNN (time-axis)	95.21%	90.08%	
1D-CNN (frequency-axis)	<u>98.48%</u>	90.24%	
2D-CNN	97.59%	<u>94.29%</u>	

Model Architecture

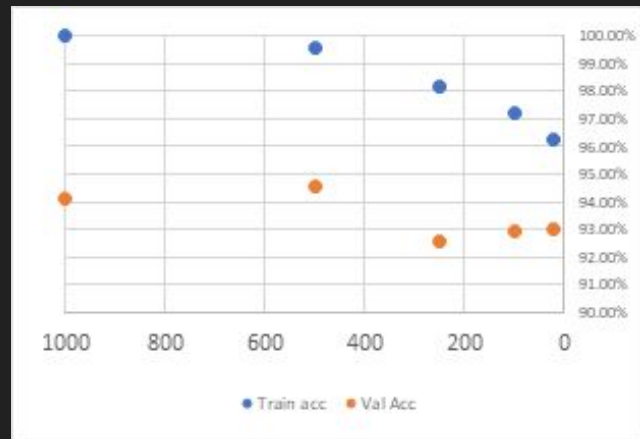


Methodology - Real-time capabilities

- 20ms stft = (1025 x 4)
 - Yet, Interval acc. 90+
- 20ms << flight of the bumblebee note interval (~67ms)



Audio interval (ms)	Training Acc	Validation Acc
1000	100.00%	94.14%
500	99.59%	94.53%
250	98.16%	92.60%
100	97.22%	92.95%
20	96.22%	93.01%



Methodology - Real-time capabilities

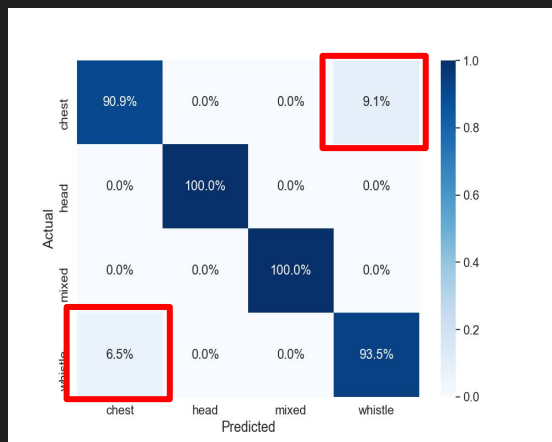
- Low inference time
- Allow continuous prediction in real-time

Audio interval (ms)	Audio processing time (ms)	Inference time (ms)	
		CPU	GPU
100	0.41	1.15	2.21
20	0.47	1.08	2.10

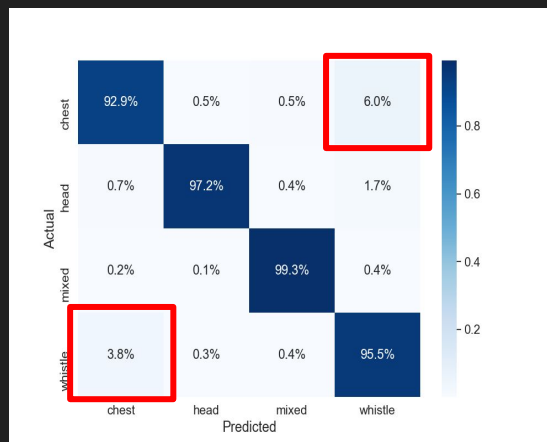
Evaluation

1. Error analysis
2. Acc. on different dataset

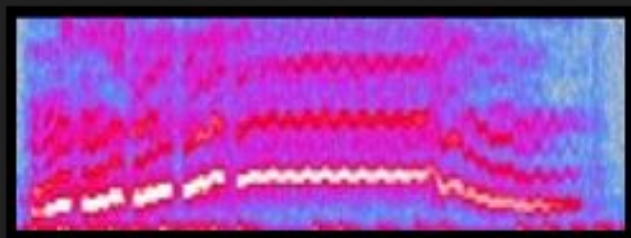
Error analysis



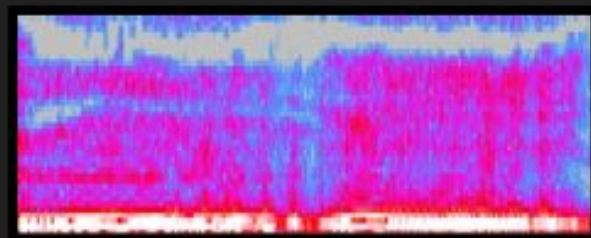
20ms



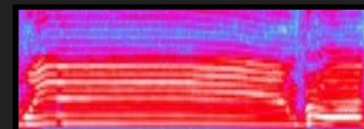
100ms



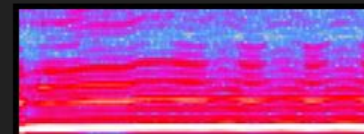
Whistle



Chest



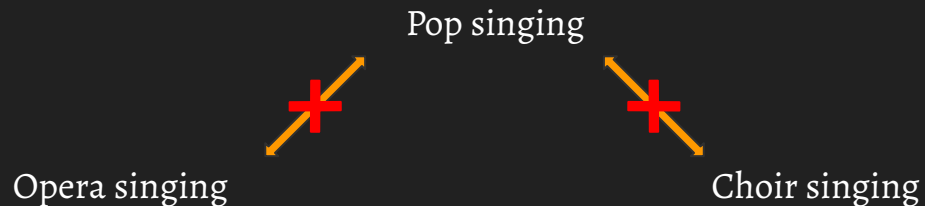
Mixed



Head

Acc. on different datasets

	Test Acc.		
Model	YouTube Pop	Opera	Choir
20ms 2D-CNN	92.01%	30.54%	40.49%



Contributions

- Dataset with labelled vocal registers (Pop / Opera / Choir)
- Discovered the time-frequency characteristics of timbre
 - Resulting in the selection of 2D kernels
- Discovered time resolution has minimal effects on model performance
 - Resulting in successful real-time prediction
- Produced a SOTA model with great performance in pop-singing

Future works -- data

Problem:

- P Dataset lack variety in melody/tempo
- P Only female pop singing voice prediction

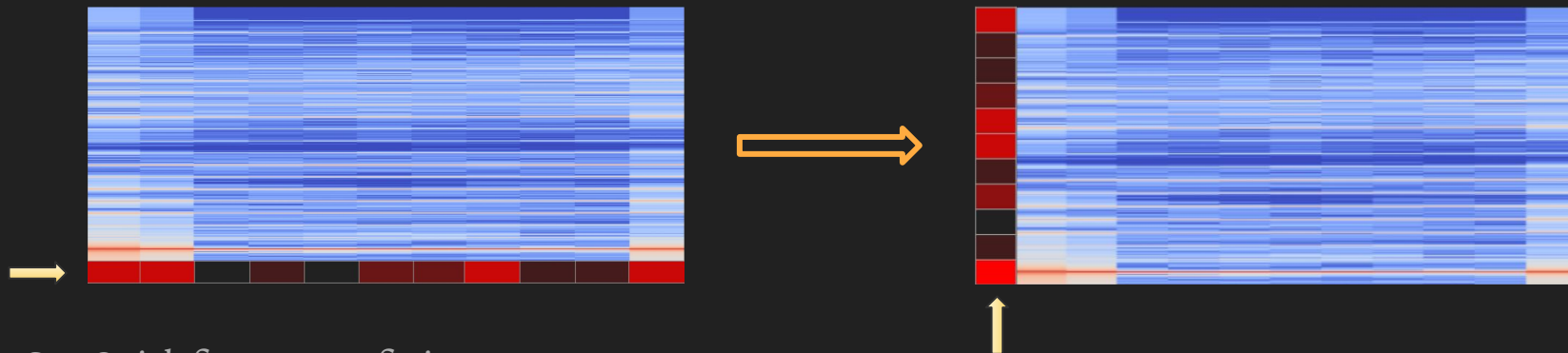
Solution:

- S **Dataset size** (65 min → 10+ hours)
- S Variety in data source

Future works -- methodology

S Attention along the timbral axis

- Reason: convolutions guard local regions, assuming that “image” patterns are usually adjacent
- Yet, timbral patterns are more irregular



S Quick fixes to overfitting:

- Pitch invariance
- Phoneme classification

THE END