Deep Learning Approaches for Sung Vowel Classification

Parker Carlson and Patrick J. Donnelly

evoMUSART 2024



Oregon State University

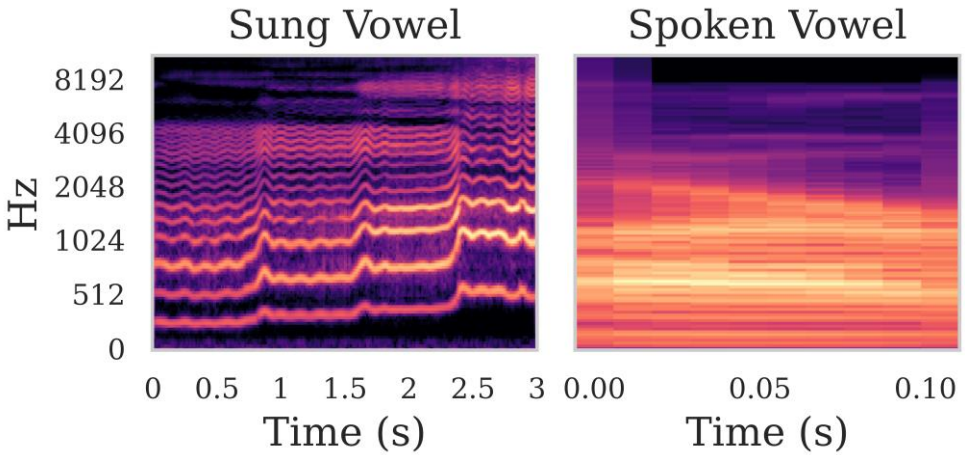
Introduction

Problem: Transcribe Spoken Words

- Spoken words => Automatic Speech Recognition (ASR)
  - Well-studied
  - Strong results on many languages
  - Struggle with changes in pronunciation and rate of speech
- Sung words => Automatic Lyric Transcription (ALT)
  - Less-studied, but quickly growing
  - Most approaches are word-level classification
- Phoneme Classification
  - Recognize the sounds, then find appropriate words using a language model

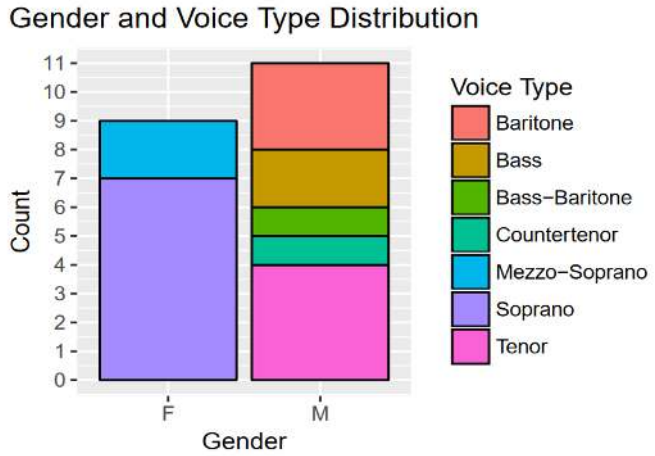
Speech Versus Song

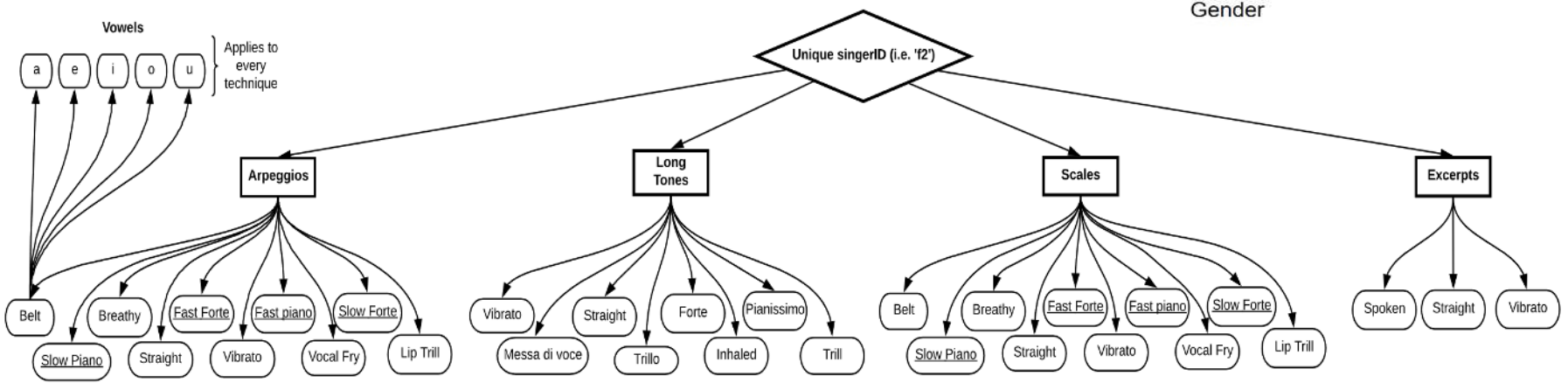
- Duration
- Acoustics (Formants)
- Male vs Female
- Altered Pronunciation



VocalSet

- Five Italian vowels per technique
  - /a/, /e/, /i/, /o/, /u/





Wilkins, J., Seetharaman, P., Wahl, A., Pardo, B.: Vocalset: A singing voice dataset. In: Proceedings of the 19th International Conference on Music Information Retrieval (ISMIR). pp. 468–474 (2018).

Dataset Preprocessing

- Include all techniques but “spoken”
- Remove silence
- Split into 3 second examples
  - Discard examples under 3s
- Standardize (0 mean, 1 variance)
- 5821 examples, 4.85 hrs

Vowel	Examples
a	1148
e	1147
i	1149
o	1190
u	1187

Tasks

Singer ID

- Classify which of the twenty singers is singing

Vocal Technique

- Classify which of the ten vocal techniques is being used

Vowel Classification

- Classify the vowel being sung

“VocalSet Replication” Tasks

Models

Convolutional Model

- Identical to VocalSet's convolutional baseline model
- 3 layers, ~112,000 parameters
- Selected to extend VocalSet's baseline findings to vowel classification task

Layer Name	Input	Conv1	BatchNorm1	MaxPool1	Conv2	BatchNorm2	MaxPool2	Conv3	BatchNorm3	MaxPool3	Dense1	Dense2
# of Units/Filters	3*44100	16	16	-	8	8	-	32	32	-	32	10/20
Filter Size, Stride	-	(1, 128), (1, 1)	-	(1, 64), (1, 8)	(1, 64), (1, 1)	-	(1, 64), (1, 8)	(1, 256), (1, 1)	-	(1, 64), (1, 8)	-	-
Activation function	-	ReLU	-	-	ReLU	-	-	ReLU	-	-	ReLU	softmax

Table 1. Network architecture. The input to the network is 3 seconds of time series audio samples from VocalSet. The output is a 10-way classification for vocal technique classification and a 20-way classification for Singer ID. The architecture for both classifiers is identical except for the output size of the final dense layer. For the dense layers, L2 regularization was set to .001.

Wilkins, J., et al.: Vocalset: A singing voice dataset. In: Proceedings of the 19th International Conference on Music Information Retrieval (ISMIR). pp. 468–474 (2018).

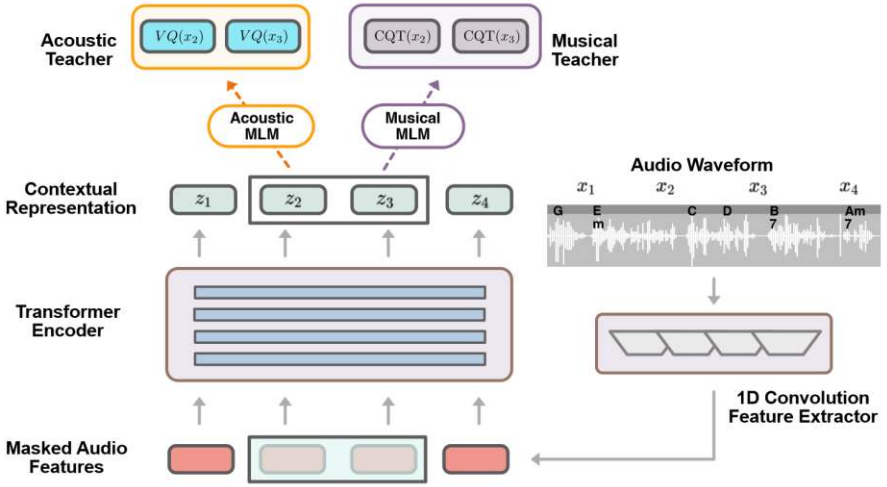
Recurrent Model

- Selected to be of similar depth and size to the convolutional model
- 3 layer BiLSTM, ~70,000 parameters
- 20% dropout, dense layers use L2 regularization of 1e-3
- Trained 200,000 batches of size 64 (same as CNN)



Transformer (MERT)

- 95M parameter pretrained musical transformer
- Pre-trained on music4all
  - 910 hrs of public music
- Use MERT to generate music embeddings, then a small feedforward network
- Fine-tune for 5 epochs

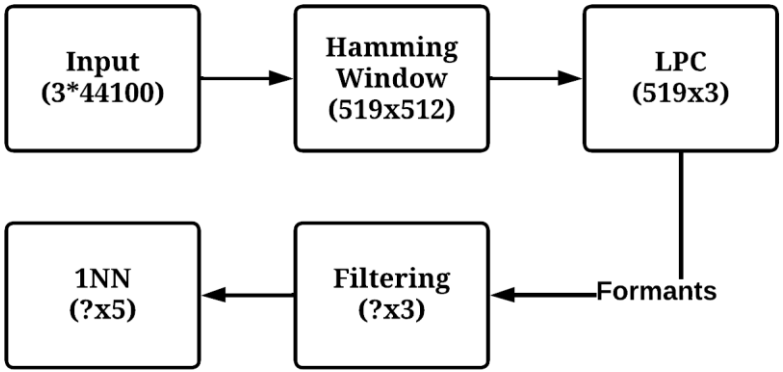




Li, Y., et al.: MERT: Acoustic music understanding model with large-scale self-supervised training. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). (2024).

Formant Model

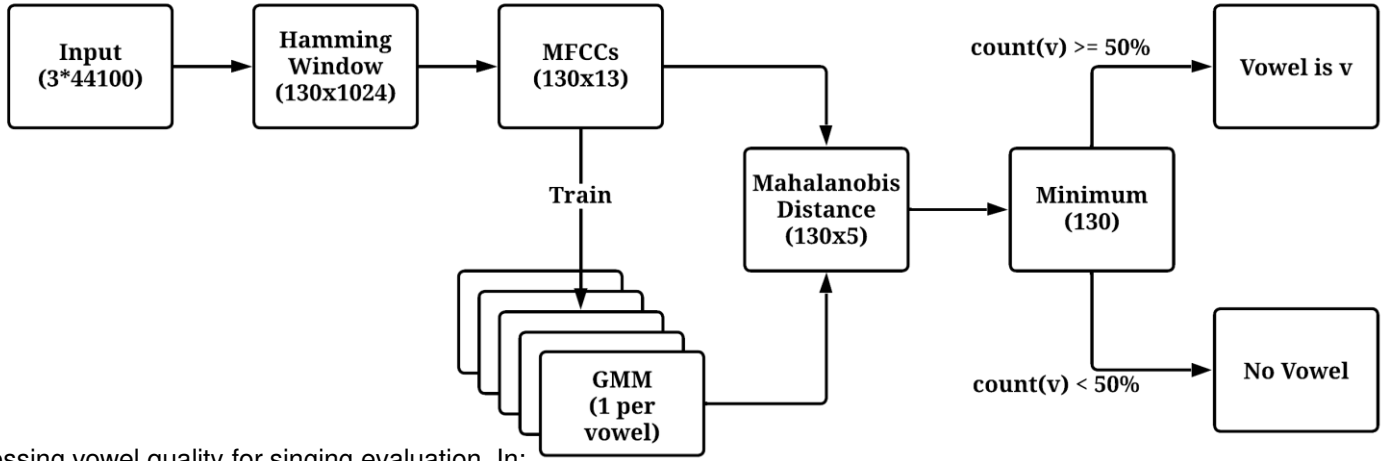
- Successfully applied to classify eight (spoken) Turkish vowels
- First three formants are extracted
- K-NN (k=1) classification with Manhattan distance
- If fewer than two formants were found, we discard the frame



Korkmaz, Y., Boyacı, A., Tuncer, T.: Turkish vowel classification based on acoustical and decompositional features optimized by genetic algorithm. Applied Acoustics 154, 28–35 (2019)

MFCC Model

- Successfully applied to classify five Hindi vowels
- Only other work on sung vowel classification to our knowledge
- One GMM trained per vowel, classification occurs on frame-level, not example level



Jha, M.V., Rao, P.: Assessing vowel quality for singing evaluation. In: National Conference on Communications (NCC 2012). pp. 1–5. IEEE (2012).

Results

VocalSet Replication

Model	Prec.	Recall	Accuracy				
			Top-1	Top-2	Top-3	M.	F.
Vocal Technique							
VocalSet [38]	0.676	0.619	-	0.801	0.867	-	-
Conv. (Ours)	0.628	0.607	0.619	0.753	0.814	-	-
MERT [24]	-	-	0.756	-	-	-	-
Singer ID							
VocalSet [38]	0.473	0.516	-	0.638	0.700	0.684	0.351
Conv. (Ours)	0.698	0.660	0.673	0.808	0.863	0.701	0.632
MERT [24]	-	-	0.780	-	-	-	-

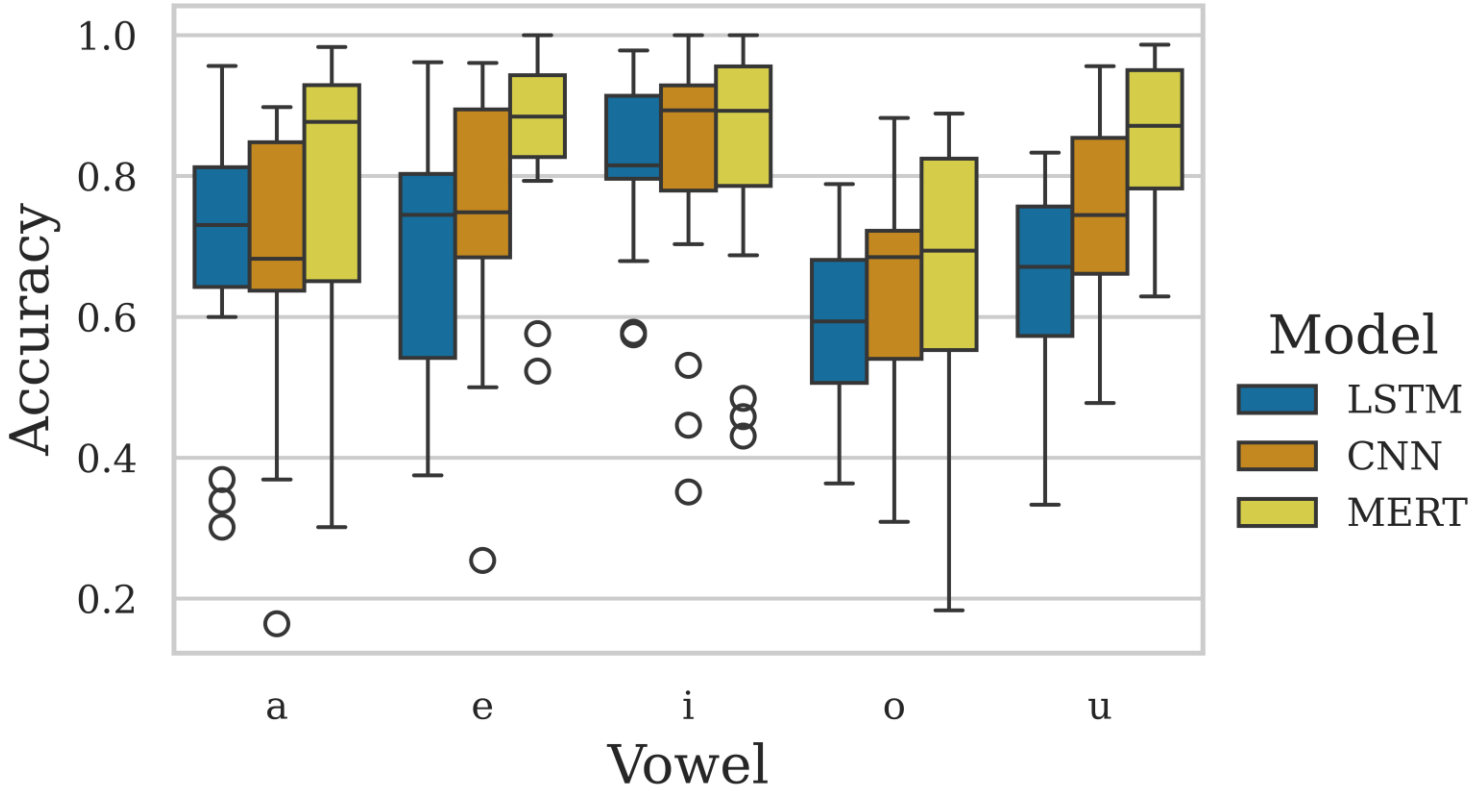
Worse

Better

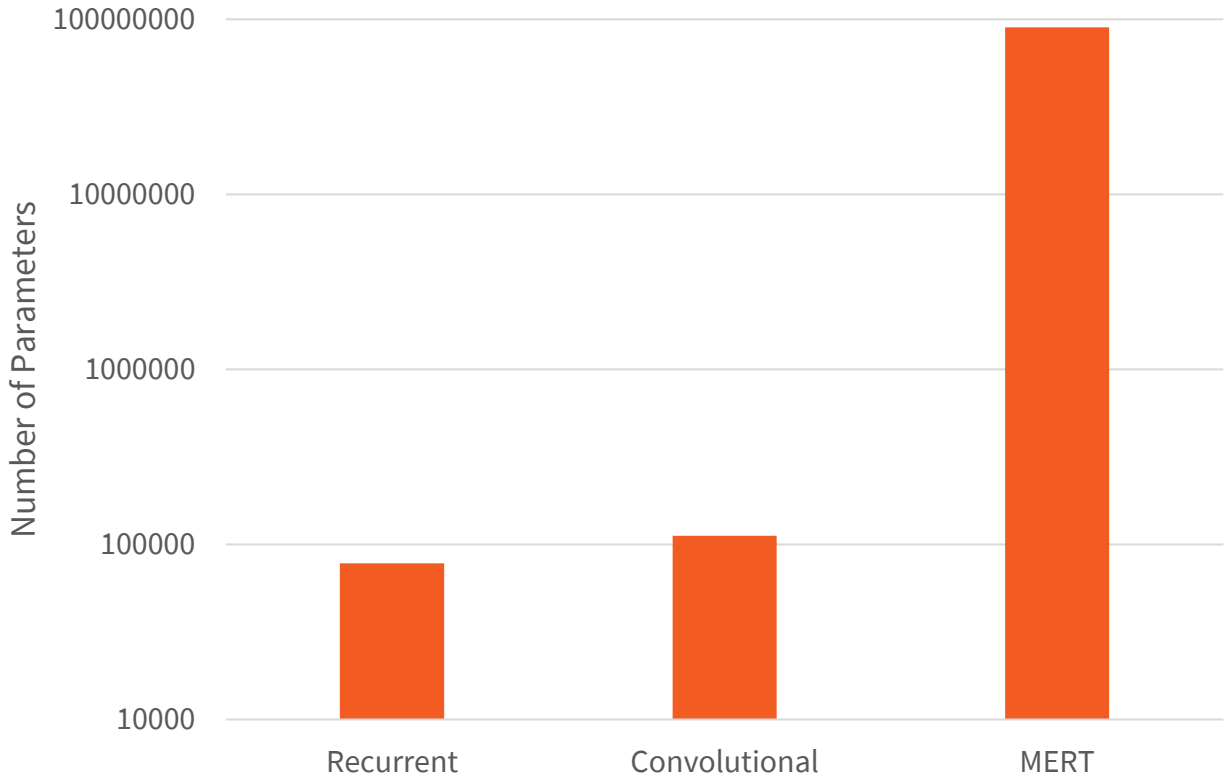
Vowel Classification

Model	Precision	Recall	Accuracy		
			Total	Male	Female
Neural					
Convolutional	0.723	0.719	0.718	0.749	0.674
Recurrent	0.685	0.685	0.684	0.714	0.640
MERT	0.799	0.797	0.796	0.827	0.752
Spectral					
Formants	0.200	0.200	0.201	-	-
MFCCs	0.041	0.199	0.203	-	-

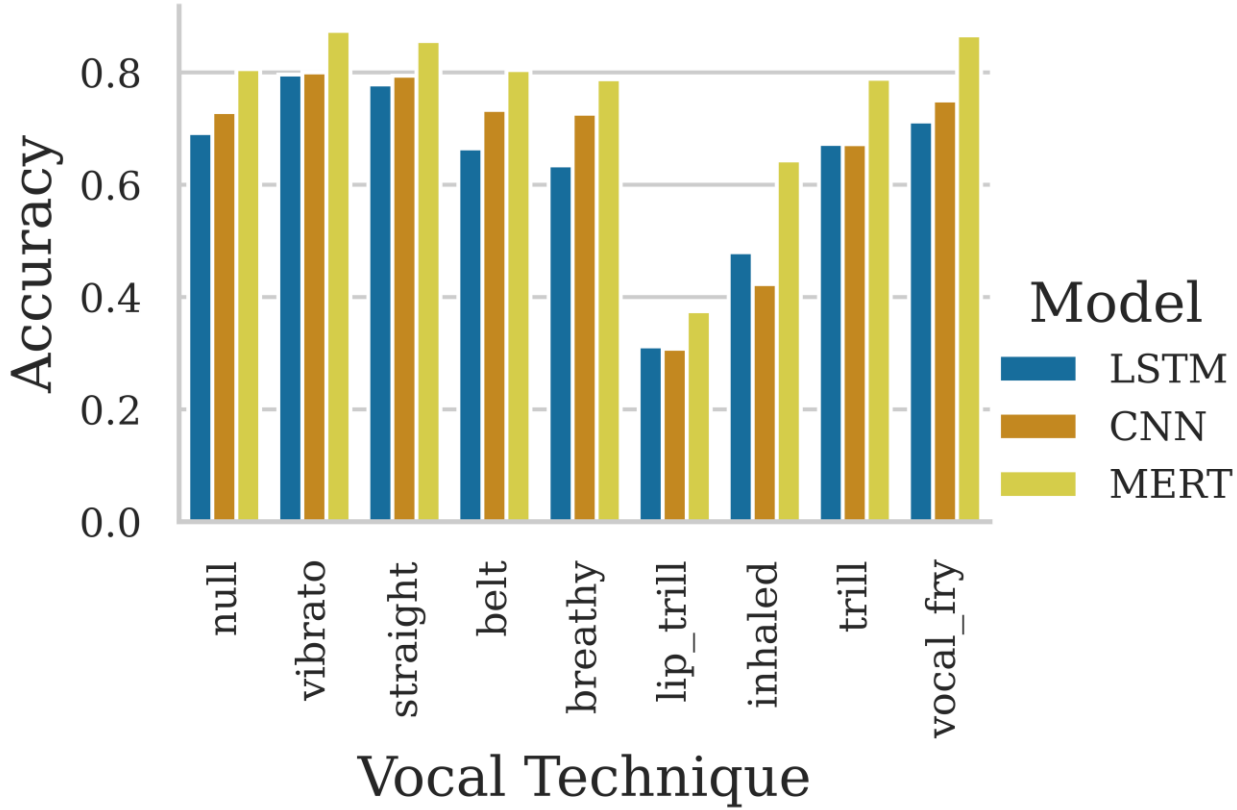
Vowel Classification



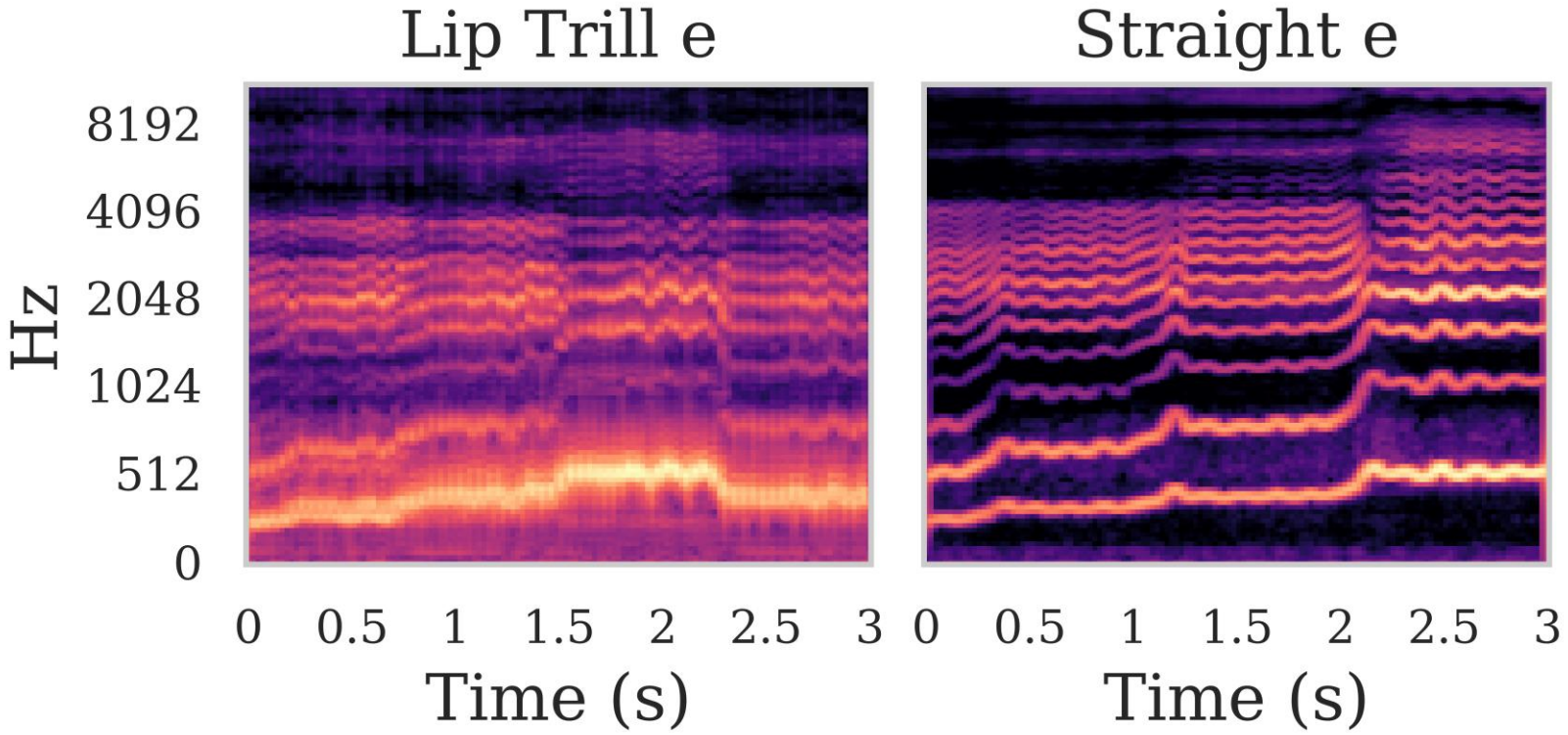
Model Size


100000000
10000000
1000000
100000
10000
Number of Parameters
Recurrent
Convolutional
MERT

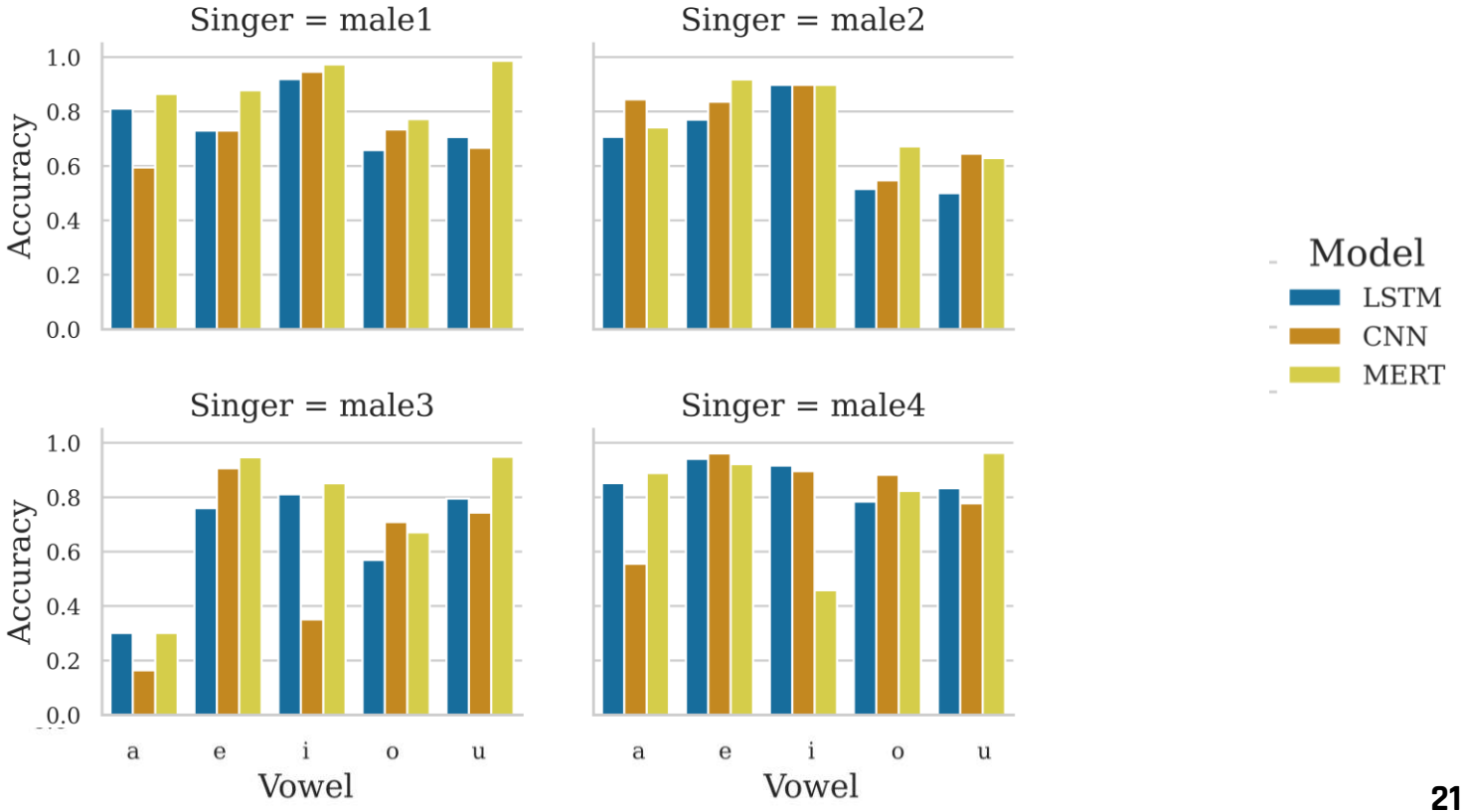
Vowel Classification



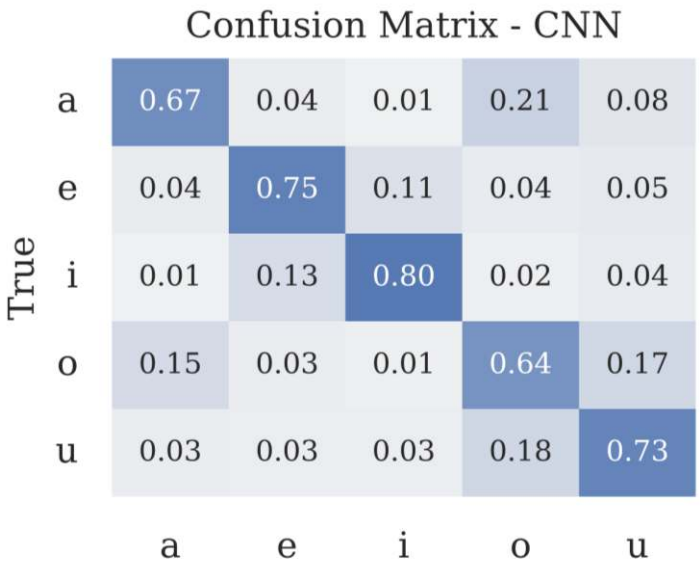
A Note on Lip Trill

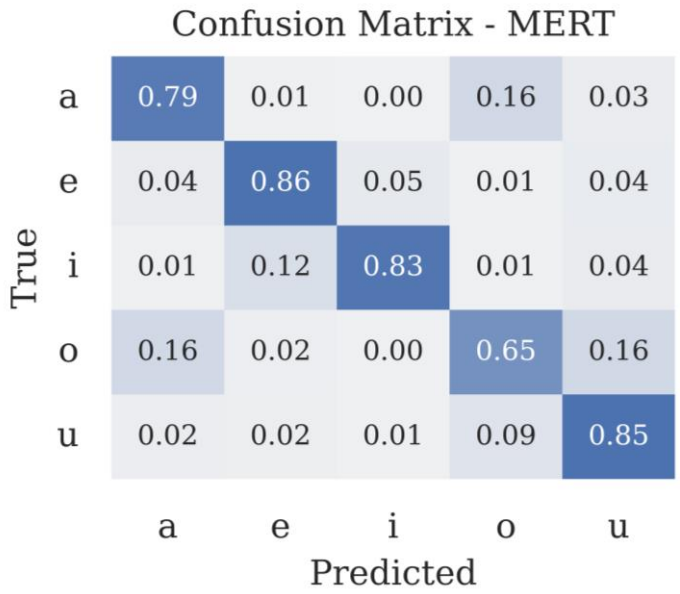


MERT Is Not Always Better



Models Make Similar Errors





Conclusions & Future Work

- Existing methods failed for sung vowel classification
- Neural methods were moderately successful, but room for improvement

Future Work

- Neural models trained on spectral features
- Larger recurrent and convolutional models to explore Pareto frontier
- Model performance gap on male and female singers
- More data, more vowels, more languages
- Test on speech datasets for generalization

Questions?