Speaking Professor Recognition

BA865 #7

Jiadai Yu, Shu Wang, Yulu Jiang

Table of Contents

Objective

Recognize QST professor through speaker classification

Motivation

1 Enhance student learning experience

- navigate speaking professor

2 New project challenges

- audio inputs, new packages

3 High data accessibility

- Echo360 lecture recording

Dataset & Pre-processing

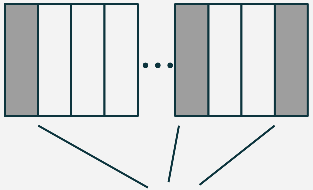
9 classes

- BA810 Sahoo
- BA775/780 Soltanieh-Ha
- BA875 Bellamy
- BA830 Fradkin
- ES710 Hutchinson
- BA865 Burtch
- BA820 Lee
- BA860 Lin
- ES720 McGinnis

Sample audio segment:

Segment For each class, ~20 minutes split into ~120 audio pieces
-> 1222 samples in total

Per audio piece:
10 seconds



Original vector:
[[left_1, right_1],
[left_2, right_2],
...
[left_48k, right_48k]]

Downsample 48 kHz -> 8 kHz

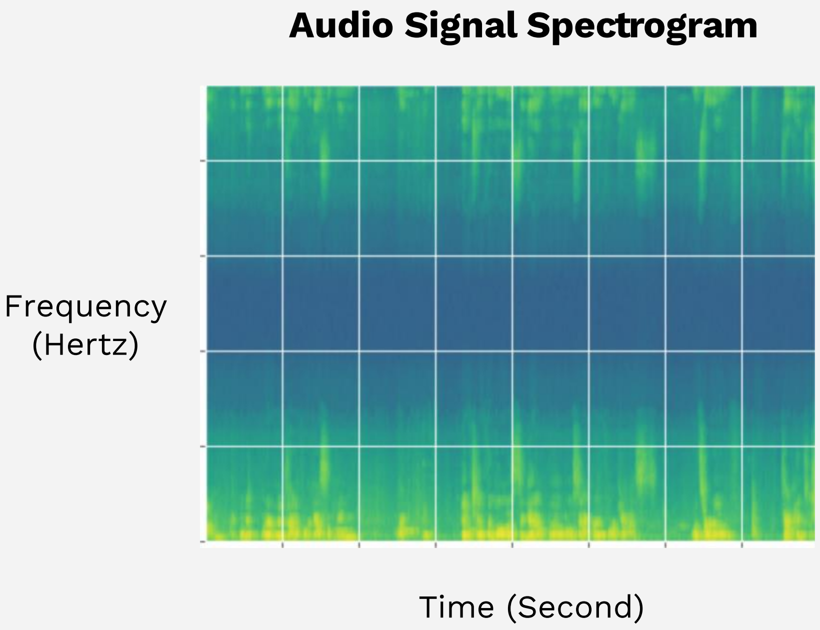
Same length,
Different amount of info stored

Pooling (LSTM only) 80k -> 4k

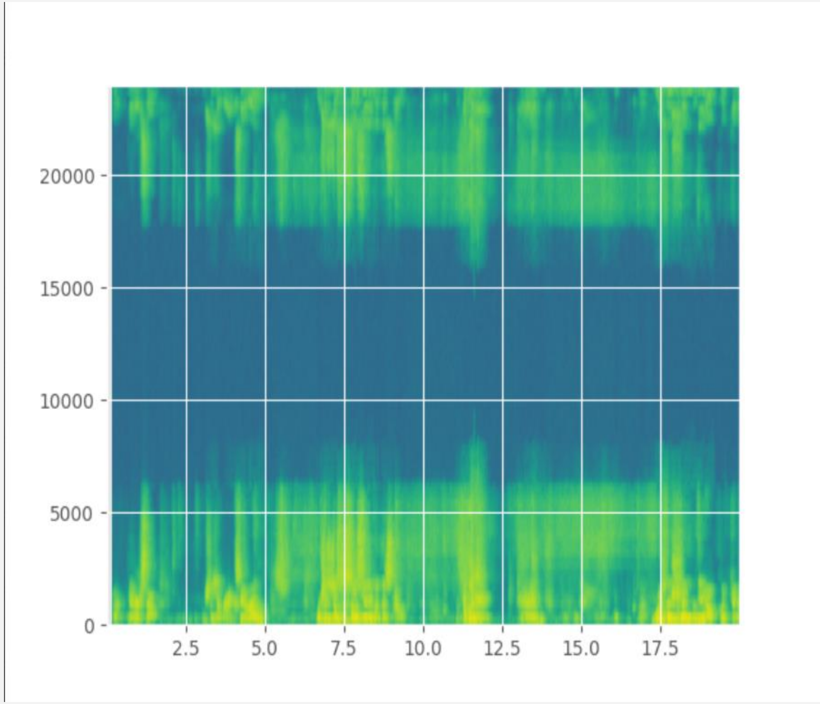
Padding Make sure equal length

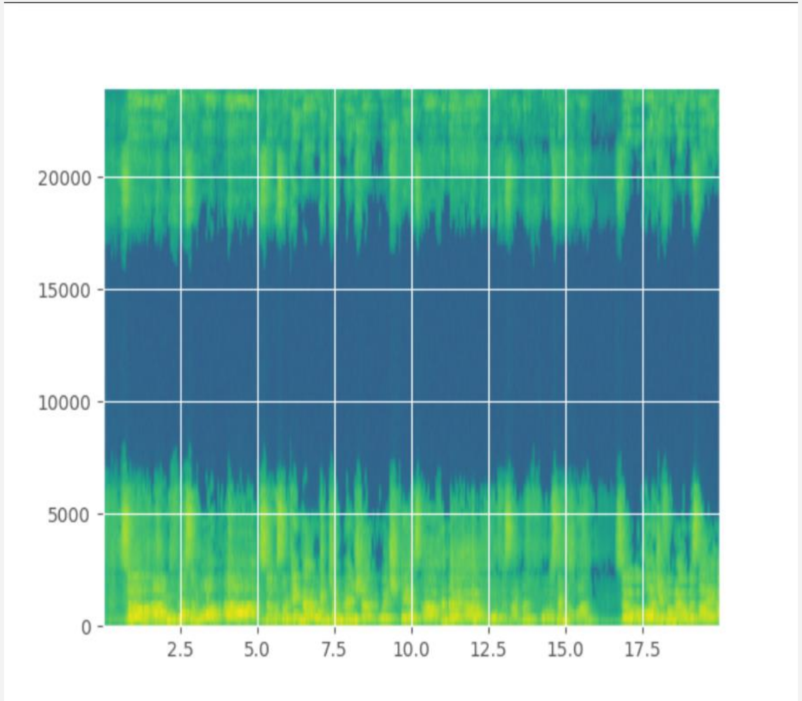
2D CNN Model Preprocessing

```
# visualization package
import pylab
# create a spectrogram of an audio signal
pylab.specgram()
```



Spectrogram Comparison

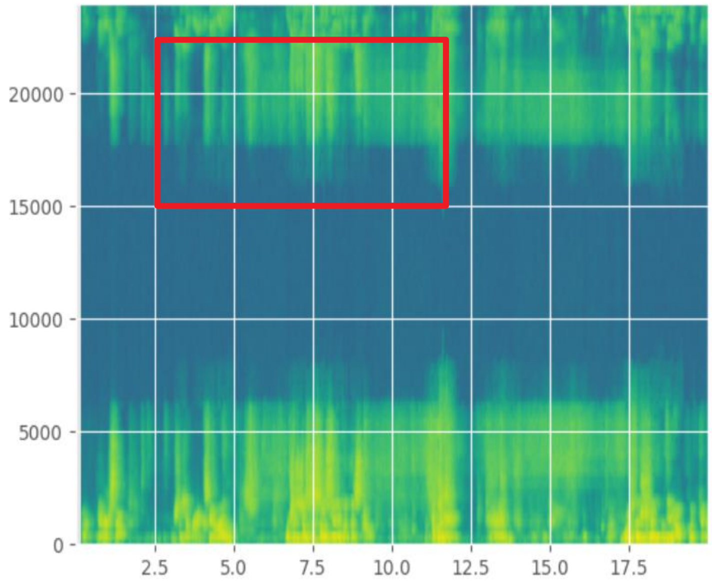




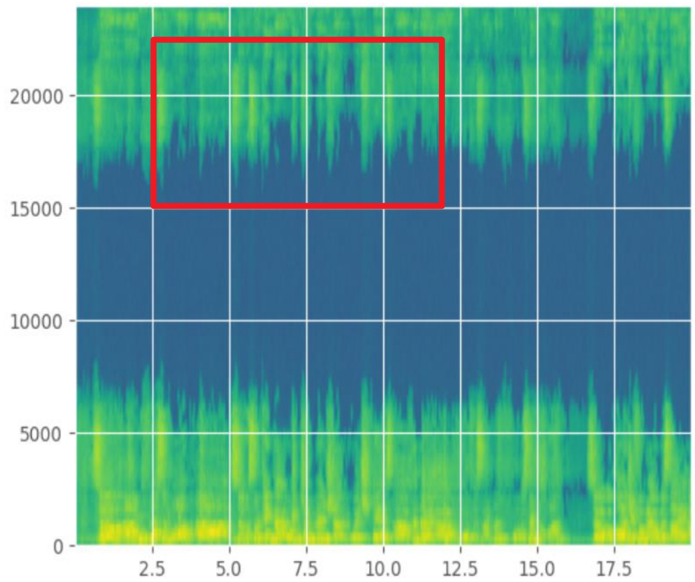
Takeaways:

1. Well-defined border for female voice spectrogram
2. Female vocal tracts are shorter and narrower than male ones.
3. Higher frequencies and shorter wavelengths.

Professor Burtch



Professor Lin



1D CNN Model

Sequence of numbers

Activation function:
ReLu for first/hidden layers;
Softmax for output layer

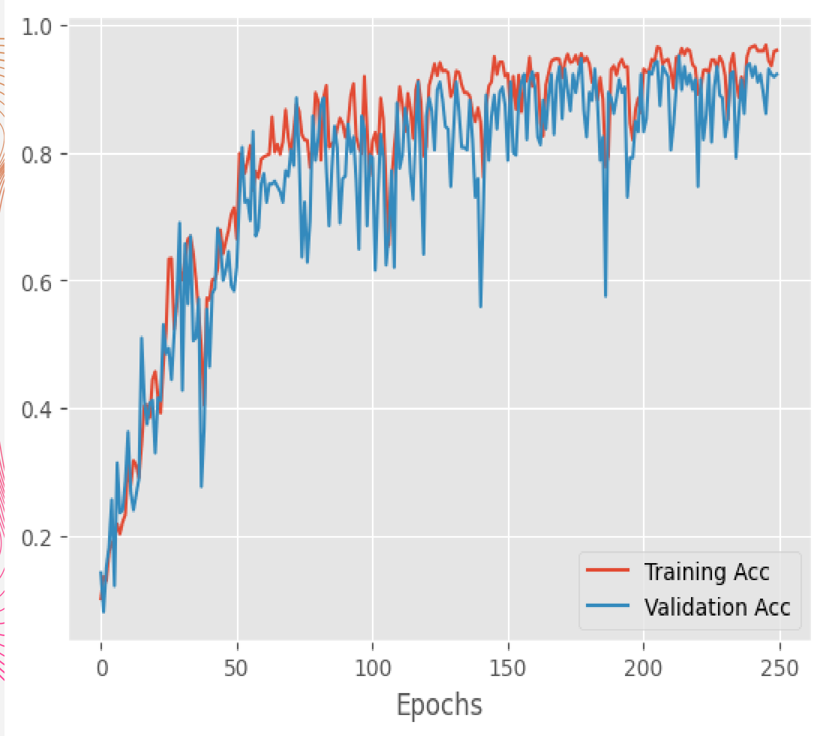
Loss function and Metrics:
Sparse_categorical_crossentropy
;
Sparse_categorical_accuracy

For **multiclass-label** classification

Structure of 1D CNN:
Conv1D->Max pooling
Conv1D->Average pooling
Flatten->Hidden Layers
Output Layer

Labels:
Give each professor a number from 0-8

Model Performance

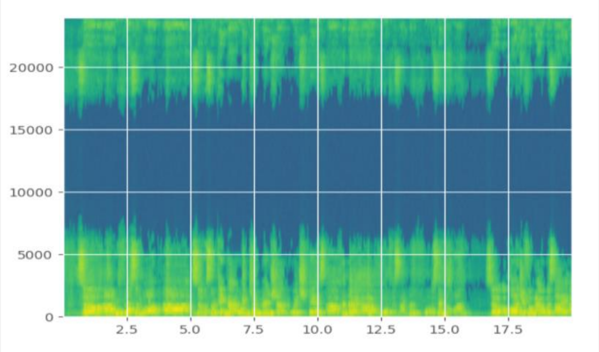


	precision	recall	f1-score	support
0	0.98	0.98	0.98	127
1	0.97	0.84	0.90	134
2	1.00	0.99	1.00	118
3	0.98	0.98	0.98	128
4	0.88	0.94	0.91	129
5	0.92	0.99	0.95	140
6	0.92	0.93	0.92	129
7	0.99	0.97	0.98	159
8	0.99	0.99	0.99	158
accuracy			0.96	1222
macro avg	0.96	0.96	0.96	1222
weighted avg	0.96	0.96	0.96	1222

2D CNN Model

Spectrogram

visual representation of the spectrum of frequencies of a signal as it varies with time



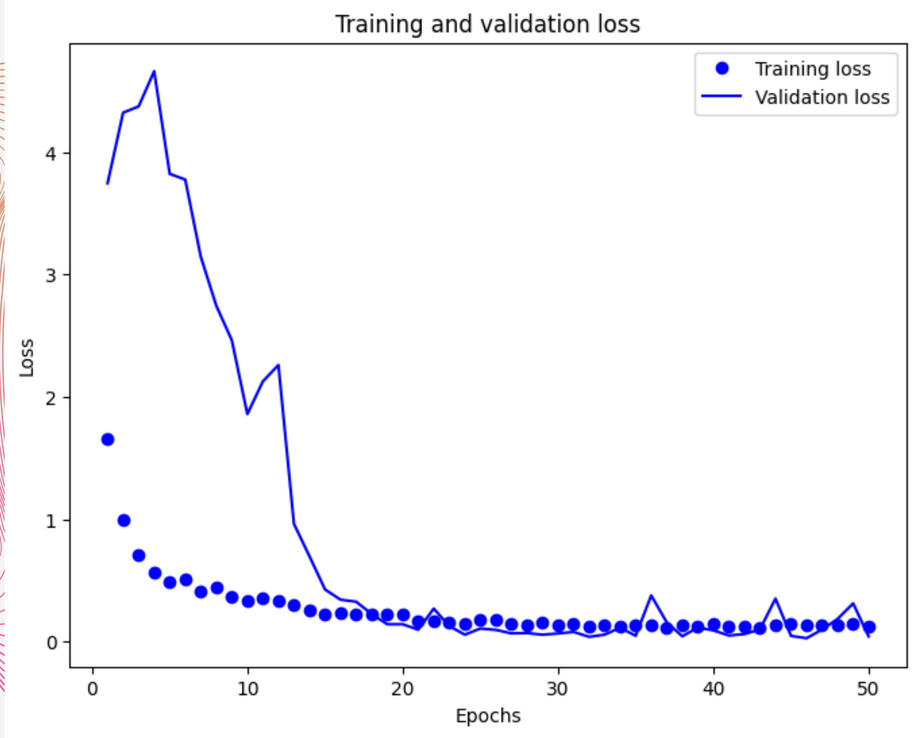
Activation/Loss function:
Same as in 1D CNN

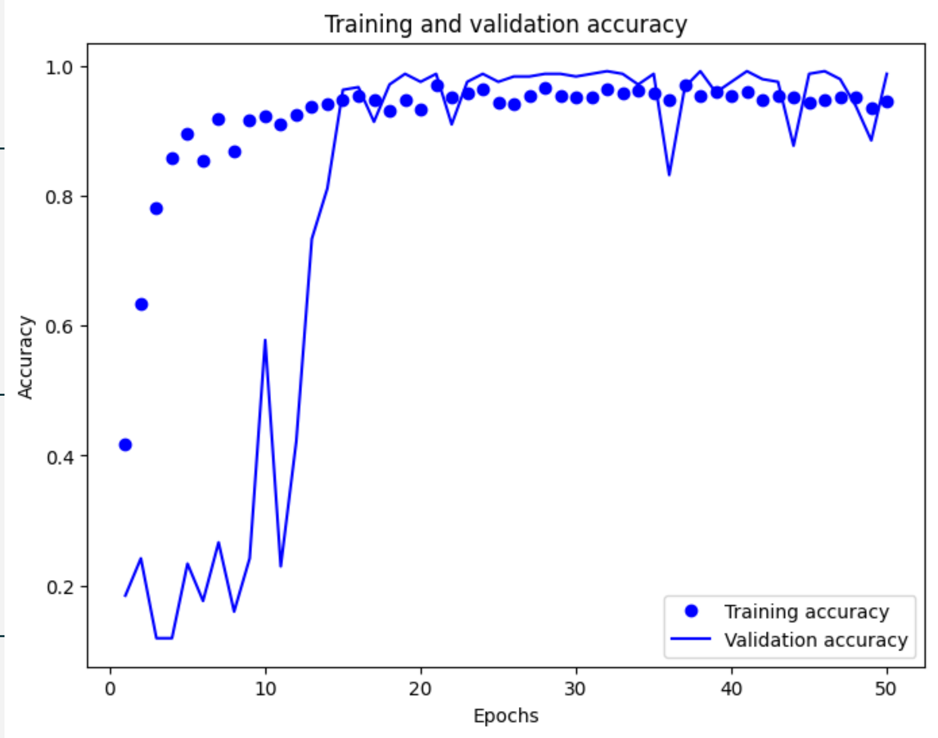
For **multiclass-label** classification

Image Augmentation:
Rescaling,
RandomFlip/Rotation

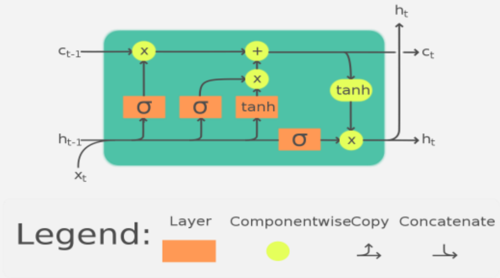
Structure of 2D CNN:
Conv2D->Batch Normalization->Max Pooling 2D->Batch Normalization
Flatten -> Dropout -> Output

Model Performance





LSTM Model

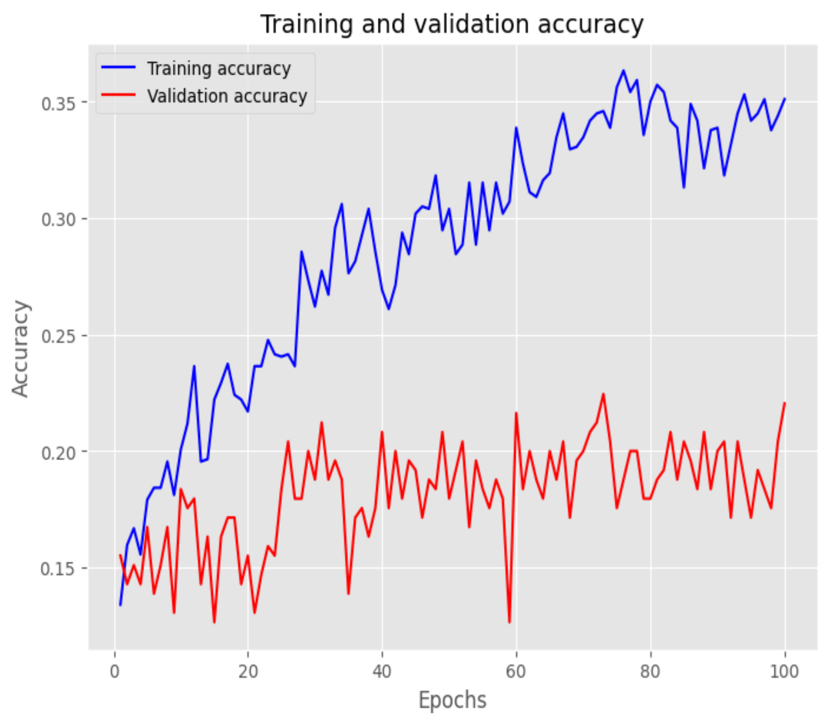


Activation/Loss function:
Same as in 1D CNN

For **multiclass-label** classification

Structure of LSTM:
LSTM -> Dropout -> Hidden Layer -> Output Layer

Model Performance



	precision	recall	f1-score	support
0	0.24	0.35	0.28	127
1	0.19	0.48	0.27	134
2	0.28	0.31	0.30	118
3	0.50	0.44	0.47	128
4	0.44	0.40	0.42	129
5	0.67	0.38	0.48	140
6	0.57	0.18	0.27	129
7	0.72	0.29	0.41	159
8	0.53	0.51	0.52	158
accuracy			0.37	1222
macro avg	0.46	0.37	0.38	1222
weighted avg	0.47	0.37	0.39	1222

Fine-tuning on 1D CNN

2D CNN takes much longer time to fine-tuning
1D CNN has a more stable performance compared to 2D-CNN

Hyperparameters:
Kernel Size,Dropout Rate,Batch Size

Grid Search results for best model:
Training accuracy -> 95%
Validation accuracy -> 92%
More complex models lead to overfitting

Conclusions

	1D CNN	2D CNN	LSTM
Validation Accuracy	96%	93%	20%
Pros	Efficient, and growing accuracy with fine tuning and more epoches	Good at capturing both local and global dependencies, high accuracy	Simple, easy to use
Cons	Not be as effective for capturing global dependencies in data	Computationally expensive	Low accuracy because of the length of sequences

Implications

- Biometrics & Security authentication (voice prints)
  Confirm the identity of the speaker
- Voice-controlled interfaces
  Customized services
- Natural language processing
  Take the accents or speaking habit of the speaker into consideration, make speech-to-text more accurate
- Academic tool
  Automatically identify and verify the course to which the recording file belongs, and help to manage the learning materials better

Limitations & Improvements

- Data collection
  The audio files are manually extracted from Echo360, so the efficiency and data size is limited.
- Preprocessing
  Professor’s speech is mixed with noise, reverberation and classroom discussions of students
- Computational power
  2D CNN has high accuracy, but GPU quickly ran out of RAM
- Network structure
  CRNNs, GANs, etc.
- Performance evaluation
  Precision, recall and robustness

Bonus Speaker Test

Audio **mis**classified as Mohammad's voice: FP/(FP+TP) = 0.03

Gordon's audio being **in**accurately classified: FN/(TP+FN) = 0.01

	precision	recall	f1-score	support
0	0.98	0.98	0.98	127
1	0.97	0.84	0.90	134
2	1.00	0.99	1.00	118
3	0.98	0.98	0.98	128
4	0.88	0.94	0.91	129
5	0.92	0.99	0.95	140
6	0.92	0.93	0.92	129
7	0.99	0.97	0.98	159
8	0.99	0.99	0.99	158
accuracy			0.96	1222
macro avg	0.96	0.96	0.96	1222
weighted avg	0.96	0.96	0.96	1222

BA810 Sahoo	8.4573139e-06
BA775/780 Soltanieh-Ha	9.9033117e-01
BA875 Bellamy	3.5032605e-05
BA830 Fradkin	8.5746113e-04
ES710 Hutchinson	3.1385373e-04
BA865 Burtch	4.6199220e-05
BA820 Lee	8.4024193e-03
BA860 Lin	1.2701168e-08
ES720 McGinnis	5.3097924e-06

Takeaways:

1. Similar voice feature across male professors
2. Limited sample size to identify feature in detail

References

- Beigi, H. (2011). Speaker Recognition. In: Fundamentals of Speaker Recognition. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-77592-0_17 [Accessed on: April 25th]
- Burtch, G. (2023). Github resource page for BA865. https://github.com/gburtch/BA865-2023 [Accessed on: April 20th]
- Ramgire, J. B., & Jagdale, S. M. (2016). A survey on speaker recognition with various feature extraction and classification techniques. International Research Journal of Engineering and Technology, 3(04), 709-712.
- Echo360 https://echo360.org/ [Accessed on: April 15th]

Thanks!

Any questions?