Frame-level Speaker Embeddings for Text-Independent Speaker Recognition and Analysis of End-to-end Model

Suwon Shon, Hao Tang, James Glass

MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), Cambridge, MA, USA

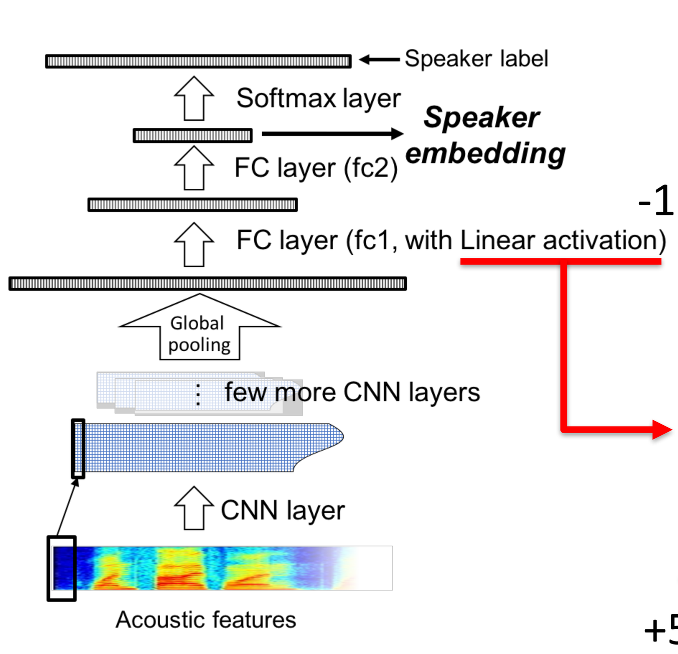
Motivation

- Speaker embeddings from neural network based End-to-end models shows impressive performance on speaker verification
- This paper analyzes how neural network model identify a speaker's characteristic when non-parallel speech input (text-independent) is given
- We modified a typical neural network-based end-to-end model to extract frame-level speaker embeddings from every layer
- After training is done, we fed the TIMIT dataset to analyze the model at the phoneme and broad class level with auxiliary tasks

We hypothesized that the network will pay more attention to how the phonemes are pronounced than what the phonemes are

Speaker Embeddings with Linear Activation

- CNN and Fully Connected layers
  - Remove ReLU activation function to extract robust speaker embeddings



<Network architecture for end-to-end system>

Table 1: Results on the Voxceleb1 test set. ReLU is applied after every layer.

	EER	$DCF_{p=0.01}$	$DCF_{p=0.001}$
fc1 (LDA+PLDA)	**6.2**	**0.53**	0.70
fc2 (LDA+PLDA)	6.9	0.55	**0.65**

-11%

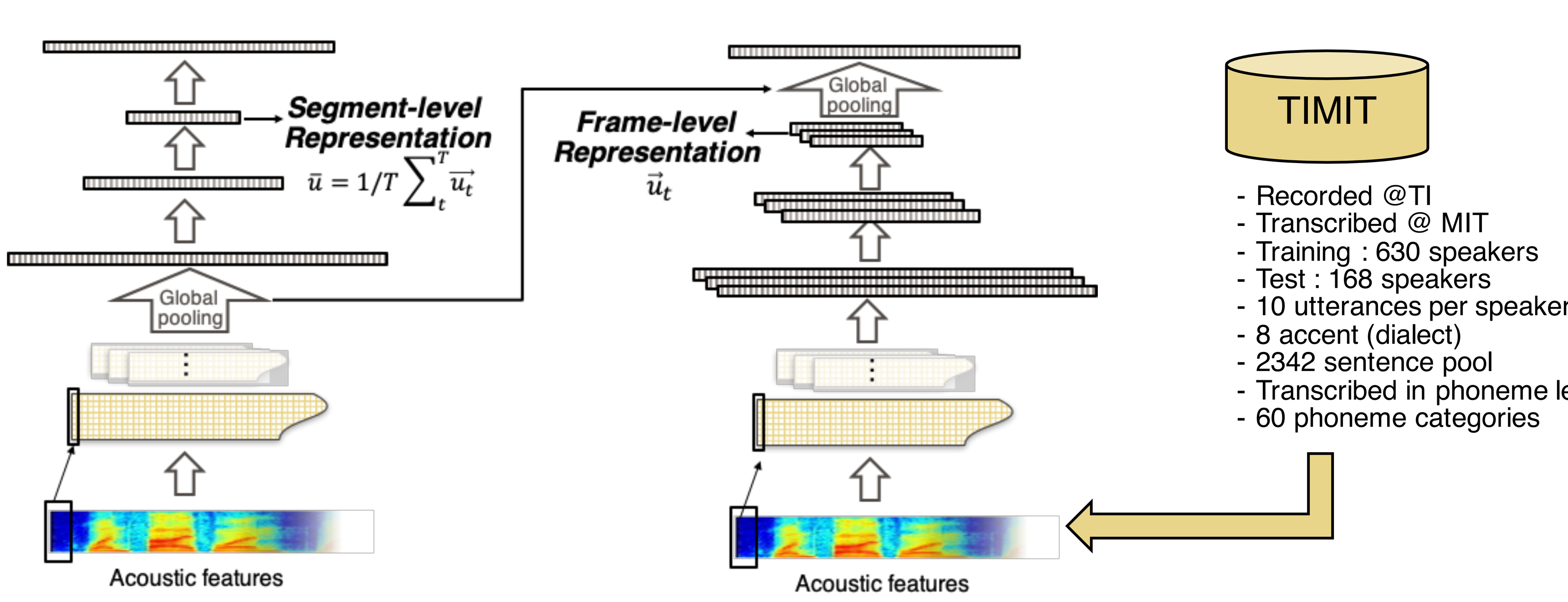
Table 2: Results on the Voxceleb1 test set. ReLU is applied after every layer except after fc1.

	EER	$DCF_{p=0.01}$	$DCF_{p=0.001}$
fc1 (LDA+PLDA)	6.2	0.51	0.69
fc2 (LDA+PLDA)	**5.9**	**0.50**	**0.62**

+5%

Modifying Structure for Frame-level Representation

- After training is done, the global average pooling layer is moved to be after the last hidden layer
  - *So, we can get a frame-level representation in every layer*
  - *Use TIMIT dataset to analyze the network layer by layer, epoch by epoch*

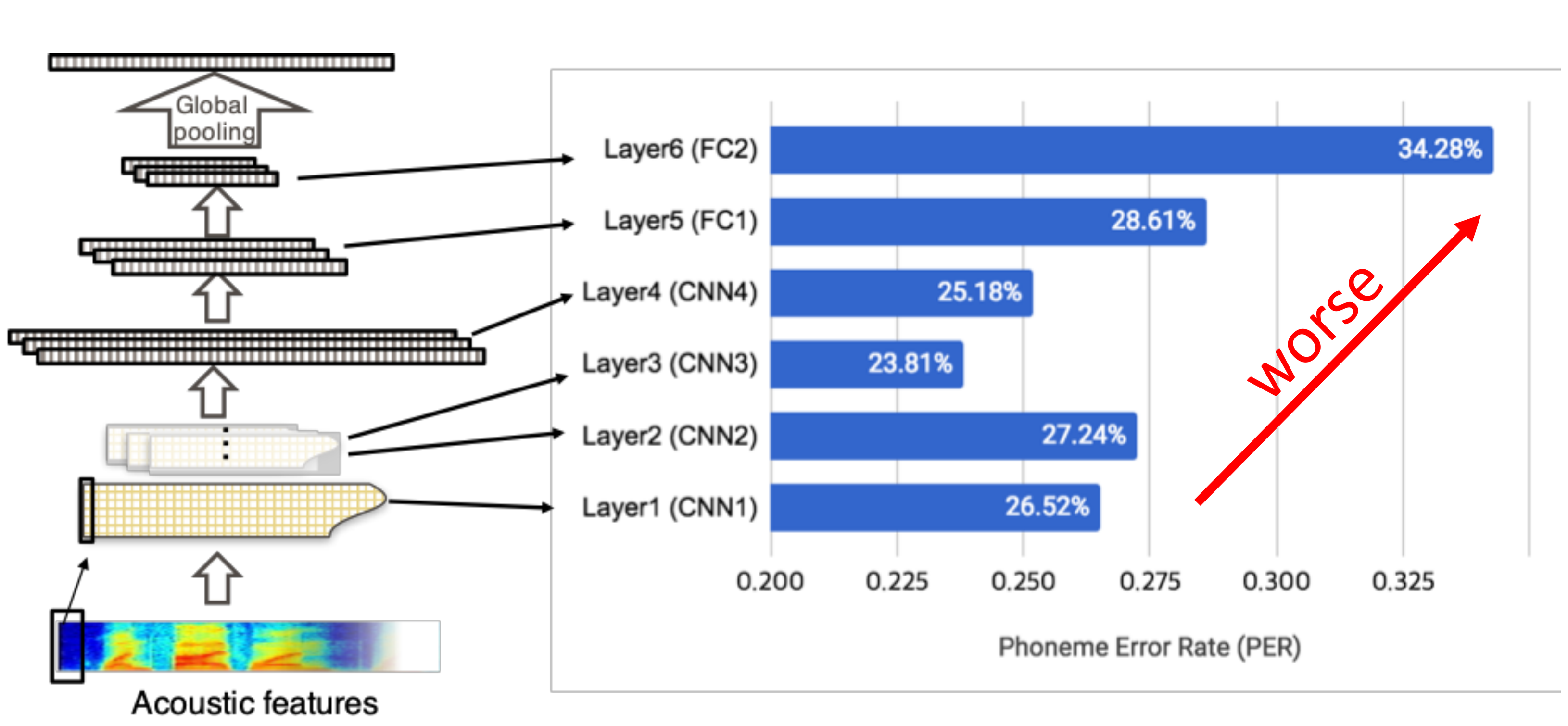


- Recorded @TI
- Transcribed @ MIT
- Training : 630 speakers
- Test : 168 speakers
- 10 utterances per speaker
- 8 accent (dialect)
- 2342 sentence pool
- Transcribed in phoneme level
- 60 phoneme categories

Experiments

<Phoneme Recognition>

- Evaluate the Phonetic Error Rate (PER) using a representation from each layer

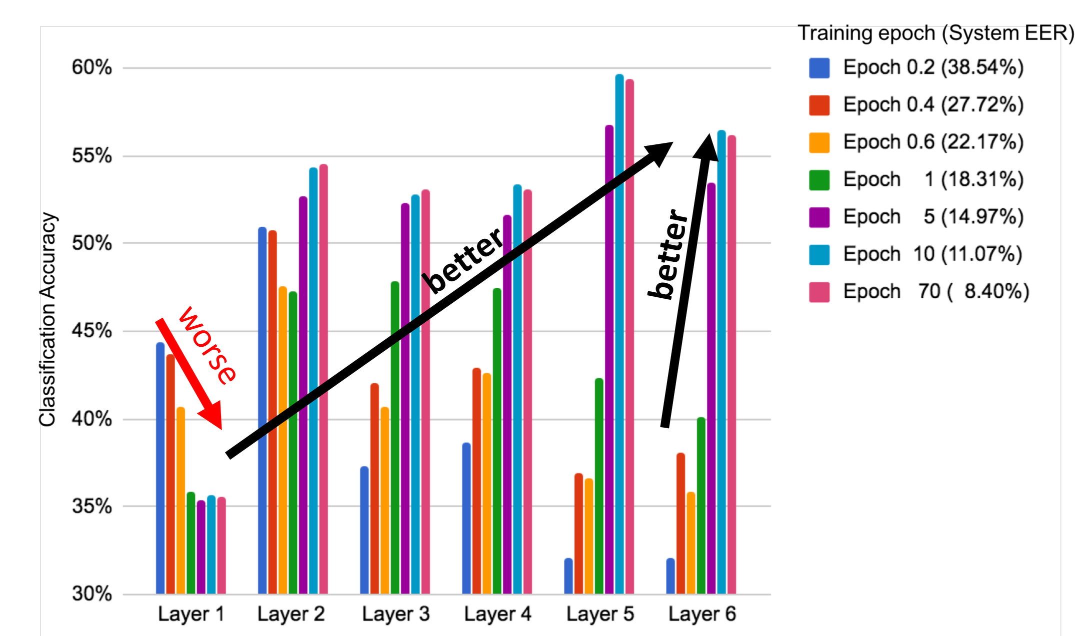


Phonetic identification does not seem to be important for discriminating speakers

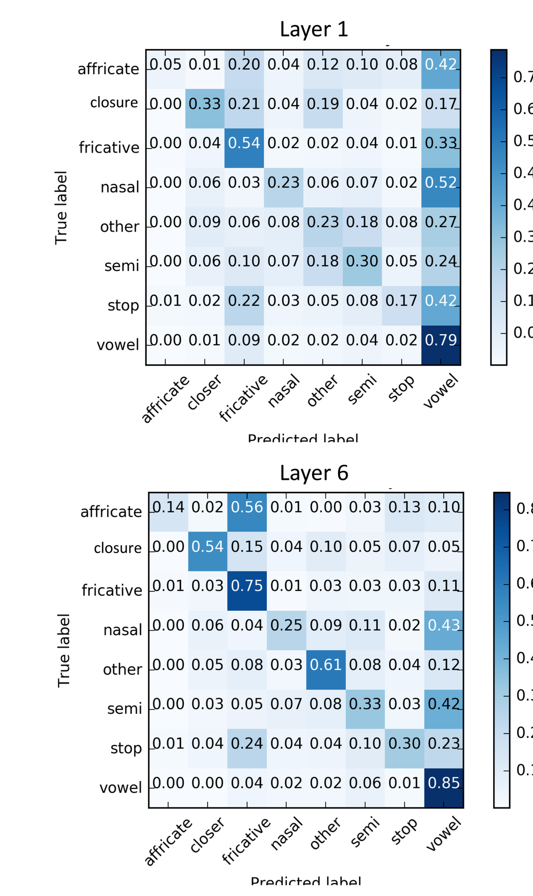
Experiments

<Broad-class Phonetic Classification>

- Segment TIMIT dataset to have a single phoneme in each segment

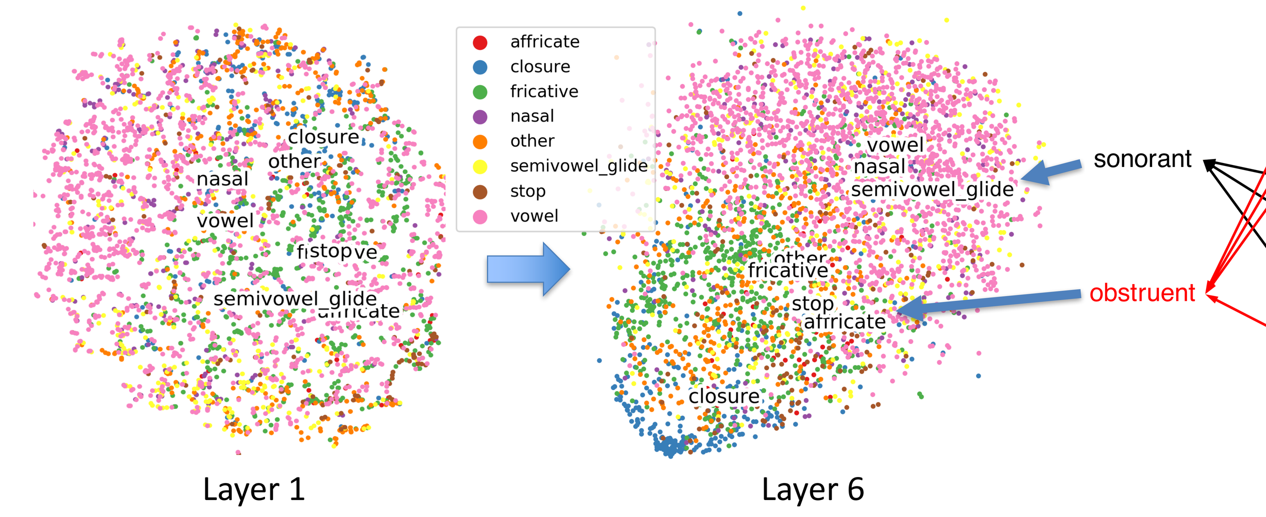


<Broad-class classification accuracy>



<confusion matrix @epoch 70 >

After training, the model learns to distinguish phonetic classes well

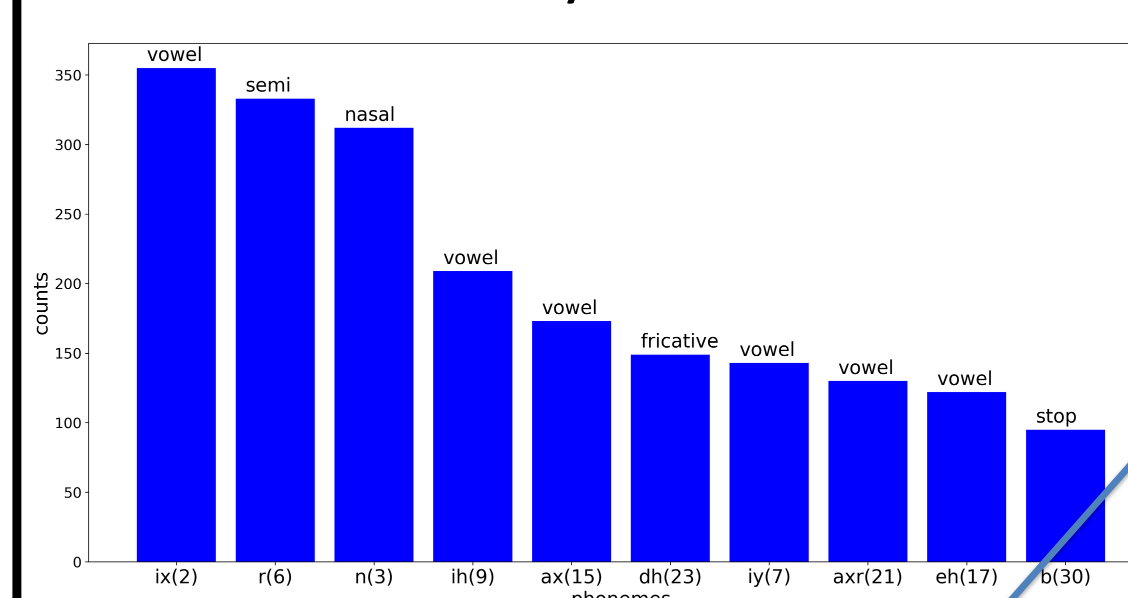


Broad-class	Symbol
Affricate	jh, ch
Closures	bcl,dcl,gcl,pcl,tck,kcl
fricative	s,sh,z,zh,f,th,v,dh
Nasals	m,n,ng,em,en,eng,nx
Semivowels and Glides	l, r, w, y, hh, hv, el
Vowels	iy, ih, eh, ey, ae, aa, aw, ay, ah, ao, oy, ow, uh, uw, ux, er, ax, ix, axr, ax-h
Stops	b, d, g, p, t, k, dx, q
others	pause(pau), epenthetic silence(epi) , start and end silence (h)

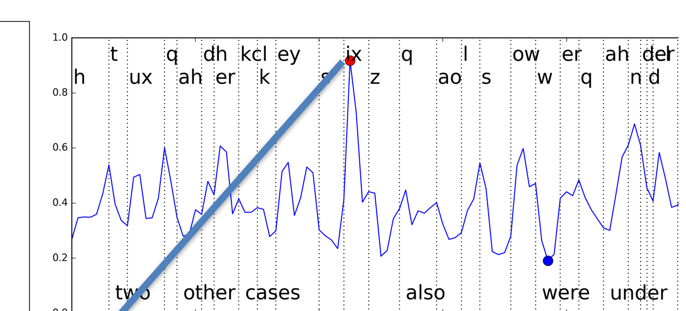
<t-SNE scatter plot of the representation>

The model classifies the phones into broad categories distinguished by degree of constriction

<Critical Phones and analysis in frame-level >

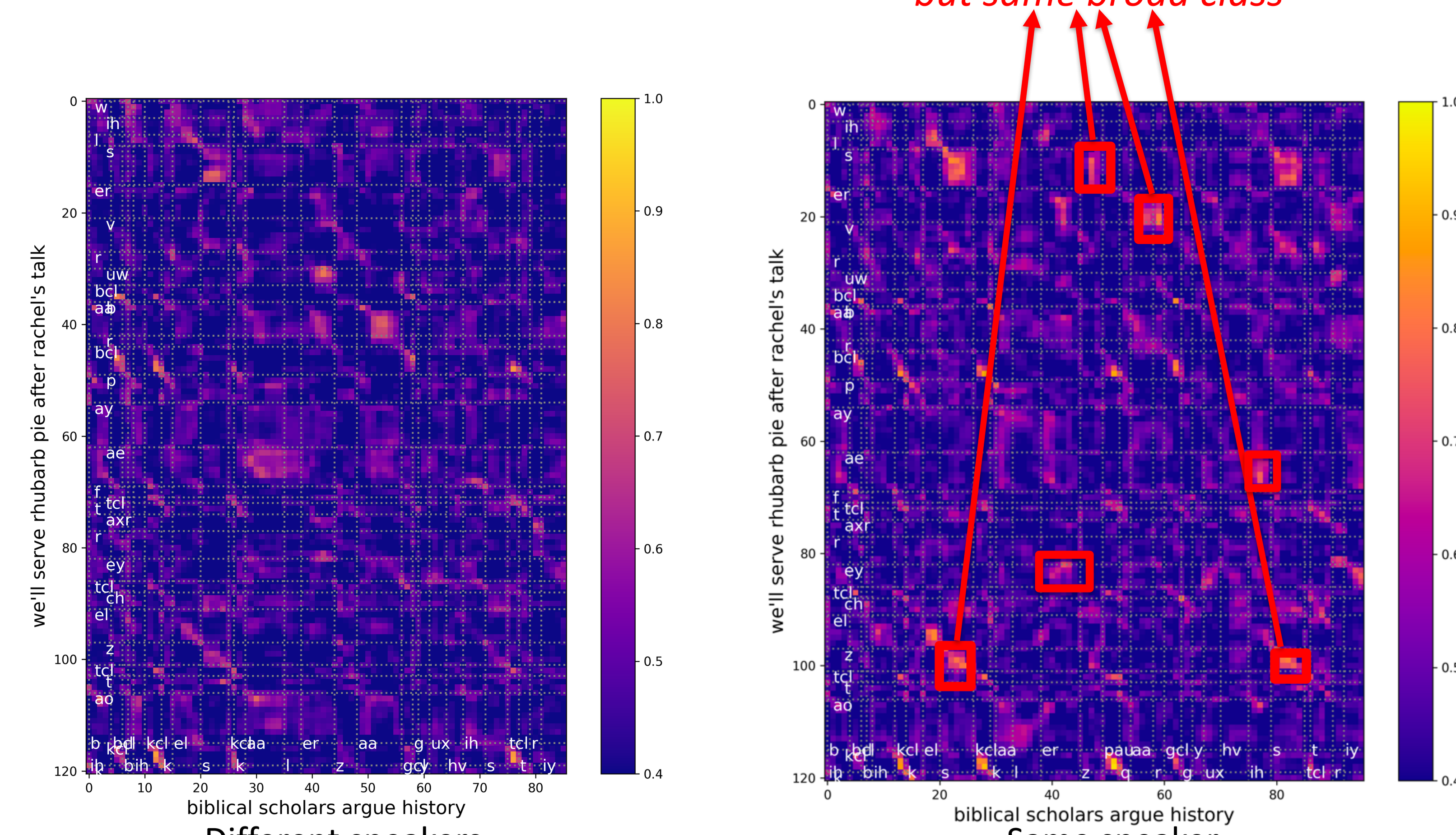


<Histogram of highest cosine similarity of phones in TIMIT test set>



<Frame by sequence cosine similarity>

Finding similarity between different phones but same broad class



Different speakers

Same speaker

<Frame by frame cosine similarity>

Conclusion

- We modified an end-to-end model to obtain a frame-level representation of the speaker embedding
- From our analysis, we attempt to better understand how the speaker recognition model extracts a discriminative representation
- The analysis provides some insight on the model and also is an important tool to assess the quality of the trained models
- The frame-level speaker embedding has other possible uses for applications such as acoustic modeling, text-to-speech synthesis and so on