CS 753: Explainability of ASR

Nikhil Chavanke: 170050017

Siddhesh Pawar: 17D170011

Problem Statement

Probling the transformer based architectures to understand what do the hidden layers actually represent

Using the representations from hidden layers of pretrained models for various downstream tasks

To investigate, where in the (end2end) model are attributes such as speaker accent, tone, captured

Datasets

- CMU Arctic Dataset - **gender and accents**
  18 speaker accents
  Around 1300 sentences per speaker
  Data for native as well as non native english speakers 16000 Hz

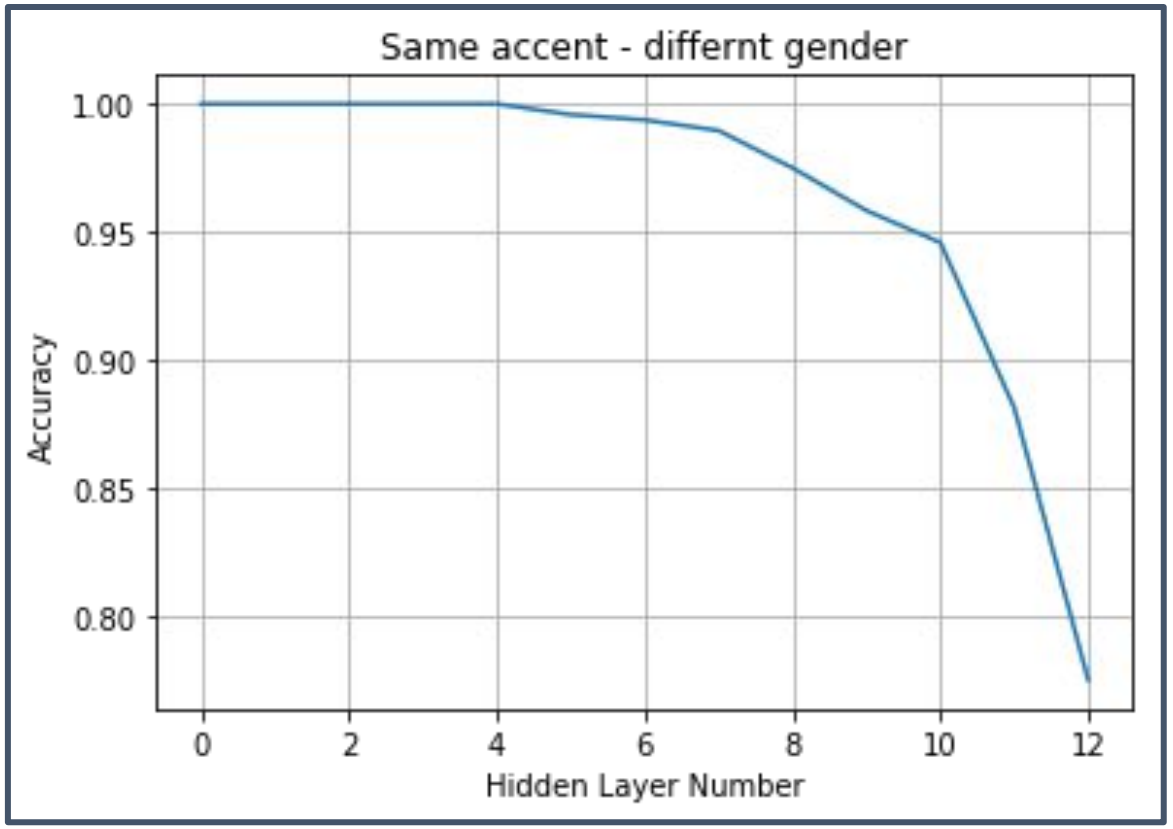
- TIMIT Dataset - **phones and speakers**
  630 speakers ,8 dialects(we used around 50 speakers)
  Word level and phonetic detail for each sentence
  Per speaker - 10 sentences

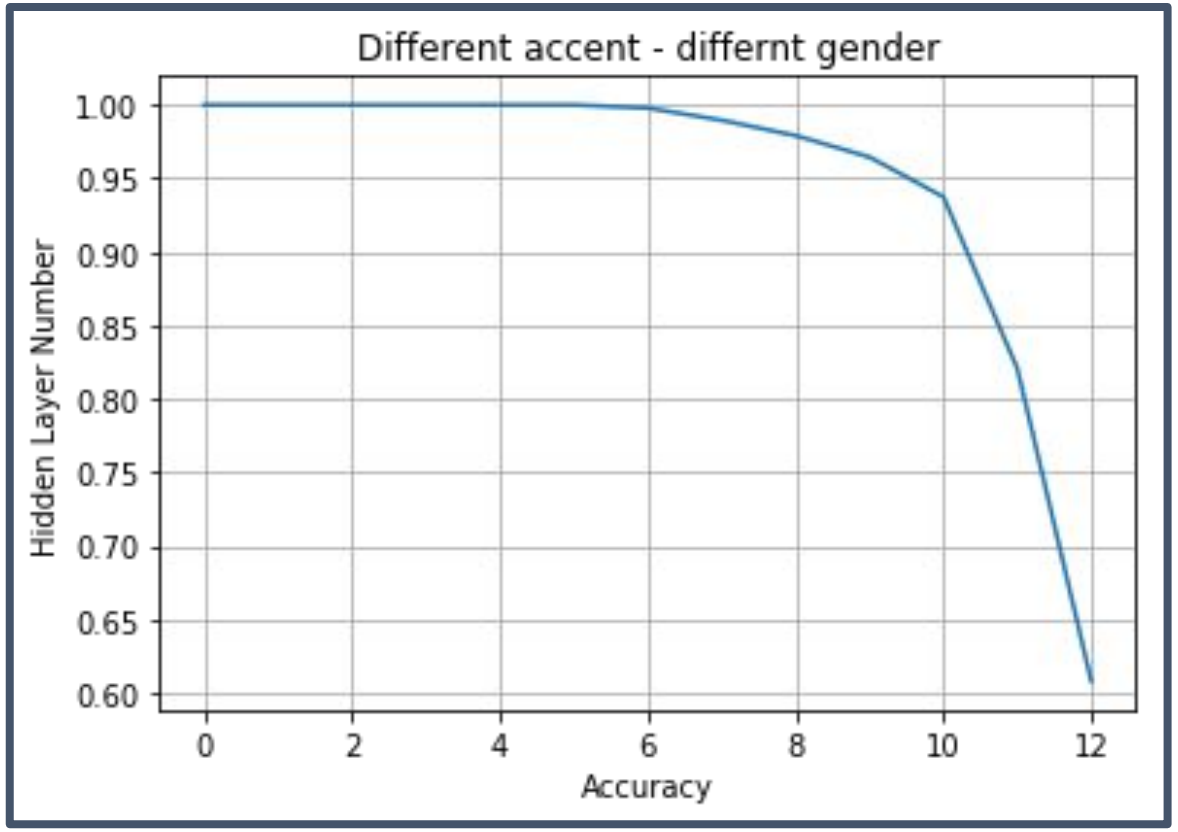
- Speech Commands - **words**
  35 Commands

Techniques used

- Probing the attention models to get activations Wav2Vec2.0
- SVC (Support vector Classifiers) and other primitive classifiers
- t-SNE (t-distributed Stochastic Neighbor Embedding) for visualization
- PCA (Principal Component Analysis) for assisting t-SNE
- JS divergence

Results - Determining Gender based on attentions in different layers

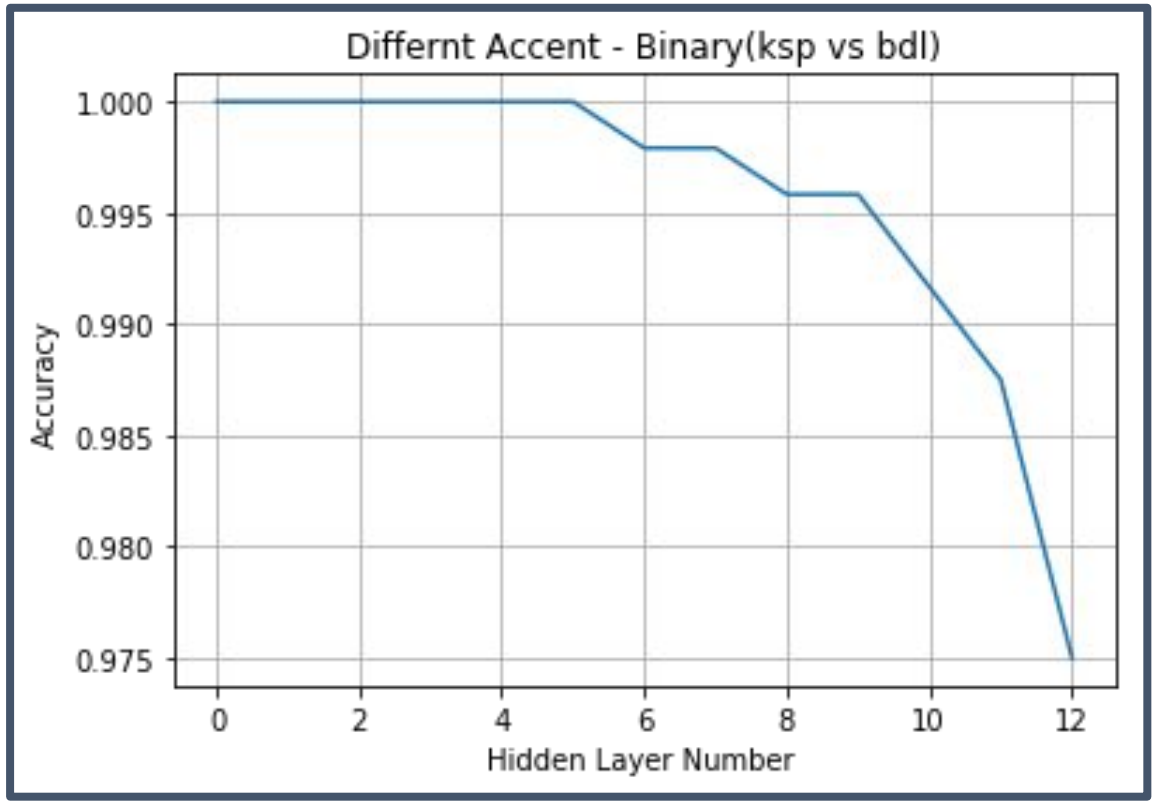




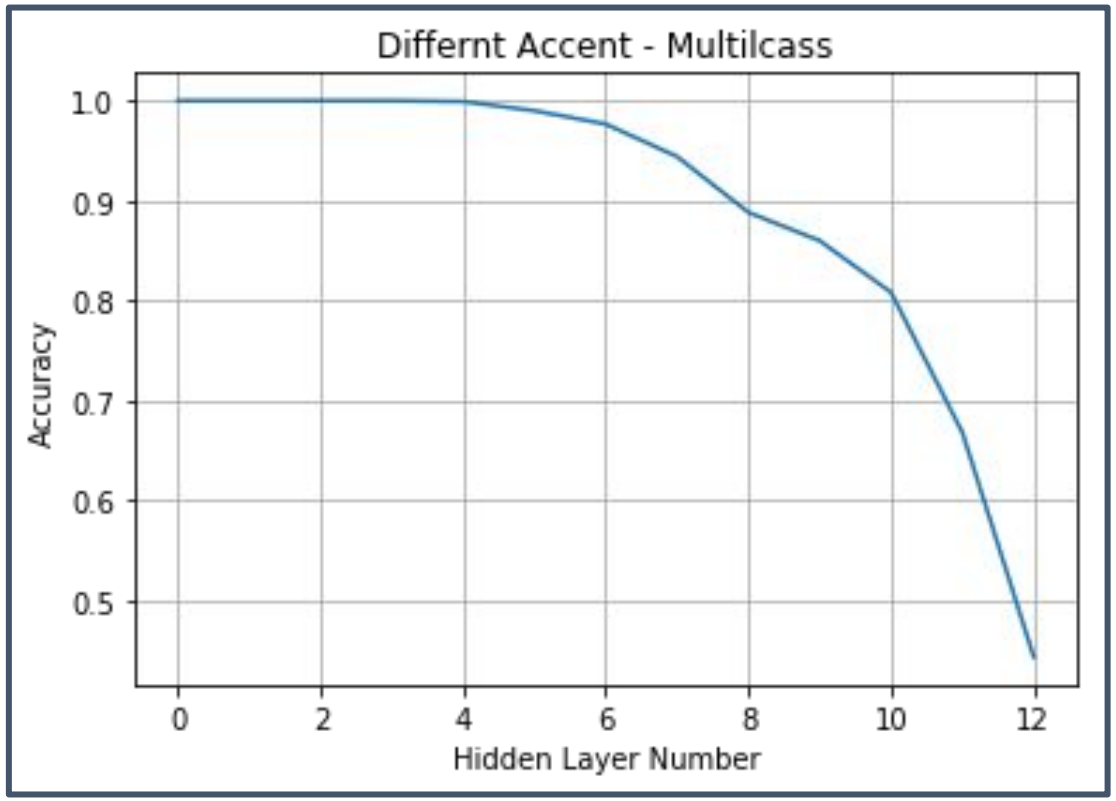
North midland American
(bdl-slt)

(JMK-stl)
North midland American - Ontario Canadian

Results - **Accent** Classification based on attentions in different layers



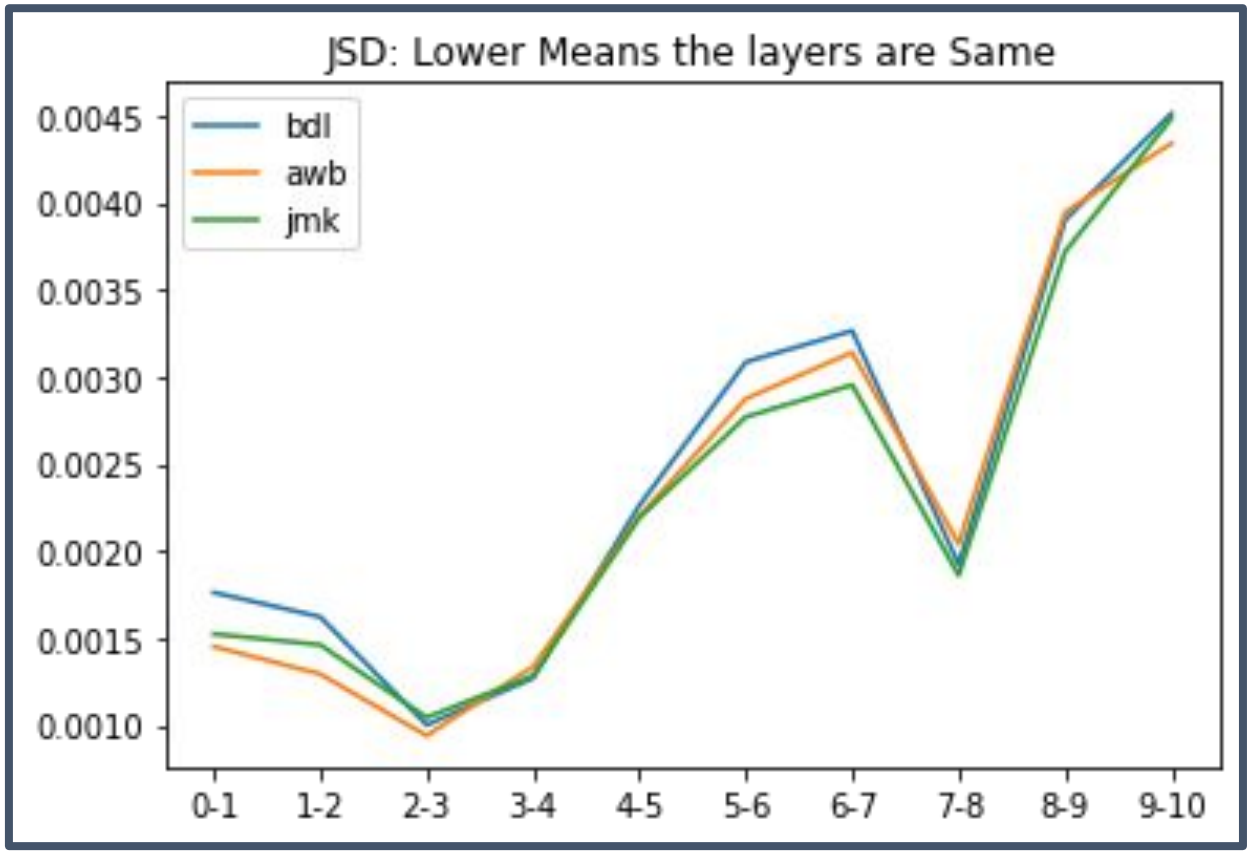
Indian Vs American Accent

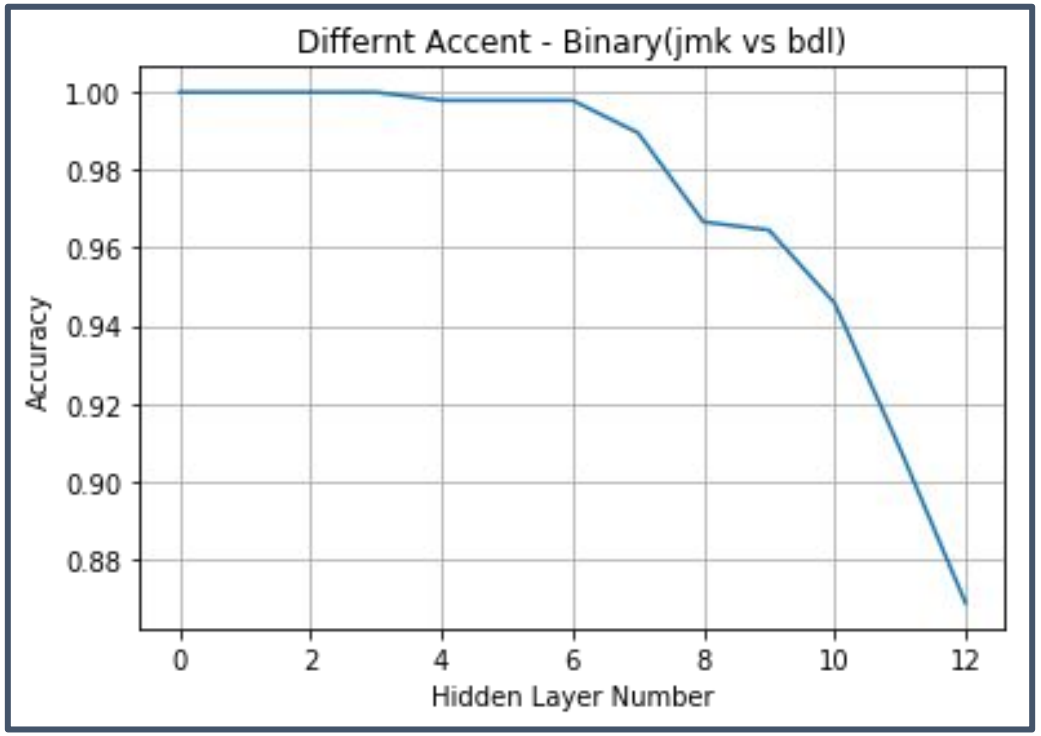


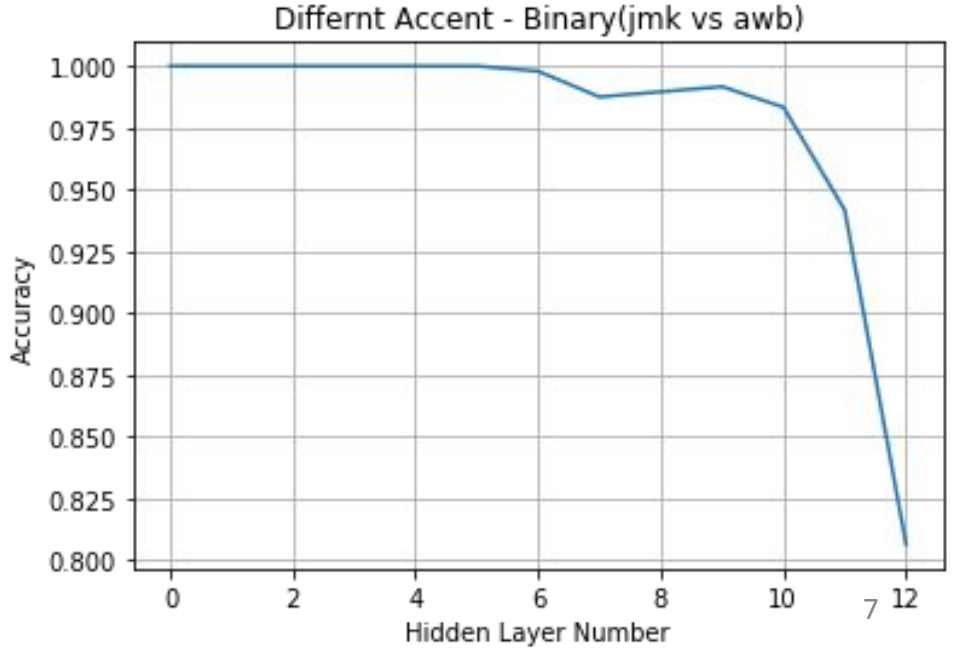
12 Various accents

Turning to JS divergence

BDL and AWB: 0.004
JMK and BDL : 0.006 (canadian and Midwestern)
AWB and JMK:0.003



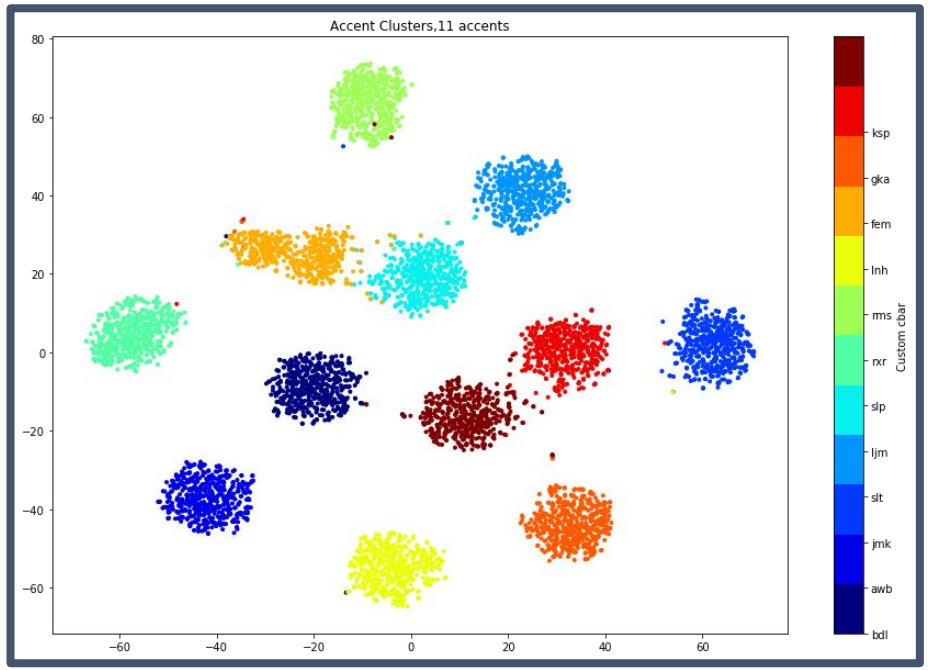


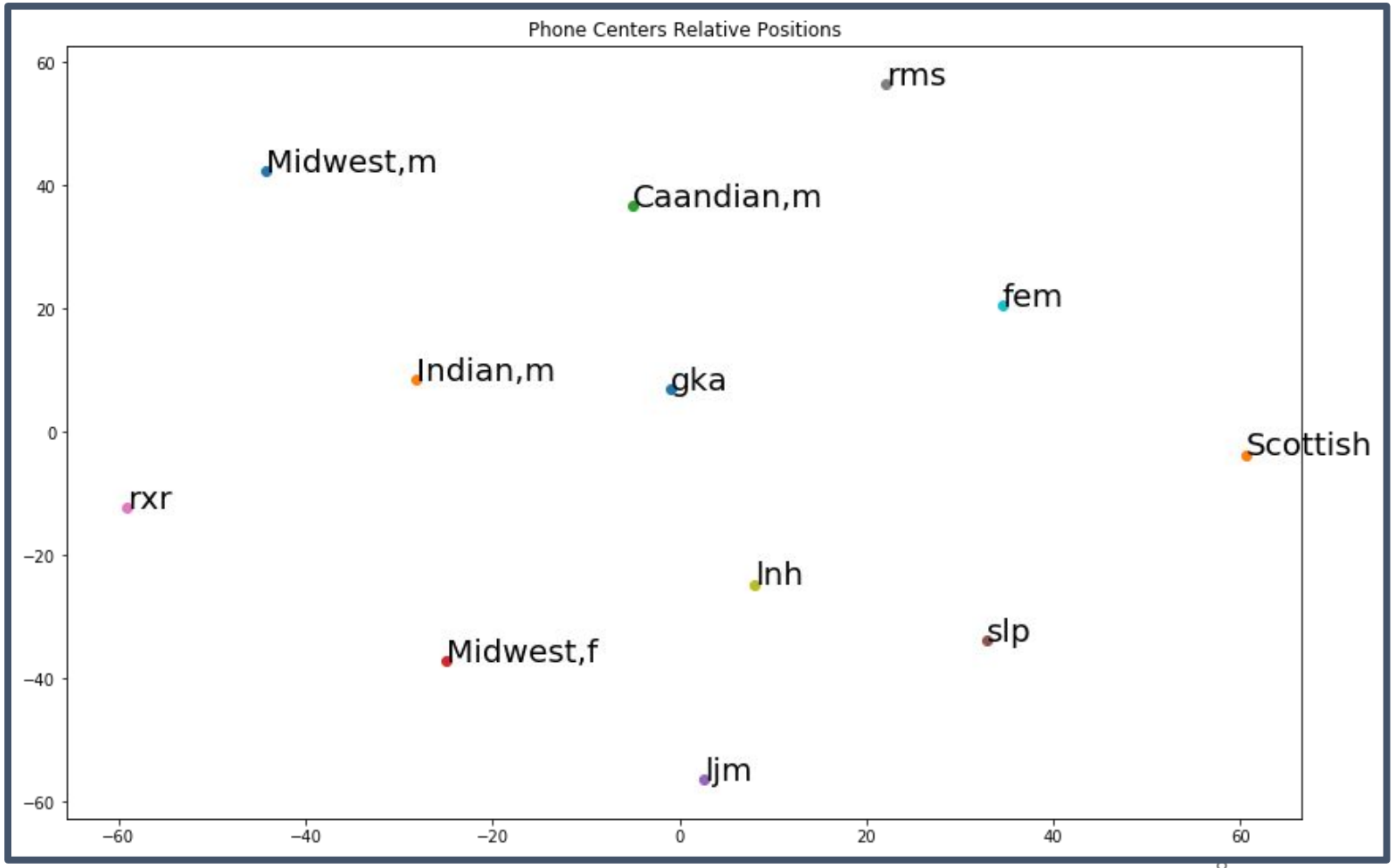


"The accent of jmk represents a minor deviation from bdl(american), while that of awb is strikingly different"

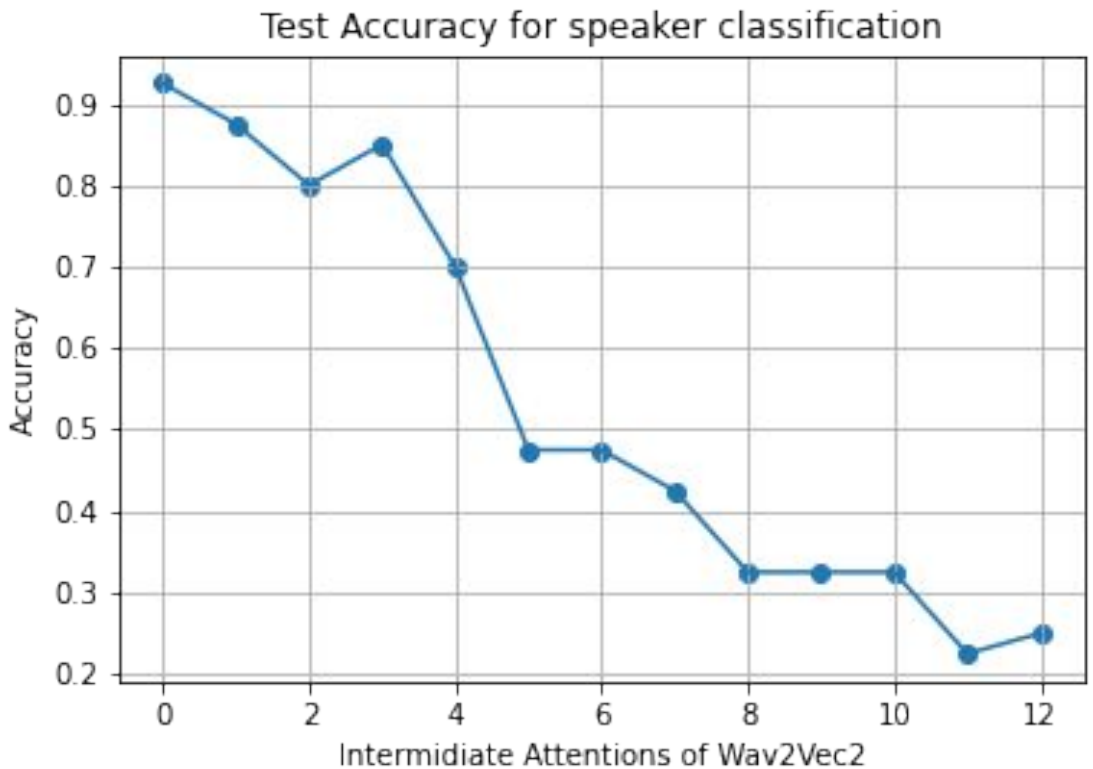
t-SNE for accents

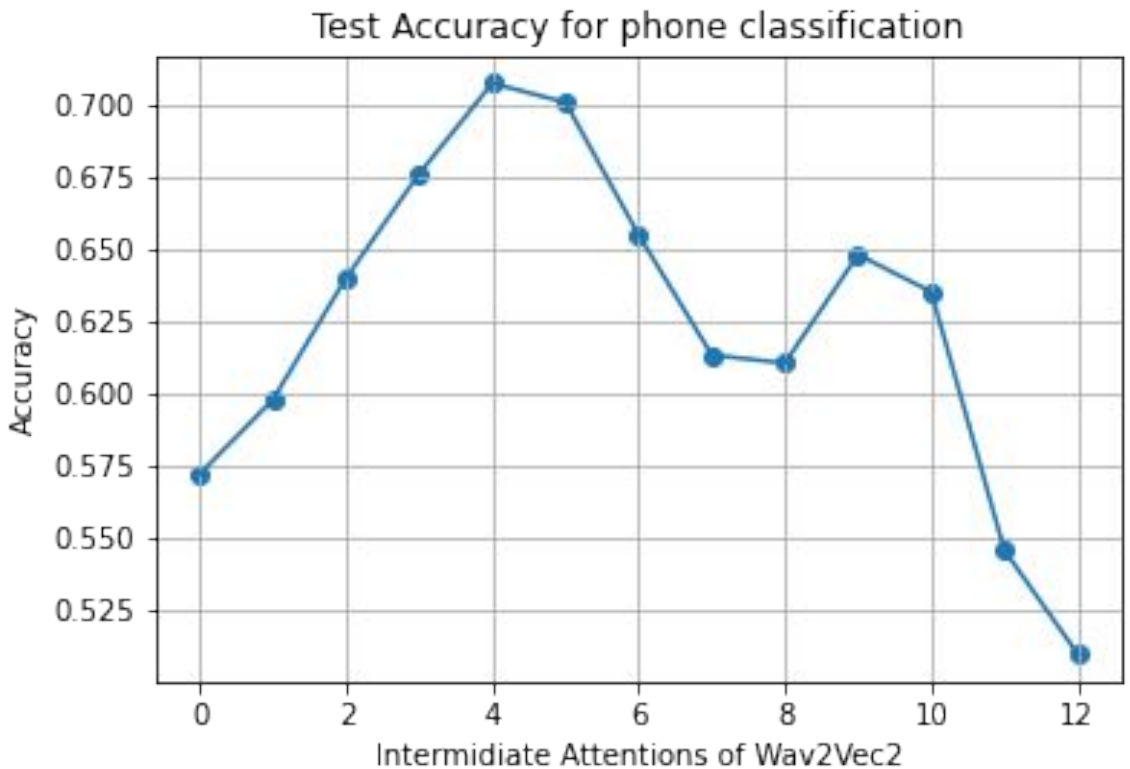
2nd Attention Layer Repr.





Results - Speaker & Phone Classification based on attentions in different layers



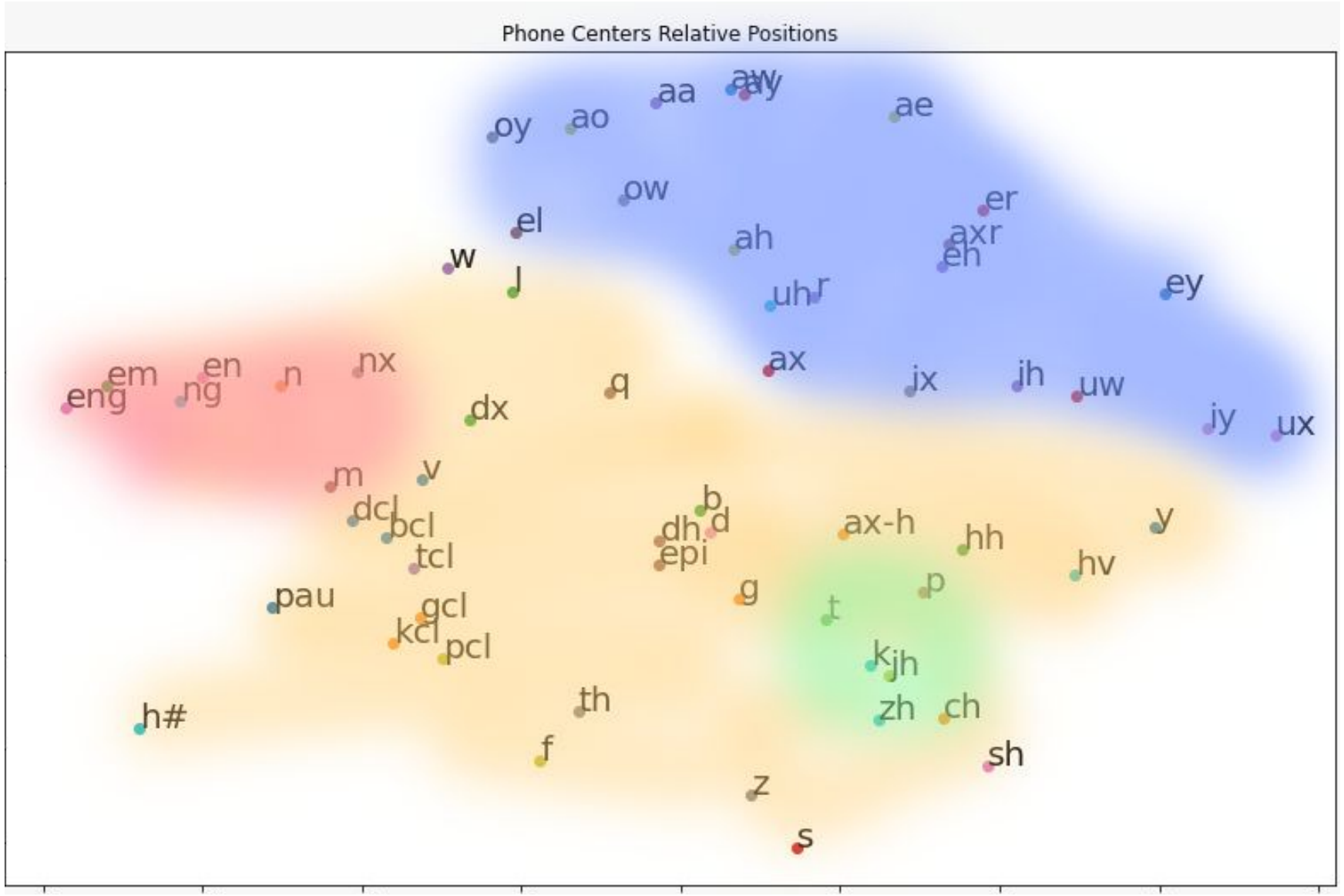


Adjacency in phones Repr.(cluster centers):

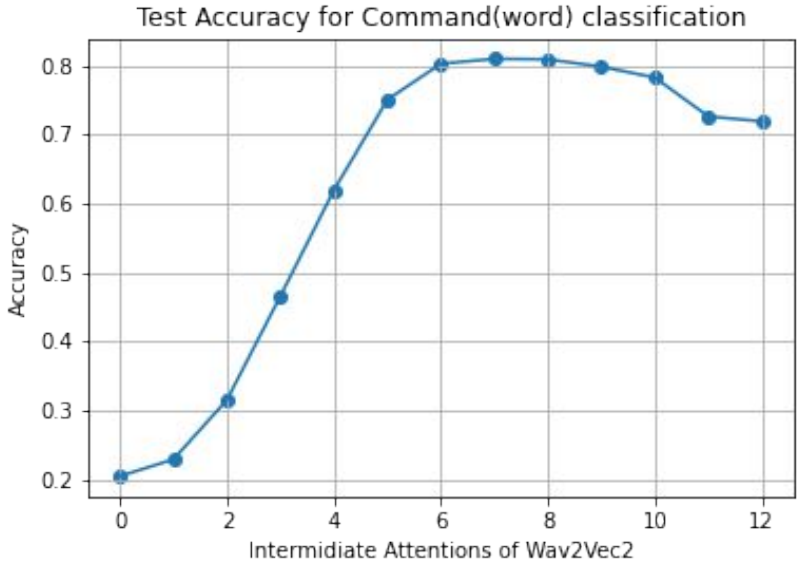
#phones = 61
(TIMIT)

Red : Nasal
Green : Voiceless
Blue : Vowel
Yellow : Consonant

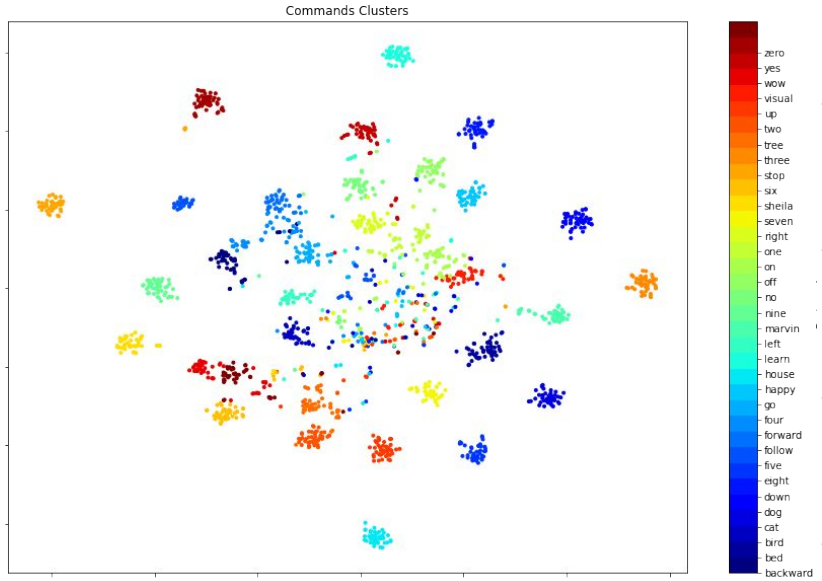
4th Attn. Layer

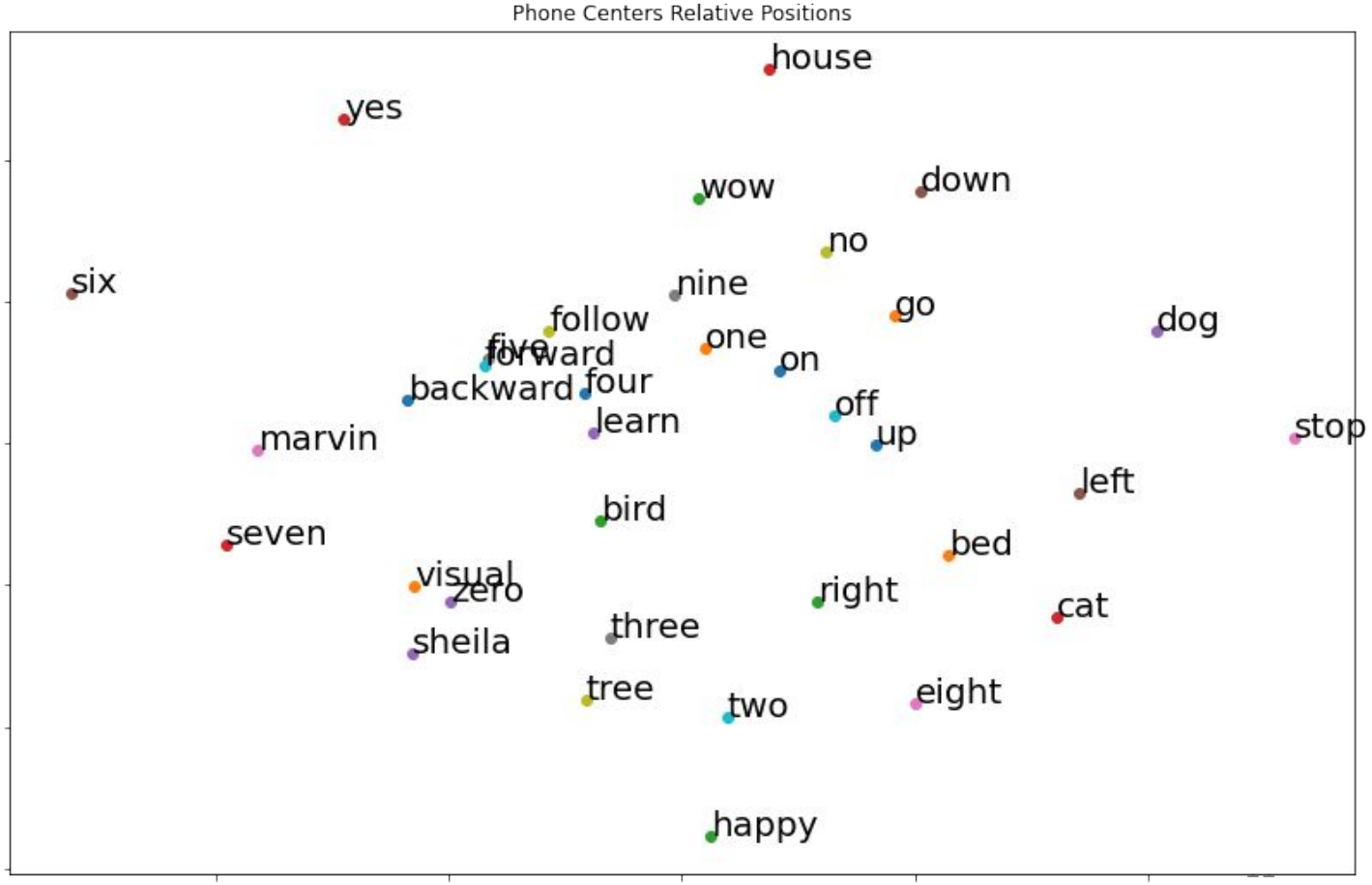


Results - Speech Command Classification based on attentions in different layers



Word representations are learned in layer 7





Conclusion and Future Work

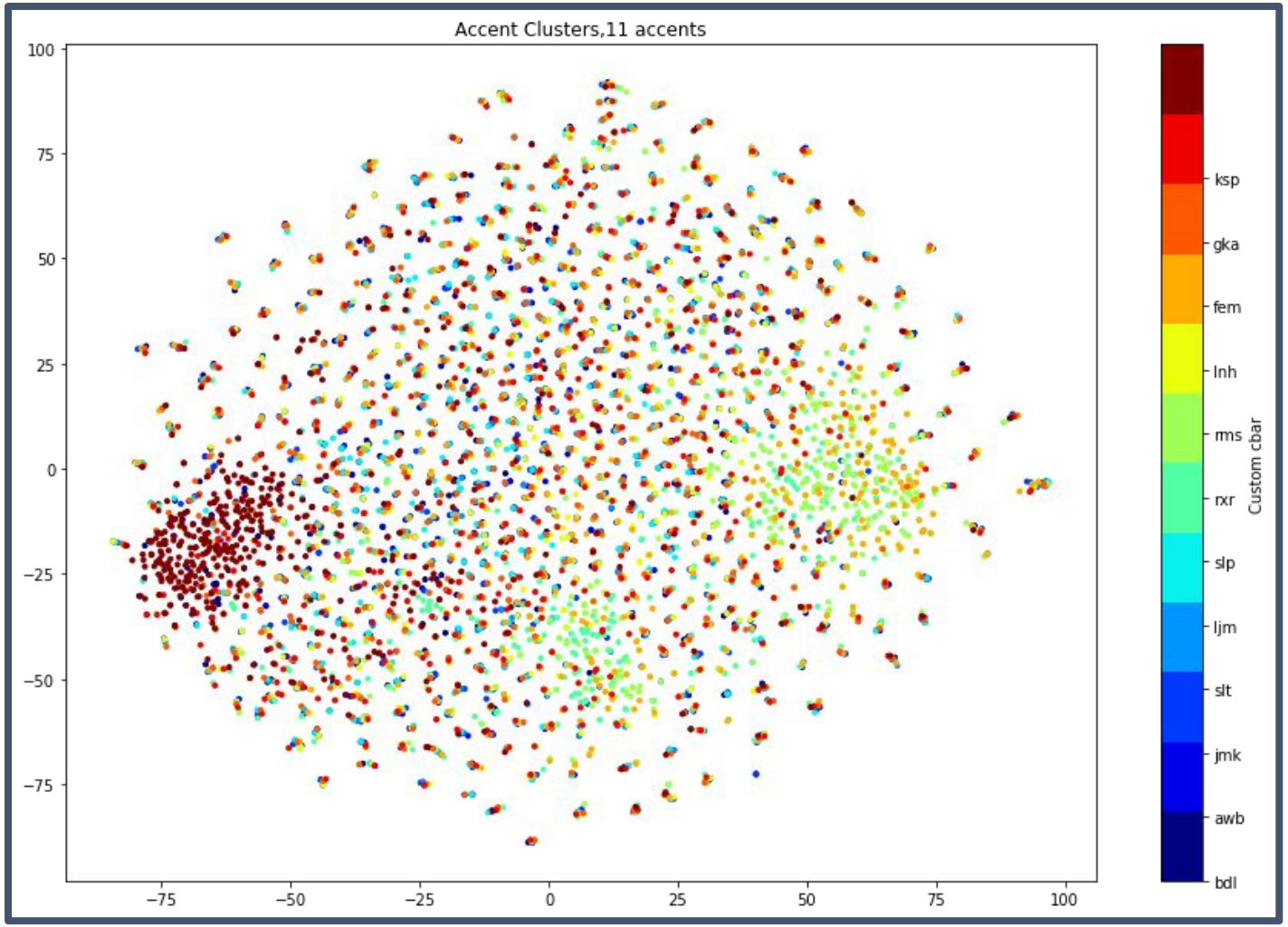
- Conducting these experiments on other representation learning models like **mockingjay** and **deepspeech**.
- Discovering in which layers other more complex speech attributes like intent, noise detection are learned
- Evaluating more metrics on the representations .

References\Literature Survey:

- Wav2vec 2.0 - https://arxiv.org/abs/2006.11477
- CMU ARCTIC Databases for Speech : http://www.festvox.org/cmu_arctic/
- TIMIT Acoustic-phonetic Continuous Speech Corpus
- Speech Commands dataset: https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html
- Adaptively Sparse Transformers,Correia,et al.: JS divergence
- Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals,Becker et. al.: The main motivation paper
- How Accents Confound: Probing for Accent Information in End-to-End Speech Recognition Systems

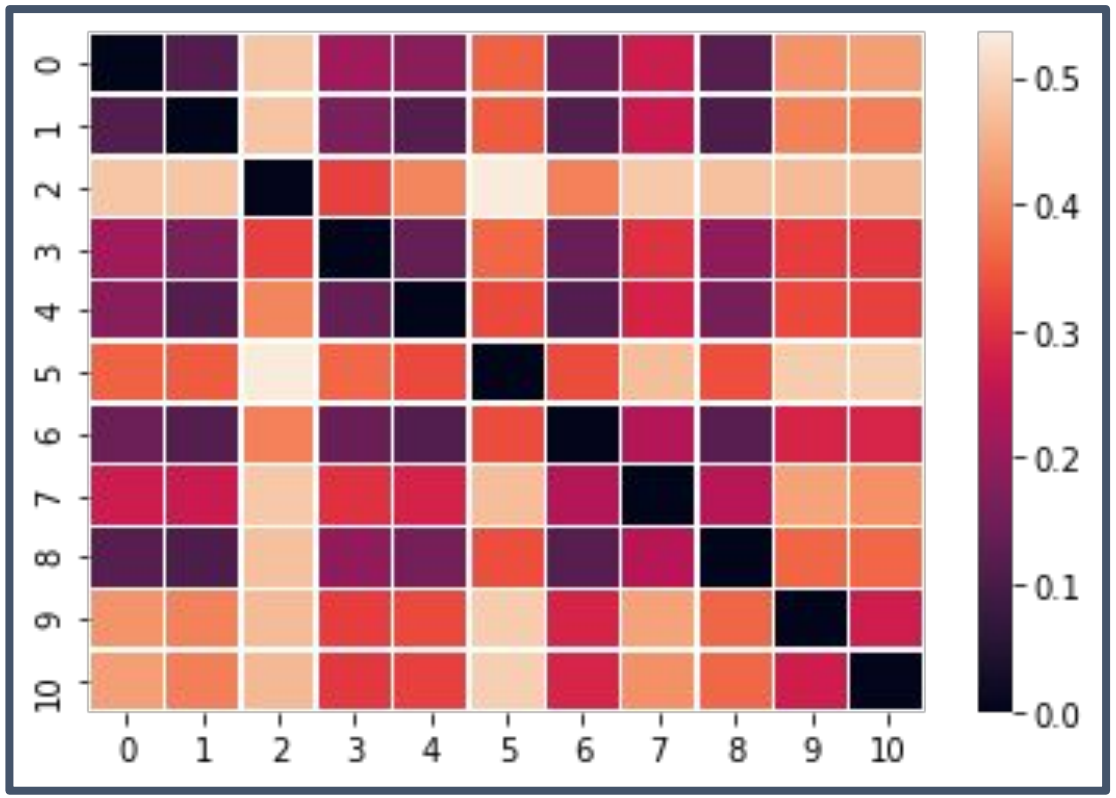
Thank You

Results - Determining Accent based on representations - T-SNE for higher hidden layers

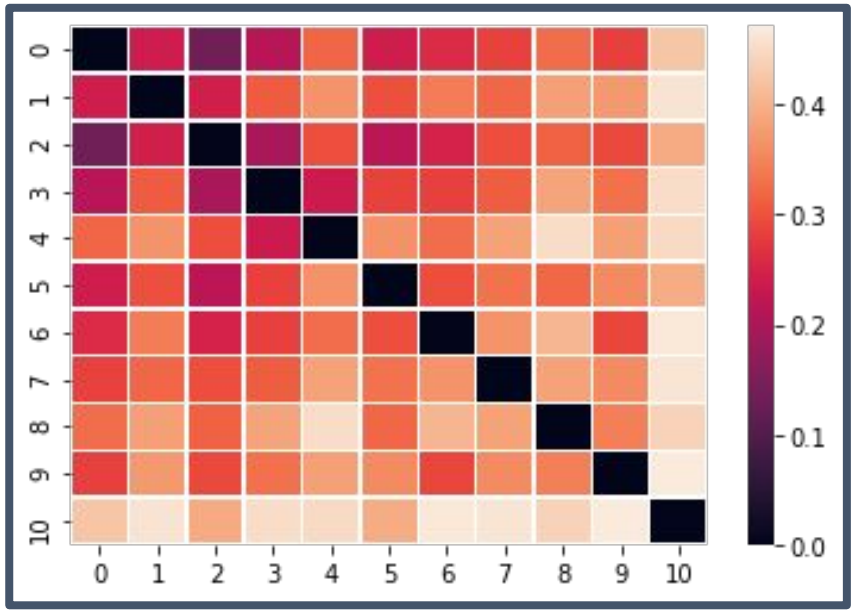


Supplementary JS - patterns

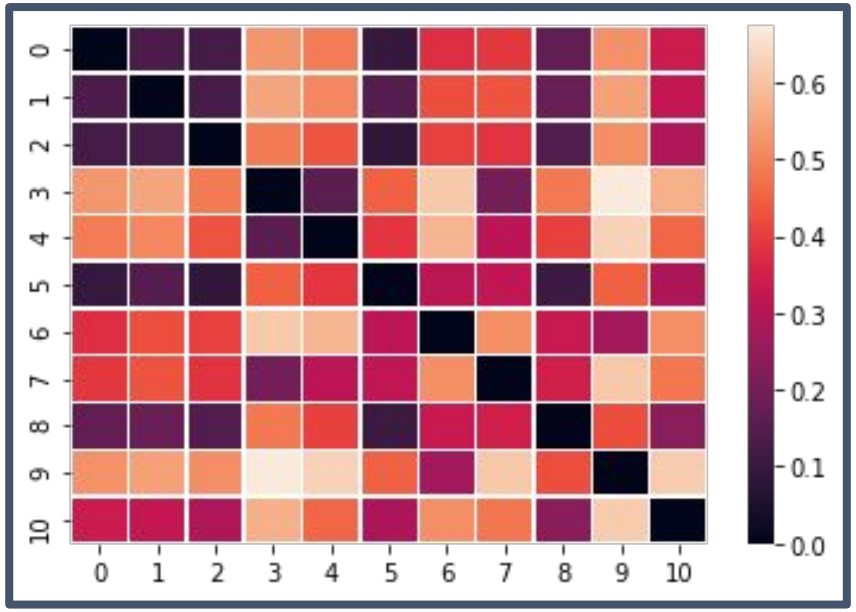
JMK



AWB-->



BDL-->



Phone Clustering using PCA & t-SNE

- Attn. Layer : 4
- PCA : d=50
- t-SNE : d=2