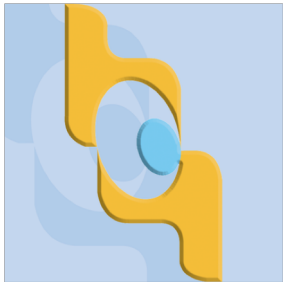
Towards Automatic Speech Identification from Vocal Tract Shape Dynamics in Real-time MRI

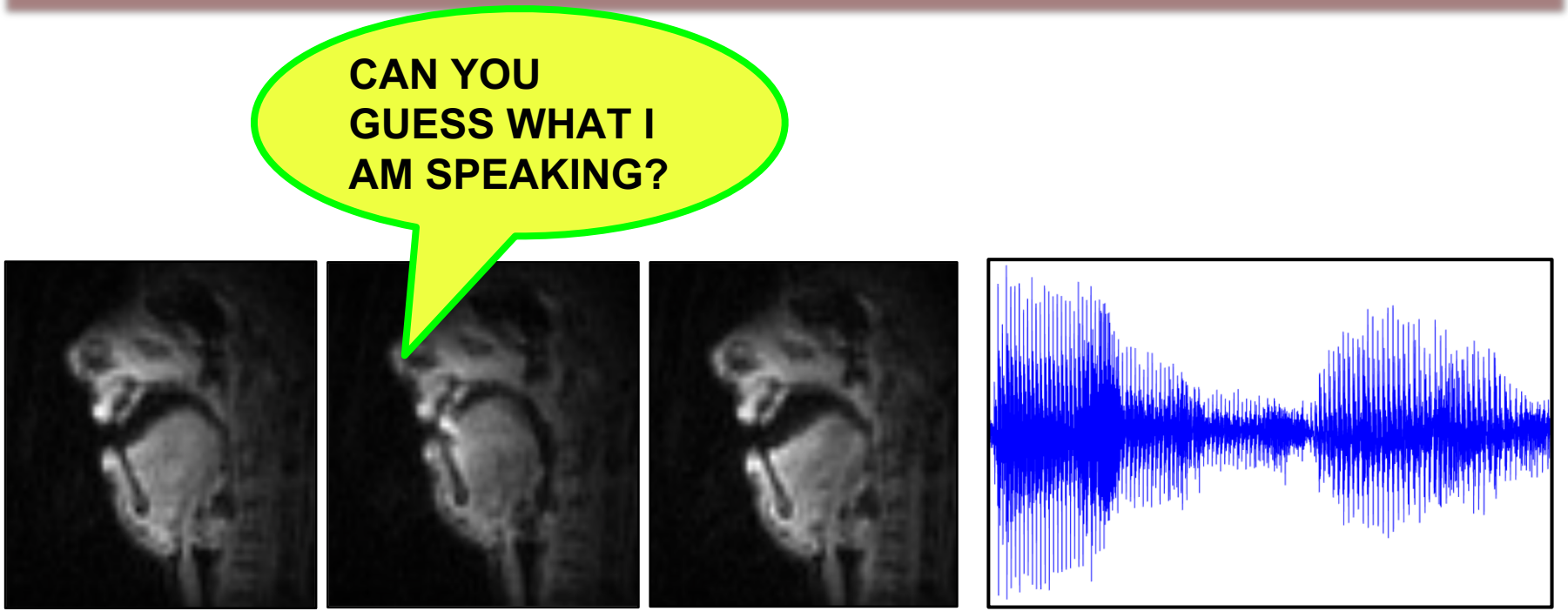
Pramit Saha*, Praneeth Srungarapu*, Sidney Fels
University of British Columbia, Vancouver, Canada



human
communication
technologies

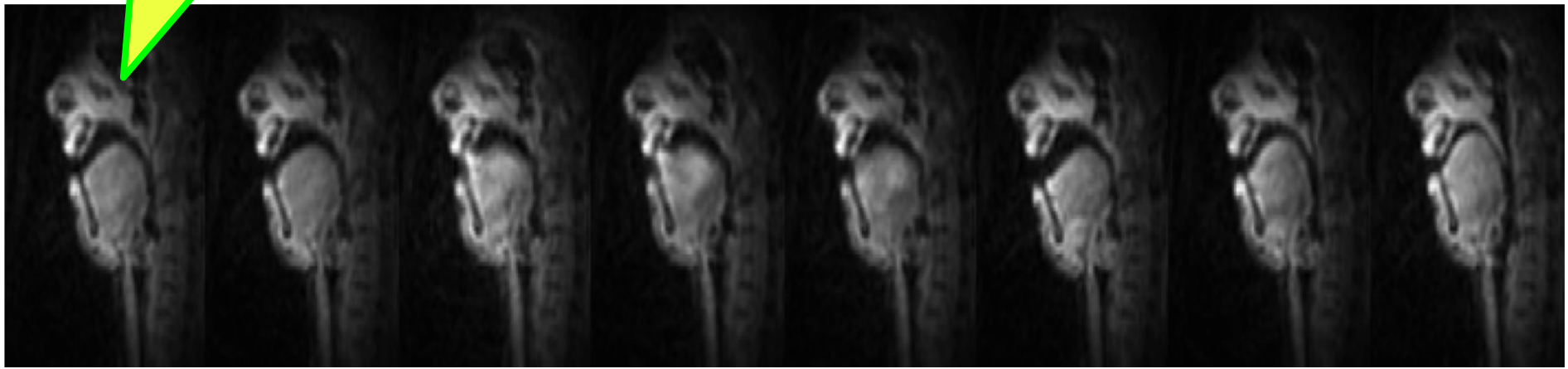
*** Indicates equal contribution**

Articulatory- Acoustic mapping from Dynamic Vocal Tract Imaging Data



Articulatory- Acoustic mapping from Dynamic Vocal Tract Imaging Data

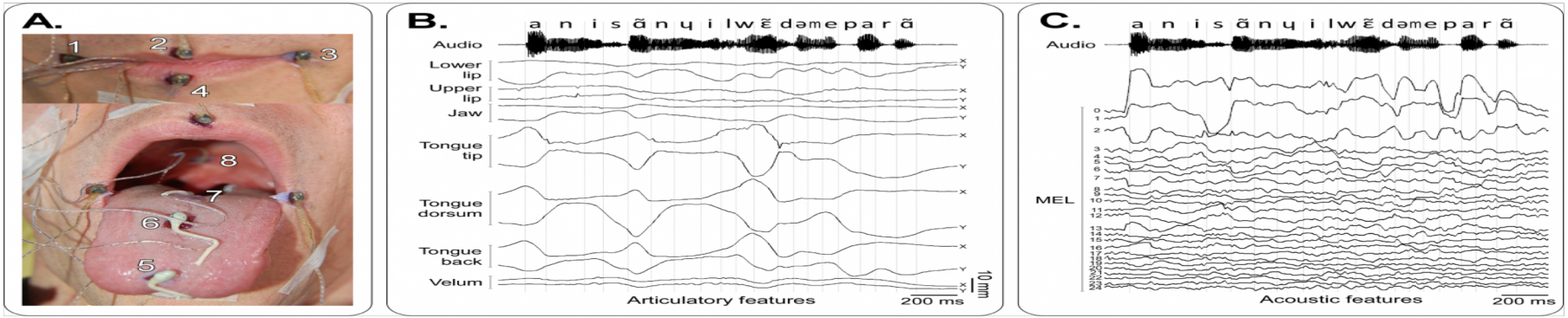
WHAT ABOUT THIS ONE?



Articulatory- Acoustic mapping from Dynamic Vocal Tract Imaging Data

- Extraction of the dynamic information of vocal tract geometry is crucial for identifying and synthesizing speech.

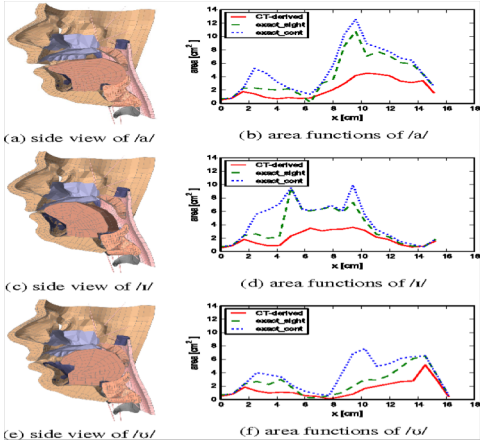
- Established methods rely solely on the articulatory **time-series positional data** from a few specified points on tongue, lip and jaw (like EMA).



Articulatory- Acoustic mapping from Dynamic Vocal Tract Imaging Data

- **Dynamic vocal tract-imaging data** can effectively record the labial, lingual and jaw motion along with the articulation of the velum, pharynx and larynx, etc.

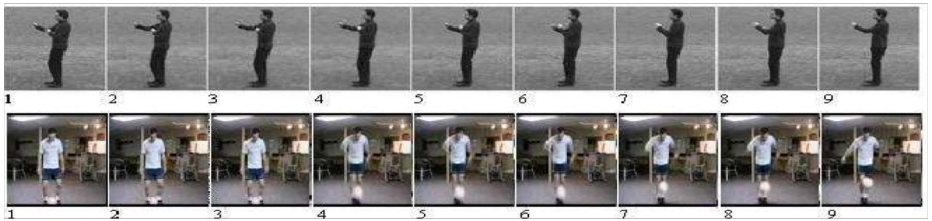
- Extremely challenging and computationally expensive to achieve a direct end-to-end mapping of the **entire vocal tract geometry utilizing dynamic information** acquired by real time imaging modalities to the corresponding speech sound output.

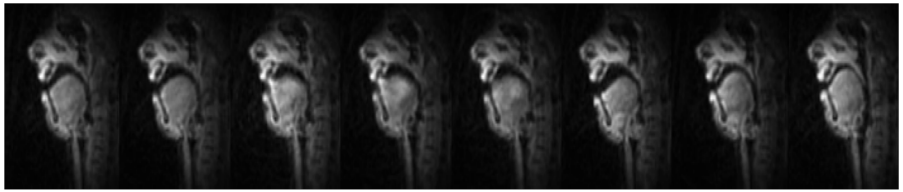


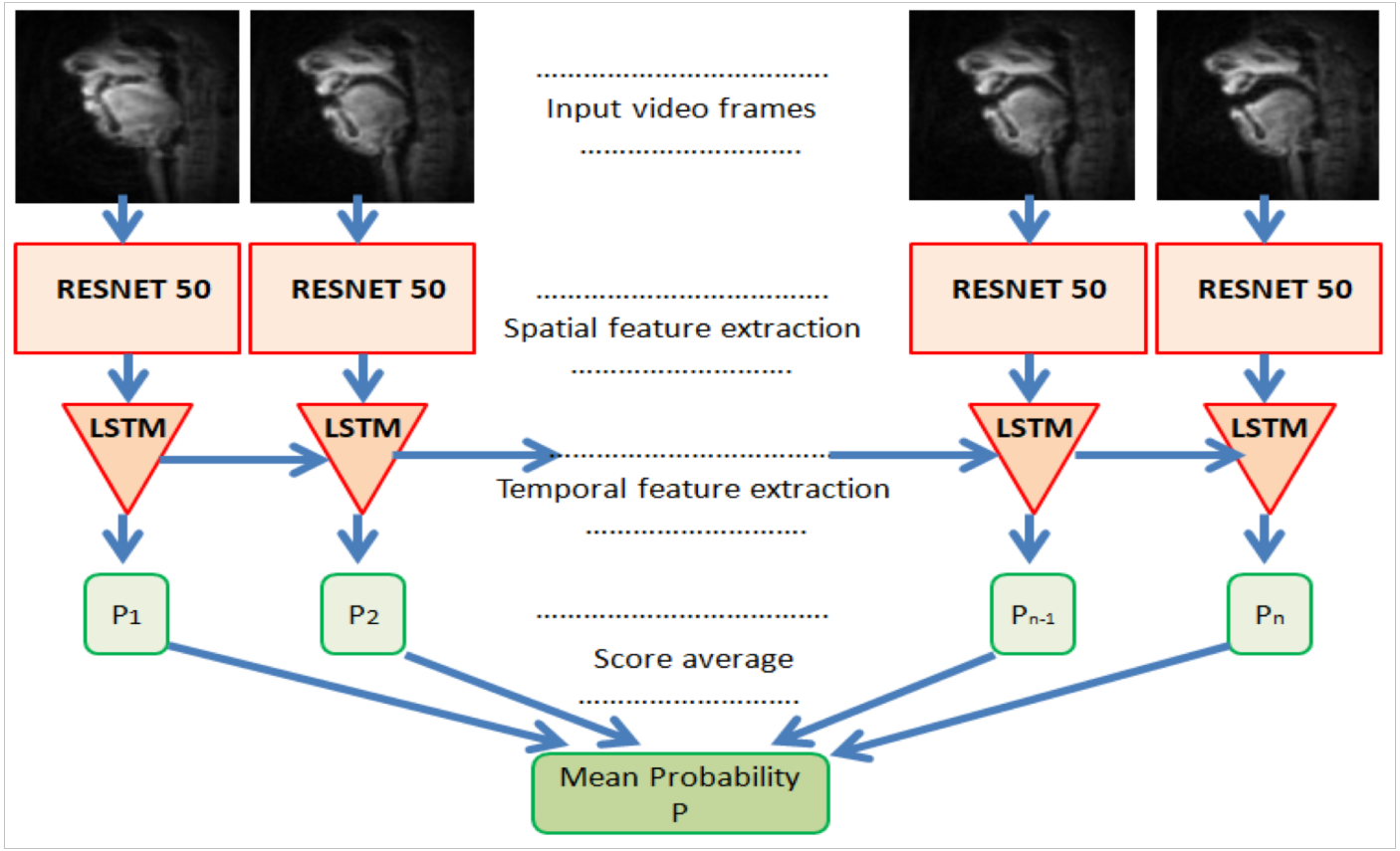
Vowel-Consonant-Vowel (VCV) Identification approach

- **Framing:** Speech token identification from the rtMRI ====> '**sequential input to fixed output**' problem with **videos of arbitrary length *T* as input** and **prediction of categories corresponding to each video, as the output**.

- **Research Question:** Can the speech recognition problem be viewed similar to a video action classification/recognition task? (Where the vocal tract movement resembles the action and the VCV sequences as the final mapped output interpreted from the action).







Approach: **Long term Recurrent Convolutional Neural Networks (LRCN)** to extract the per-frame spatial features and simultaneous inter-frame temporal features for speech identification tasks from rtMRI videos.

Speech tokens in Vocal Tract Morphology rt-MRI Dataset

- **USC Speech and Vocal Tract Morphology MRI Database** : includes 2D real-time MRI of vocal tract shaping of 17 speakers (9 female and 8 male) along with simultaneous denoised audio recording.

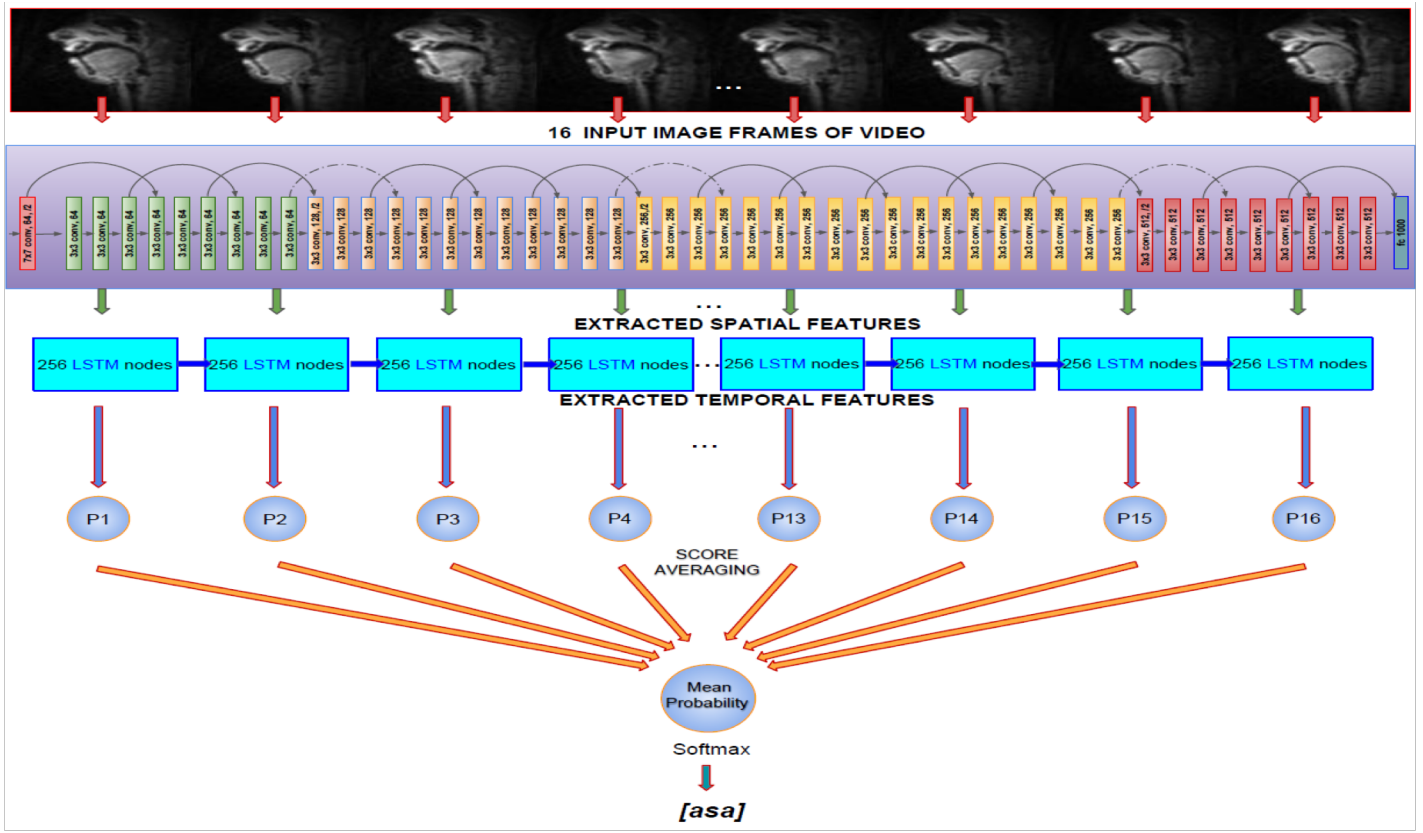
- The database contains 3 repetitions corresponding to each speaker for each of 51 VCV utterances:
  apa, upu, ipi, ata, utu, iti, atha, uthu, ithi, aka, uku, iki, aba, ubu, ibi, ada, udu, idi, aga, ugu, igi, asa, usu, isi, aʃa, uʃu, iʃi, ama, umu, imi, ana, unu, ini, ala, ulu, ili, afa, ufu, ifi, aɹa, uɹu, iɹi, aha, uhu, ihi, awa, uwu, iwi, aja, uju, iji.

- We pre-segment these into a total of **2754 videos**, each containing the entire length of single VCV utterance and then into training and testing dataset of 2268 and 486 videos for cross-subject evaluation. 16 image frames are extracted with stride of 3 excluding silent frames.

Long Term Recurrent Convolutional Neural Network Model (LRCNN) for Speech Token Identification

- We use deep (50 layered) residual learning framework known as ResNet, composed of series of Identity and Convolutional blocks.

- After ResNet based spatial feature extraction phase, it is cascaded through two fully connected layers (with 2048 and 1024 neurons) and then through RNN composed of LSTM units, capable of learning complex temporal dynamics over long sequences. Each LSTM cell is composed of memory cell, input, output, forget and input modulation gate.



Batch size: 64, **Number of steps/epoch:** 500, **Number of LSTM layer:** 1, **Number of LSTM nodes:** 256, **Total number of epochs:** 140, **Drop-out for Fully Connected Layers:** 0.5, **Drop out for LSTM network:** 0.9. **Learning rate:** .001, **Optimization algorithm:** Adam, **Loss function:** Categorical Cross-Entropy, **Activation function:** Softmax.

Speech Identification Performance

Evaluations were performed for various combinations of parameters and classes (V, C, VCV)

Table I: Top-1 categorical accuracies

Identification Task	Top-1 accuracy	Kappa value
Vowel	0.96	0.63
Consonant	0.68	0.62
VCV	0.42	0.40

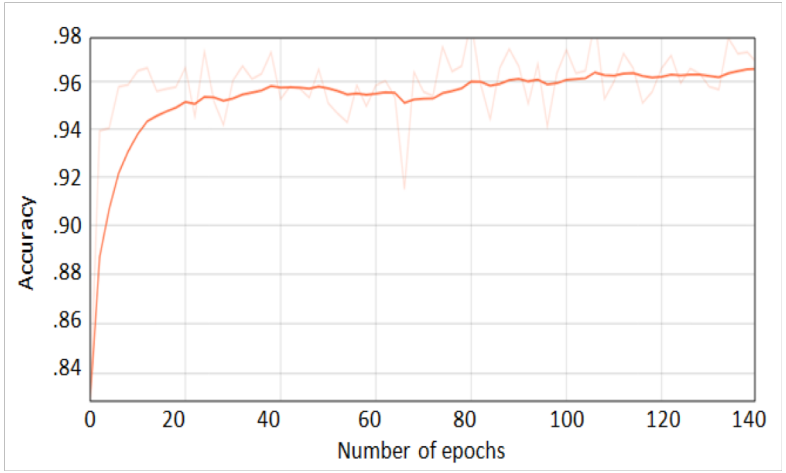
Above-chance accuracy or Cohen's Kappa value (k) for vowels and consonants are similar. Here, the '**chance**' for vowels is defined as **.33**, for consonants **.06** and for VCV, **.02** .

$$k = \frac{p_o - p_e}{1 - p_e}$$

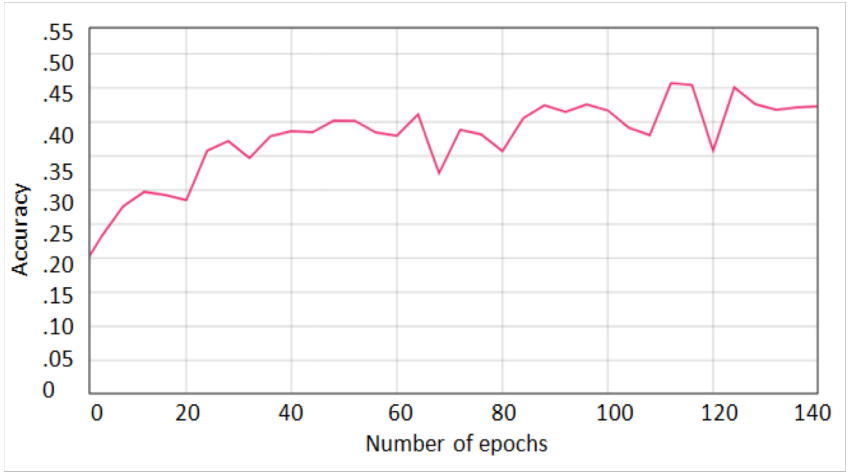
Top-1, Top-5 and Top-10 accuracies for VCV classification reach **.42, .76** and **.93** respectively.

Speech Identification Performance

- Vowel or open vocal tract sound classification task achieved maximum accuracy (**.96**), Classification of closed vowel sounds achieved accuracy of **.68**.



Vowel Classification Accuracy



VCV

Classification Accuracy

Conclusion and Future Directions

- A promising deep learning based algorithm for action classification in videos has been trained and tested on rtMRI database with 51 VCV speech utterances.
- Cross-subject classification performance was satisfactorily approximated for vowels and acceptably for consonants, but showed decrease in accuracy for VCV [transitional tokens]
- While extracting features from imaging modalities, the inter-participant anatomical differences restrict a particular model from generalizing the structural features like the shapes and lengths of the articulators
- In future, to avoid current roadblocks, we plan to couple the current model with an underlying biomechanical model and check whether augmentation from vocal tract model can assist the algorithm

Thank you

Questions???