Ultrasound Tongue Image Sequence Classification Using Deep Learning

Presented by: Khoa Tran and Shane Steinberg

Speakers

Khoa Tran

Shane Steinberg

Motivation: Silent Speech Interface

- Rehabilitation/accessibility devices, e.g., Laryngectomy patients.
- Human Machine Interface via speech-to-text control.
- Covert communication for military applications.

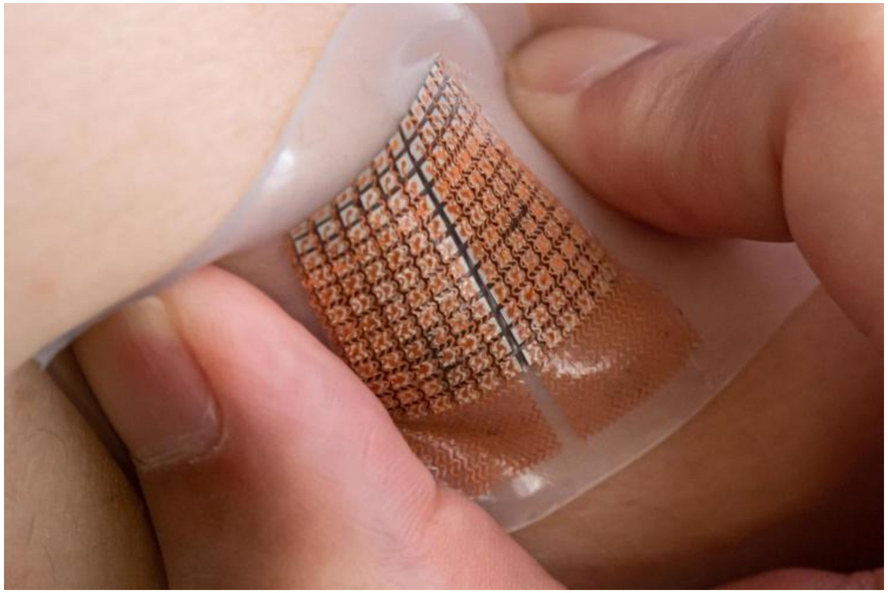
Example silent speech device proposed in [1].

Ultrasound Imaging for Articulatory Muscle Sensing

- Ultrasound imaging is non-invasive modality that can capture structural and dynamical information of the internal tissues.

- Potential for miniaturization and wearable conformal devices.

- Newer transducer technologies can be integrated on chip (CMUT, PMUT).



Wearable conformal ultrasound imaging device proposed in [2].

Objective

- Investigate the feasibility of silent speech interface using ultrasound image sequences of the tongue and palate.

- Word classification from limited vocabulary selected for preliminary study.

- Explore data pre-processing steps to improve the learning efficiency of the classification model.

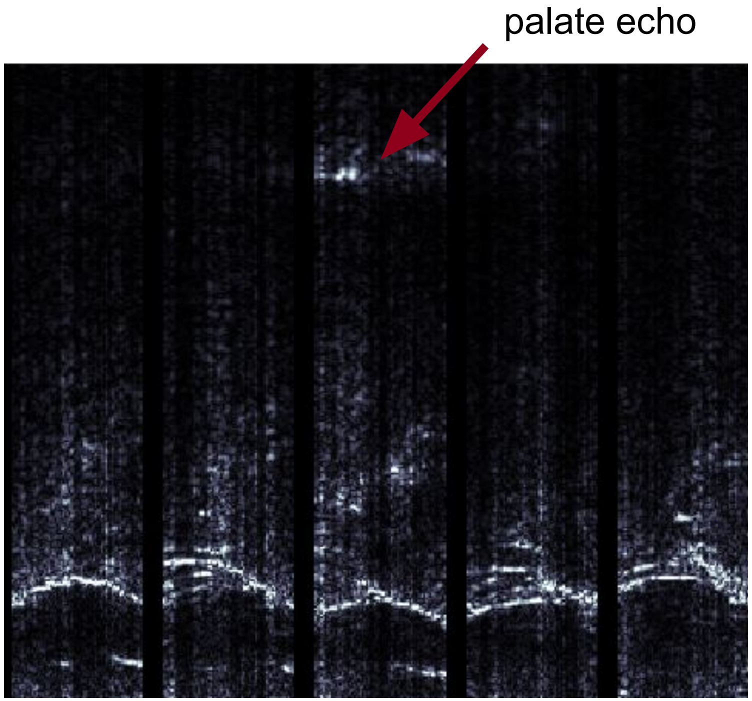
Experimental Methodology

- Ultrasound image sequences acquired from tongue and palate during word utterance.
- Four words from Spelling Alphabet selected for classification:
  'Alpha', 'Bravo', 'Charlie', 'Delta'.
- 50 samples of each utterance obtained in 3 second image sequence recordings.
- One human subject.

Experimental configuration of ultrasound image acquisition.

Data Acquisition Parameters

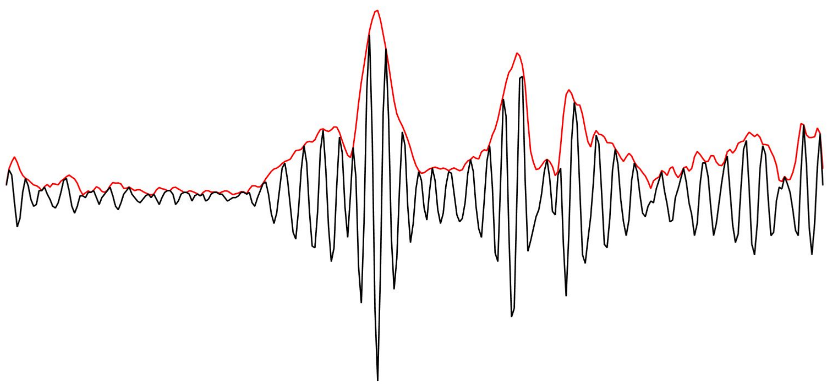
- Verasonics Vantage 64 LE Research Platform with linear probe.
- Plane-wave transmit and receive.
- 128 transmit elements, 64 receive.
- 7 MHz ultrasound center frequency.
- 100 Hz frame rate.



Ultrasound image sequence montage of articulatory tissues during utterance of 'alpha'.

Ultrasound Image Reconstruction

- Ultrasound radiofrequency signals demodulated using absolute value of the Hilbert transform.
- Gain compensation applied along depth of signal to counteract wave attenuation in pixel intensity.
- Normalized to range [-1,1].

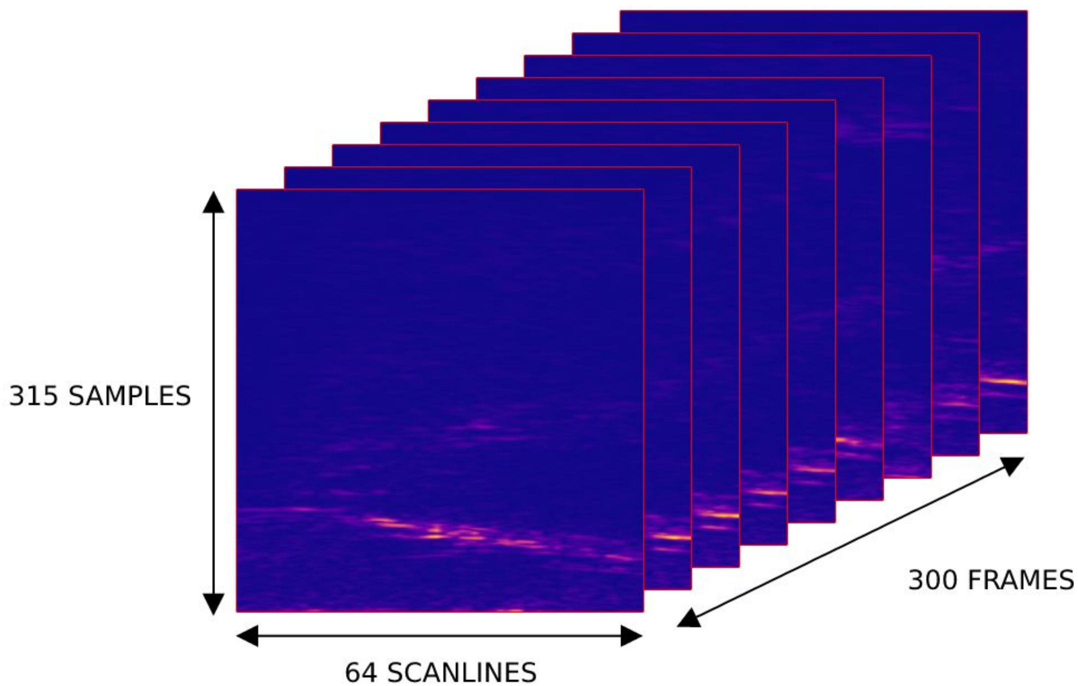


(Black) ultrasound radiofrequency signal, (red) envelope signal obtained via Hilbert transform.

Acquired Dataset

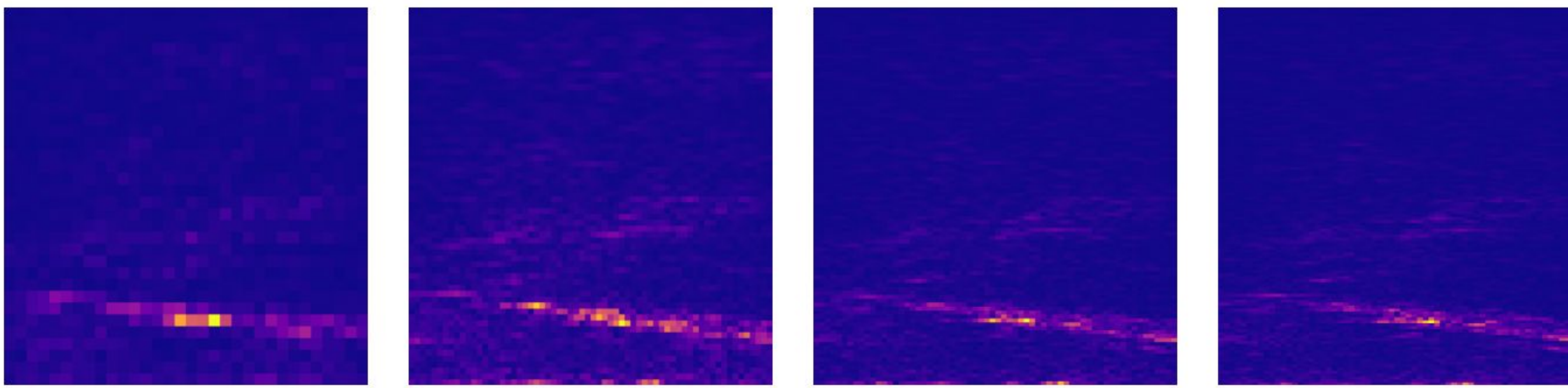
- Dataset consists of 4 classes with 50 samples each.
- Each sample is a (315, 64, 300) UTI sequence.

Class	Number of Samples
Alpha	50
Bravo	50
Charlie	50
Delta	50



Data Preprocessing - Resizing

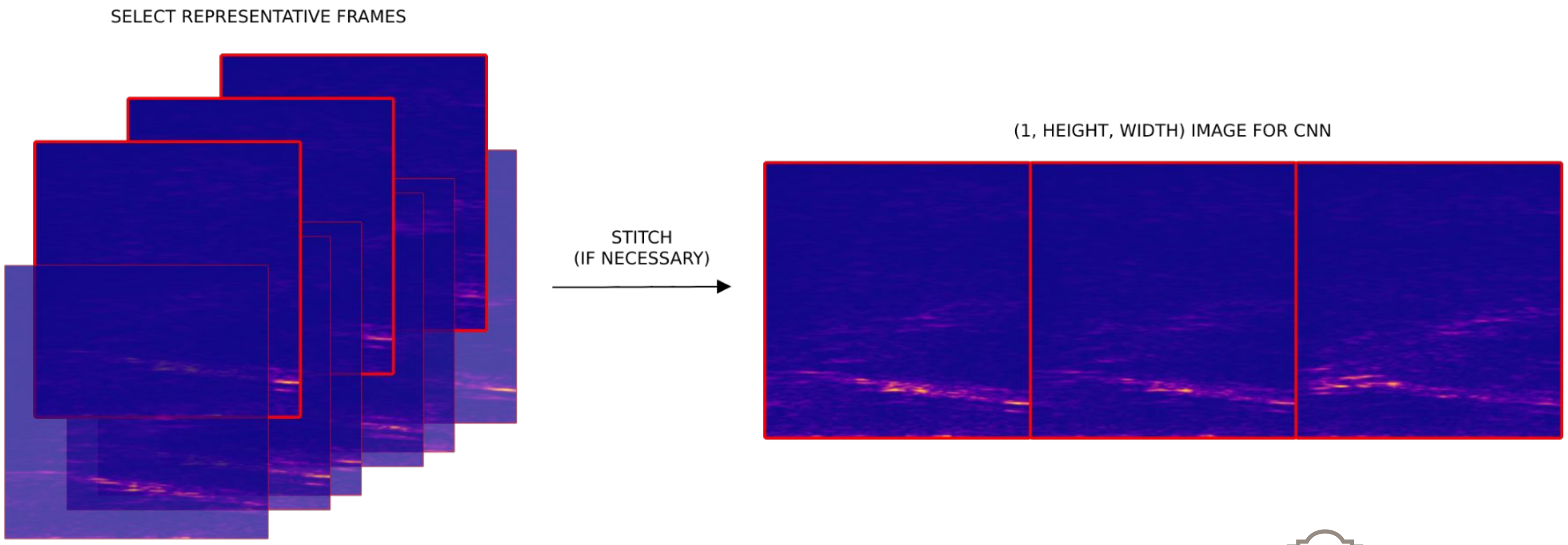
- In literature, images have been downsampled to 96x64 [3], 128x128 [4], 128x64 [6].
- We explore downsampling to the following input sizes using bilinear interpolation:



32x32 64x64 96x64 128x64

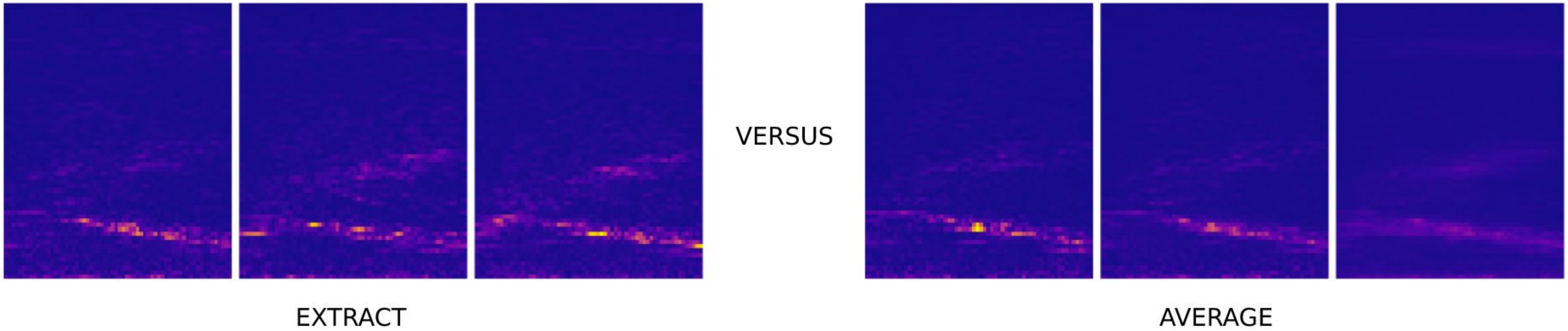
Data Preprocessing - Frame Selection

- Not all frames are necessary. Sometimes, only one single frame is enough [5, 6], or several frames are used [3, 4].
- We take [9, 16, 25, 36, 49, 64, 81, 100] evenly-spaced frames.
- Frames are stitched together for CNN-based models [3, 4].



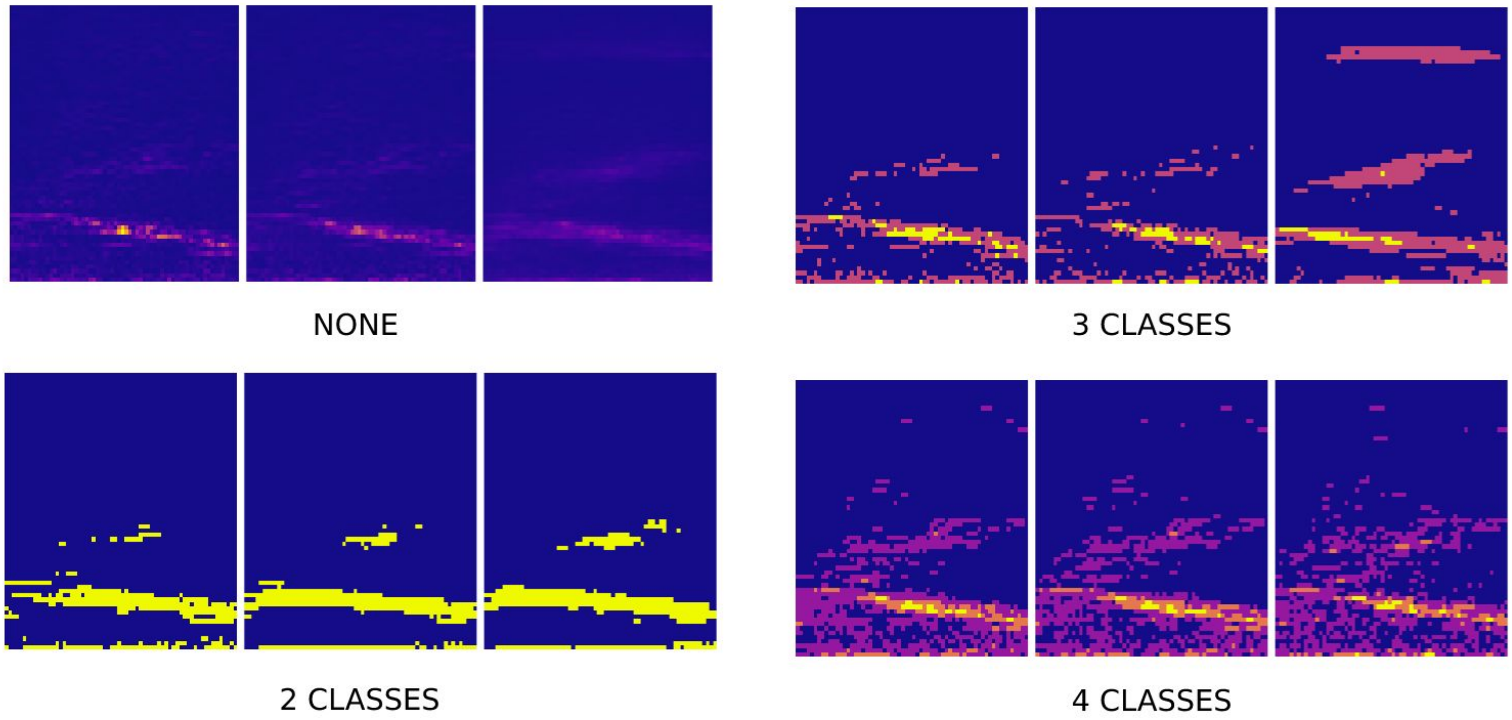
Data Preprocessing - Temporal Processing

- We explore selecting representative frames by **extraction** vs. **averaging** with a number-of-frames-dependent window size.
- Averaging may help reduce random noise.
- Training samples are **augmented** by randomly offsetting the evenly-spaced frame indexes and randomly adjusting the window size.



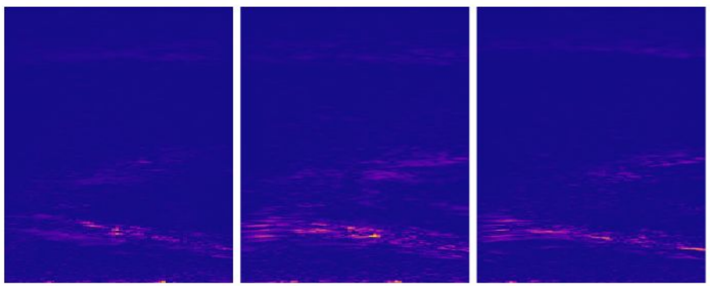
Data Preprocessing - Otsu Thresholding

- We explore **Otsu multi-threshold** intensity-based segmentation to reduce noise and simplify inputs.

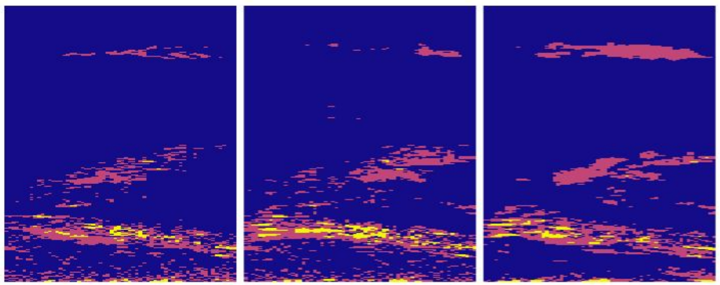


Data Preprocessing - Motion Maps

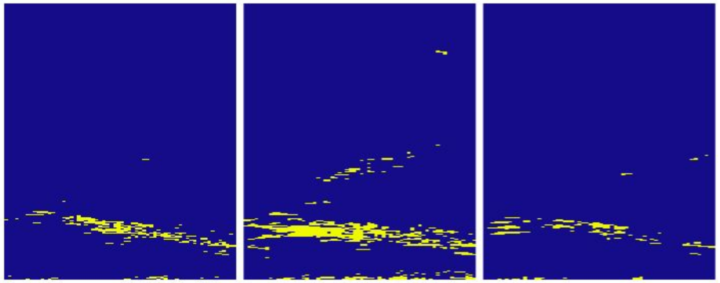
- We investigate **absolute temporal differentiation** using **multi-Otsu thresholding** to generate motion map inputs.



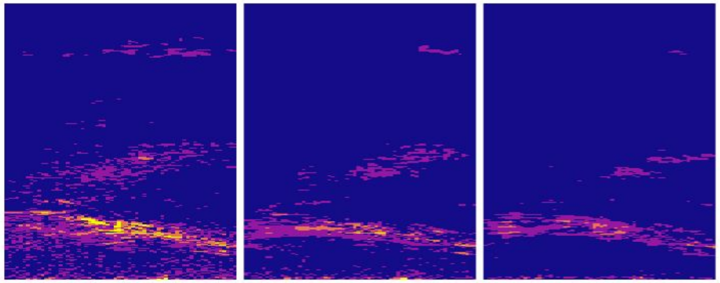
NONE



3 CLASSES



2 CLASSES



4 CLASSES

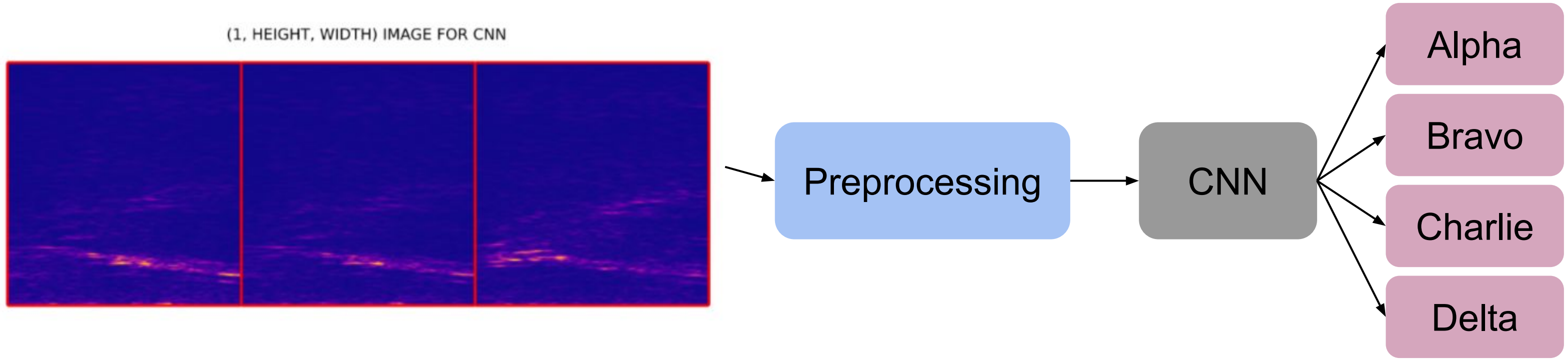
Training/Validation/Test Splits

- We split the dataset into 0.64, 0.16, 0.2 subsets for training, validation, and testing, respectively.
- Note that the training set is augmented (randomly selected frames and average-windows) to avoid overfitting.

Subset	**Percentage**	**Number of Samples**
Training (augmented during training)	0.64	128
Validation	0.16	32
Testing	0.2	40

CNN-Based Models

- Several papers apply CNNs to perform **articulation-to-acoustic conversion** using single frames [3, 5] or several frames [4].
- **Articulation-to-class conversion** has been done using a CNN with single frame inputs [5].
- We create and tune our own CNN with our dataset based on their works to perform **articulation-to-class conversion**.



CNN Model Implementation

Conv2D(in=1, out=16, kernel=3)
MaxPool2D(kernel=2, stride=2)
Leaky ReLU

↓

Conv2D(in=1, out=16, kernel=5)
MaxPool2D(kernel=2, stride=2)
Leaky ReLU

↓

Flatten

↓

Linear(in=VARIABLE, out=16)
Leaky ReLU

↓

Linear(in=16, out=16)
Leaky ReLU

↓

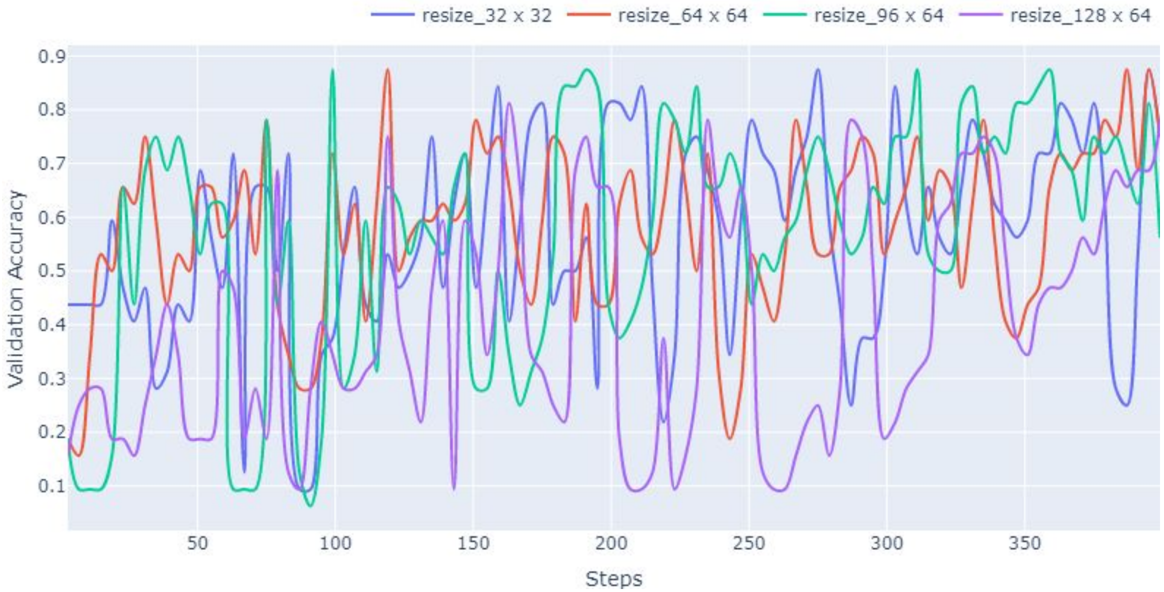
Linear(in=16, out=4)
Softmax

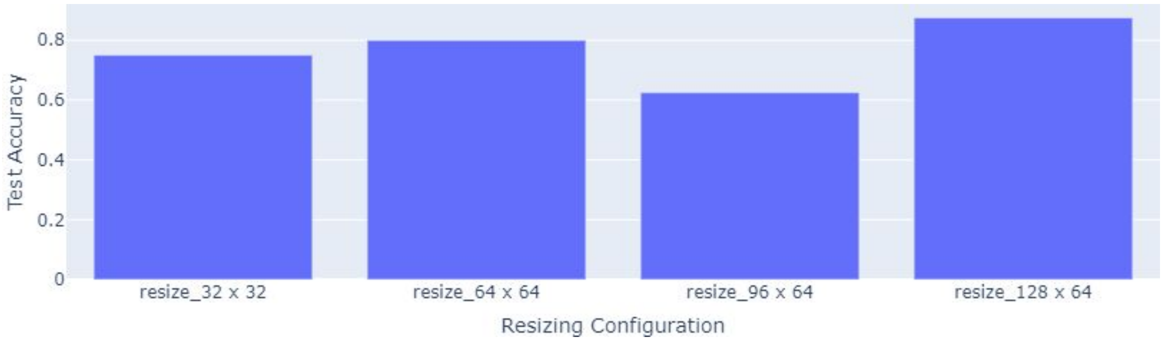
- Dropout of 0.2 after each layer with learnable parameters.
- Batch normalization after each layer with learnable parameters.

- Batch Size = **32**
- Learning Rate = **0.01**
- Optimizer = **Adam**
- Gradient Clipping = **0.5**
- Trained on **Cross Entropy Loss**
- Trained for **100 epochs**

CNN Model Tuning - Resizing

Processing Stage	**Optimal Value**
Resizing	64 x 64
Frames	
Temporal Strategy	
Motion Map	
Otsu Classes	



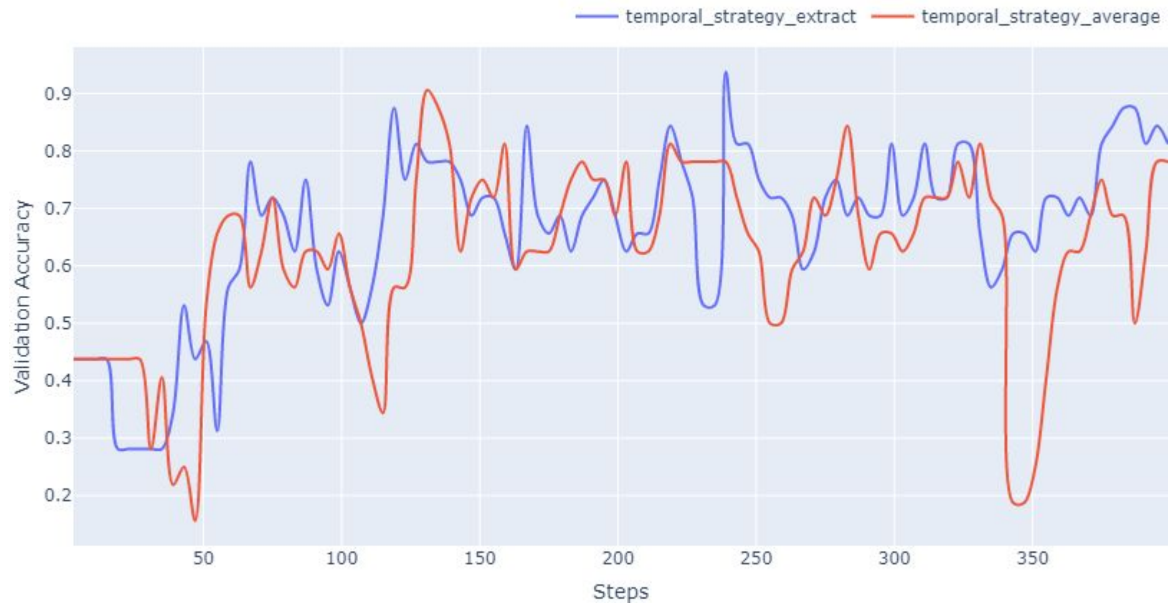


CNN Model Tuning - Frames

Processing Stage	Optimal Value
Resizing	64 x 64
Frames	49
Temporal Strategy	
Motion Map	
Otsu Classes	

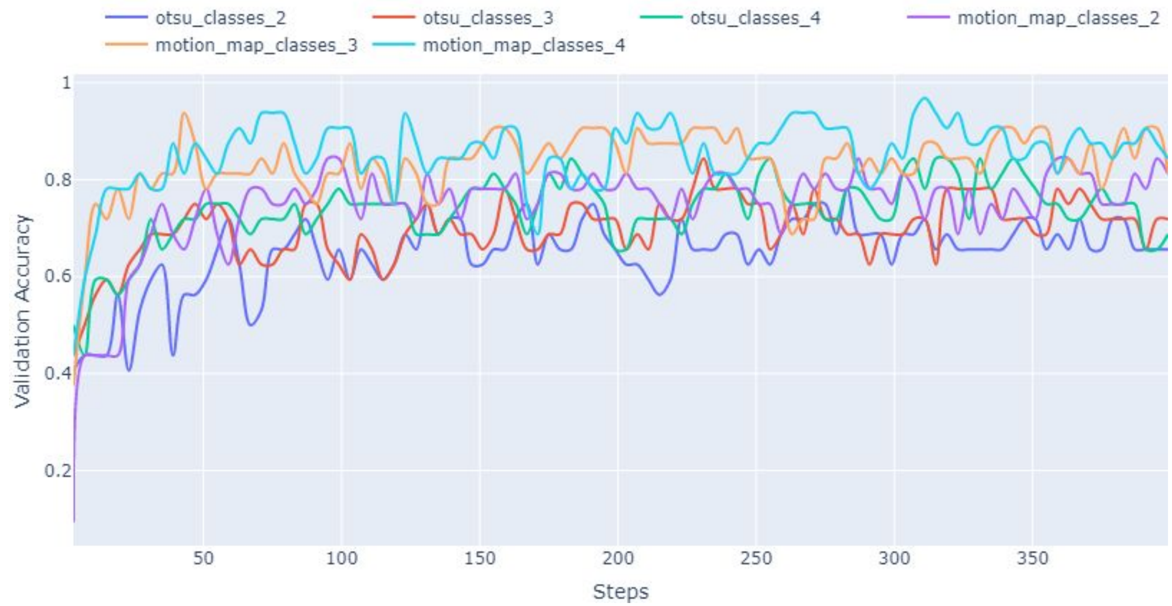
CNN Model Tuning - Temporal Strategy

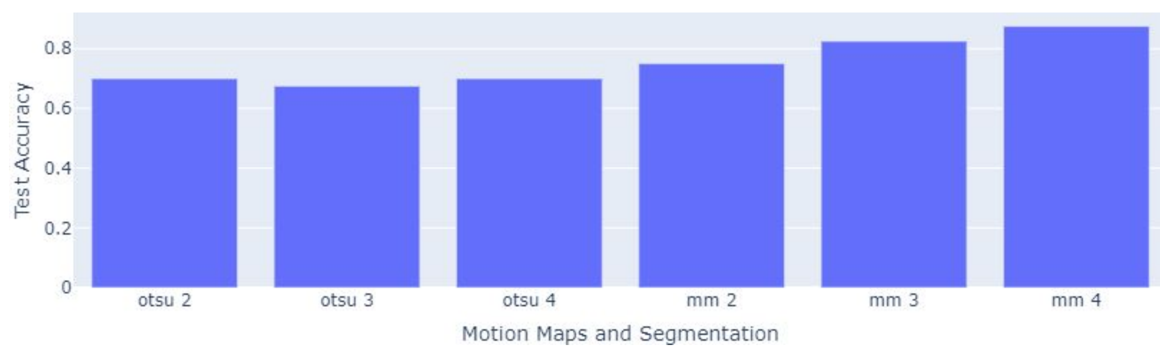
Processing Stage	Optimal Value
Resizing	64 x 64
Frames	49
Temporal Strategy	Average
Motion Map	
Otsu Classes	



CNN Model Tuning - Motion Map and Otsu

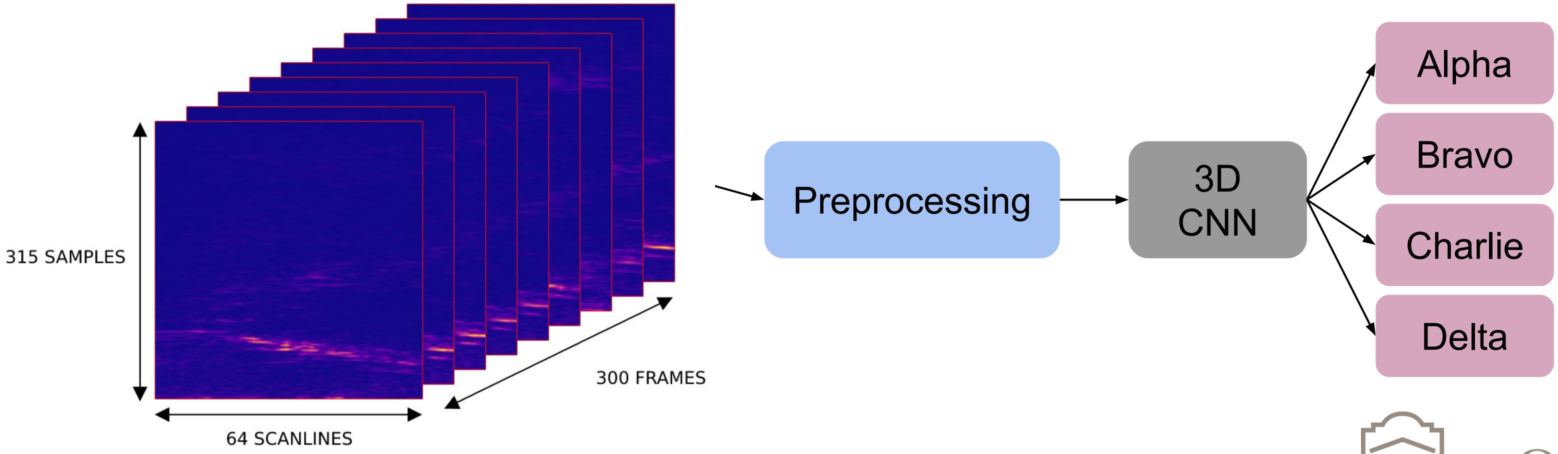
Processing Stage	**Optimal Value**
Resizing	64 x 64
Frames	49
Temporal Strategy	Average
Motion Map	True
Otsu Classes	3





3D CNN-Based Models

- Rather than stitch frames together to make one-big image, we can process sequence as a volume using 3D CNNs [6].
- 3D CNNs can learn frame-to-frame changes better (but more parameters needed).
- We implement a modified 3D CNN architecture described in [6] to compare with the CNN.



3D CNN Model Implementation

Conv3D(in=1, out=16, kernel=3)
MaxPool3D(kernel=2, stride=2)
Leaky ReLU

↓

Conv3D(in=1, out=16, kernel=5)
MaxPool3D(kernel=2, stride=2)
Leaky ReLU

↓

Flatten

↓

Linear(in=VARIABLE, out=16)
Leaky ReLU

↓

Linear(in=16, out=16)
Leaky ReLU

↓

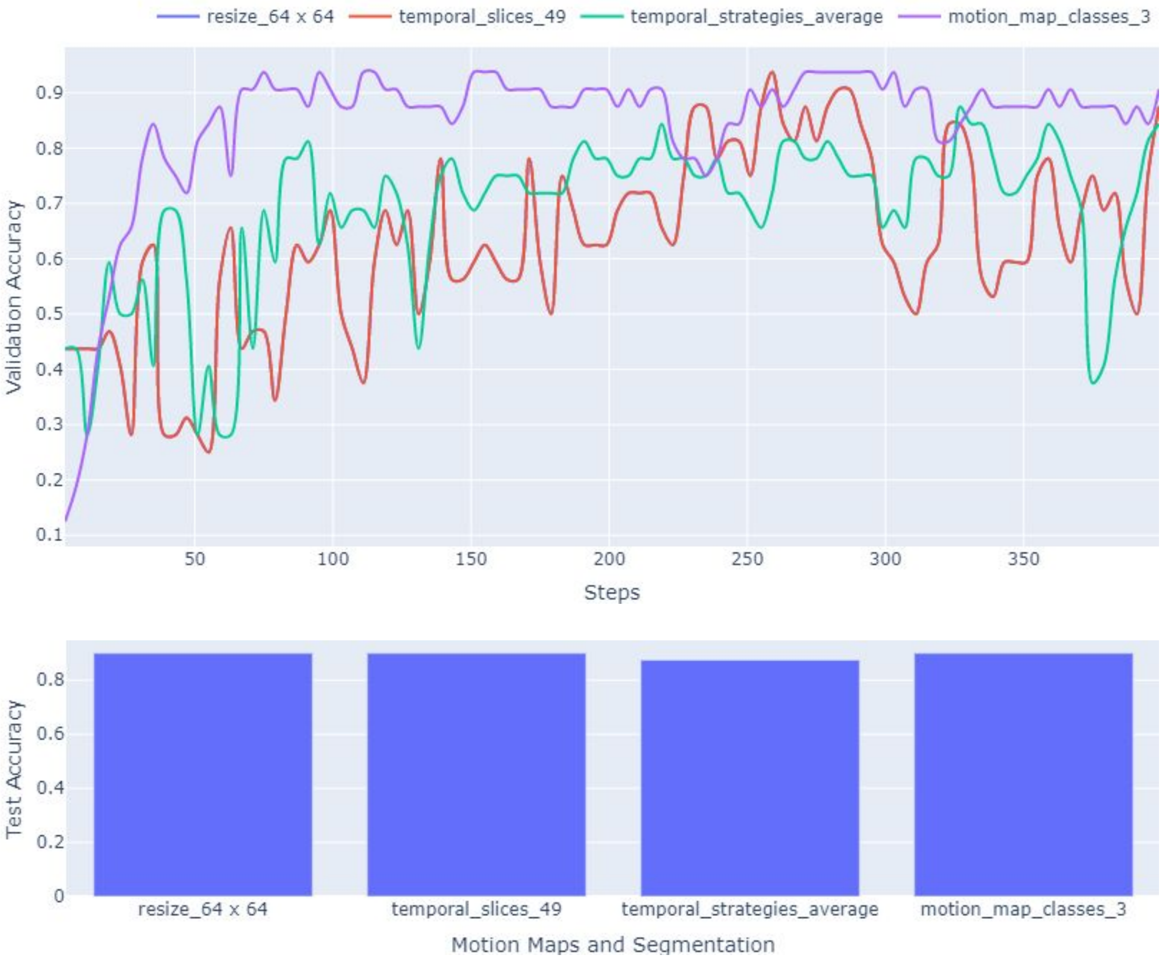
Linear(in=16, out=4)
Softmax

- Dropout of 0.2 after each layer with learnable parameters.
- Batch normalization after each layer with learnable parameters.

- Batch Size = **32**
- Learning Rate = **0.01**
- Optimizer = **Adam**
- Gradient Clipping = **0.5**
- Trained on **Cross Entropy Loss**
- Trained for **100 epochs**

3D CNN Model Tuning

Processing Stage	Optimal Value
Resizing	64 x 64
Frames	49
Temporal Strategy	Average
Motion Map	True
Otsu Classes	3



Test Results

CNN

- Accuracy: 85 %

3D CNN

- Accuracy: 92.5 %

Summary and Conclusions

- Collected a four-class dataset of UTI sequences for classification.
- CNNs and 3D CNNs were implemented to perform four-way classification on UTI sequences.
- Significant preprocessing is required for acceptable model performance, particularly noise-removal and image simplification (segmentation).
- Results show that the change in frames could be more important than the frames themselves for classification.
- 3D CNNs are better suited for the 3D volumetric data.

Future Work

- We will collect more samples from more subjects to generate more reliable test/validation performance estimates.
- Evaluate other common models in literature.
- Collect more classes (> 10) to investigate few-shot learning. techniques for classification of unseen UTI sequences classes.
- Investigate the CNN and 3D CNN models using Grad-CAM to understand why misclassifications occur and how to prevent them.

References

[1] A. Kapur, S. Kapur and Pattie Maes, "AlterEgo: A Personalized Wearable Silent Speech Interface", in *International Conference on Intelligent User Interfaces*, Tokyo, Japan, 2018, p.43-53

[2] C. Wang *et al.*, "Continuous Monitoring of Deep-Tissue Haemodynamics With Stretchable Ultrasound Phased Arrays", *Nature Biomedical Engineering*, vol. 5, no. 7, 2021

[3] E. M. Juanpere and T. G. Csapo, "Ultrasound-Based Silent Speech Interface Using Convolutional and Recurrent Neural Networks", *Acta Acustica united with Acustica,* vol. 105, no. 4, 2019

[4] N. Kimura, M. Kono and J. Rekimoto, "SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks", in *International Conference on Human-Computer Interaction*, Glasgow, Scotland, 2019, p.1-11

[5] M. S. Riberio, A. Eshky, K. Richmond and S. Renals, "Speaker-Independent Classification of Phonetic Segments From Raw Ultrasound in Child Speech", 2019, *arXiv:1907.01413v1*

[6] L. Toth and A. H. Shandiz, "3D Convolutional Neural Network for Ultrasound-Based Silent Speech Interfaces", 2014, *arXiv:2014.11532v1*