AttentionFCN: Multi-speaker Vocal Tract MRI Segmentation

Rishi Jain
UC Berkeley
rishiraij@berkeley.edu

Abstract

Hand-labeled segmentations of the vocal tract by articulatory imaging experts are extremely labor-intensive to produce. However, these segmentations provide information-dense representations that provide vital details for both research and medicine. Existing algorithms to segment vocal tract MRIs for mandibular, maxillary, and posterior regions are computationally expensive and have higher rates of error. In this paper, I propose a deep learning approach to solving this problem. I implement two different architectures, one being a more traditional fully convolutional network, and one including attention gating, to efficiently label these three regions with high accuracy without the bottleneck of requiring expert labels. I compare the two networks and demonstrate how the model can generalize well to unseen speakers.

1. Introduction

Articulatory representations are used in speech to understand the mechanics of speech production and connect the path from signals in the motor cortex to sound. A better scientific understanding of the dynamics of articulators is a key part of phonetics, and the translation of biological mechanics to audible features is a task that is at the basis of much of speech research. As a result, getting a better representation of articulatory features is integral in speech-related problems and is necessary to build fundamental speech models. In modern speech research, these articulatory representations are used in tasks including automatic speech recognition, neural speech decoding, and natural machine speech synthesis. These features are also of interest in medicine, especially in the form of speech pathology and voice coaching.

Dynamic imaging provides necessary insight into the shaping of the vocal tract. One of the most well-researched forms of dynamic imaging comes from electromagnetic articulography (EMA) where electromagnetic fields trace markers on seven key articulatory features in three-dimensional space. While this has led to breakthroughs in speech representation, the feature space is limited to the position of these points. Other works have used X-ray microbeams which collects data similar to EMA, electromyography (EMG) which directly measures electrical nerve response from external electrodes, and ultrasound for internal imaging. Real-time magnetic resonance imaging (rt-MRI) has been introduced into speech research as an alternative due to its higher spatial resolution and soft-tissue readings, providing readings in two or three-dimensional space. This is collected by generating a large-scale electromagnetic field which does not require explicit markers, instead providing a complete image of the tissue in the upper airway. For speech, we use the midsagittal plane, which gives unprecedented views of the pharynx, larynx, and pharyngeal wall on top of features such as jaw movement, lip position, and tongue articulators. MRI provides unparalleled high-dimensional signal for real-time speech without any known health consequences but lacks algorithms which extract signal relevant to speech applications. Existing approaches are unsupervised and involve hand-engineered anatomically informed object models which work in the high-noise frequency space [1], which suffers from bad cross-speaker zero-shot generalization.

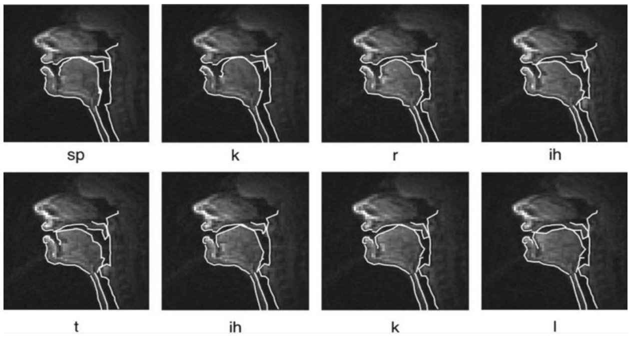

Figure 1: Visualization of region segmentation for the three articulators for a set of phonemes in the word "critical".

This paper aims to address this issue by providing a model to segment the mandibular, maxillary, and posterior regions of the rt-MRI's midsagittal plane. These segments are vital to understanding speech mechanics and are not

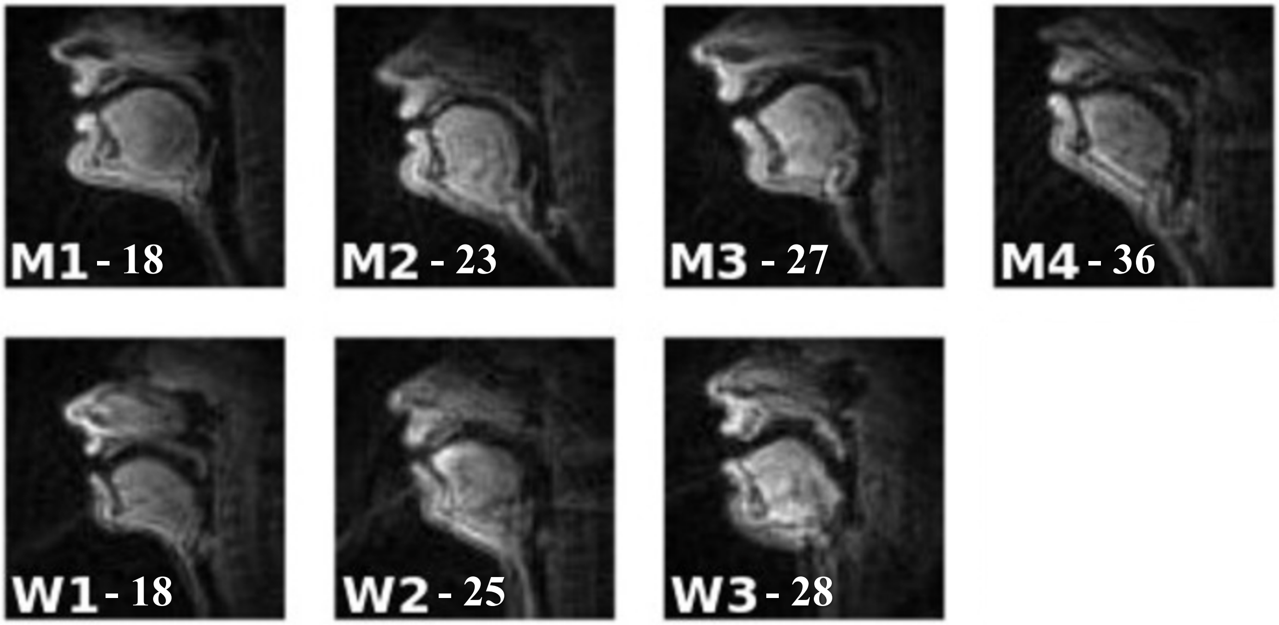

Figure 2: The seven speakers of the USC-TIMIT dataset. 'M' and 'W' indicate gender with the number on the right indicating speaker age. Speaker 'M4' was held out as validation to show the model's generalization across features.

possible to measure using the other imaging techniques mentioned above.

2. Methodology

2.1 Data Collection and Processing

The dataset being used is the University of Southern California's TIMIT database [2], available at https://sail.usc.edu/span/resources.html, which includes seven speakers (Figure 2) speaking phonetically diverse sentences with real-time MRI imaging at around 22 Hz. Subject read sentences off a projection screen as they lay supine in an MRI scanner with their heads steady. Images were reproduced from the two anterior coils, and the extraction of the midsagittal plane resulted in a spatial resolution of 84 x 84 pixels. Each of these videos have been supplemented with segmentations of the mandibular, maxillary, and posterior regions (Figure 1) by marking 40-70 points along the outlines, hand-selected for each frame by experts in articulatory imaging. For the purposes of this project, each frame was treated as a separate supervised region segmentation problem and the data was downsampled to a frequency of 2.2 Hz due to the minimal difference in signal of consecutive frames.

2.2 Architecture

Deep learning approaches to medical image segmentation have been recently outperforming more classical methods in a host of domains. After experimenting with convolutional and purely attention-based methods, the fully convolutional network was determined to have the best empirical result.

Fully Connected Network

The fully connected architecture as seen in Figure 3 is based on state-of-the-art medical region segmentation for low-resolution images. A variant of this architecture, first introduced as the U-net **Error! Reference source not found.**, has been shown to work well with reduced amount of input data. Because data was only available from seven speakers, this architecture provided the best fit while also generalizing to held out speakers.

The architecture begins with a contracting path. Each step in the contracting path includes two 3x3 convolutions followed by a 2x2 max-pooling layer to reduce the spatial dimension. Each subsequent layer has an increase in learnt feature maps in accordance with traditional CNNs. This is followed by an expansive path consisting of upsampling layers followed by two convolutional layers. Each downsampling and upsampling block ends with a batch normalization in contrast to the original U-net paper as this

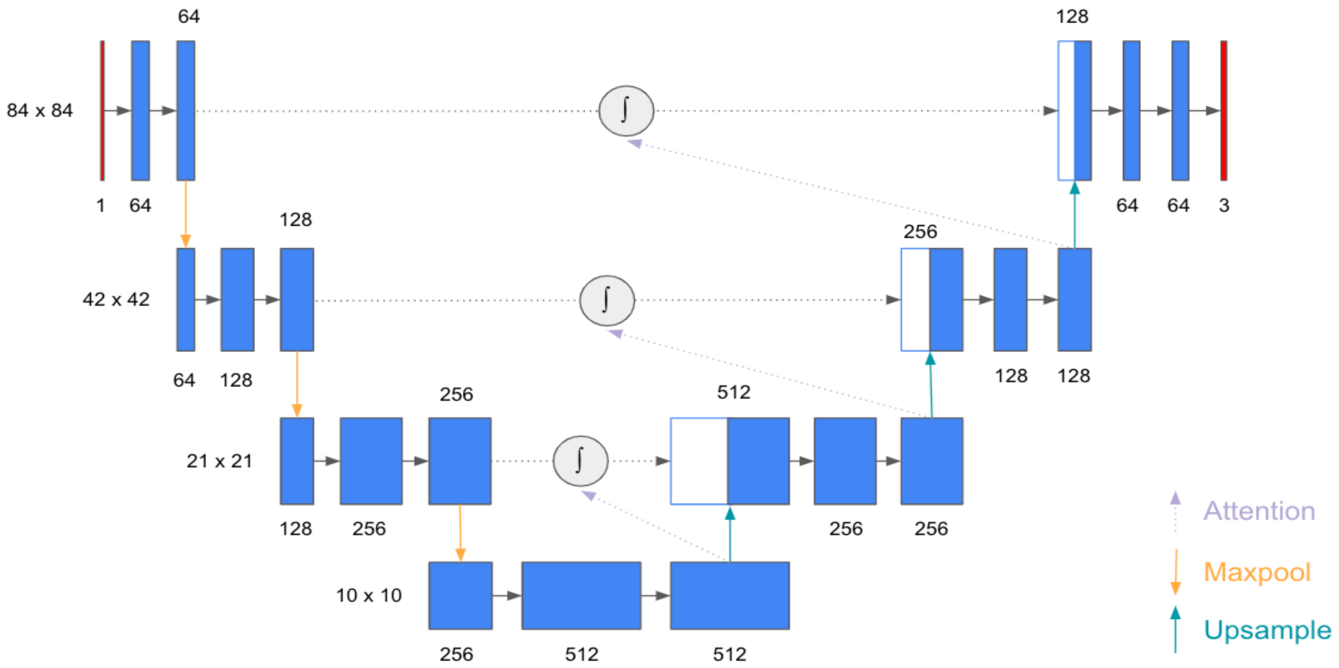

Figure 3: Fully connected network architecture used in this paper vocal tract segmentation. Attention gating is only used in the second model, but the structure of convolutions and pooling is the same. Batch normalizations exist implicitly after each block.

resulted in smoother training. Additionally, each convolutional layer is followed by a rectified linear unit (ReLU). After each upsampling block, there is a residual connection with the corresponding block of the contracting path. Both concatenation and addition were tried as the skip connection with concatenation showing better empirical results. The final layer has three feature maps of the same spatial dimension as the input. In order to get to this specific architecture, the number of downsampling and upsampling layers, as well as the number of feature maps of each layer were treated as hyperparameters that were experimented with to get the final architecture.

Attention Gating

The second model implemented maintains the same fully connected structure of the first one described above, with the addition of attention gating to the residual connection. The attention gate allows the model to add an additional weight to the activations of the residual connection from the downsizing portion of the model. The gating method considers the downstream information from the pre-upsampled activation to produce the weighting map that is then applied to the residual activation, as seen in Figure 3. The pre-upsampling activation is known as the gating vector *g* and the residual activation is the pixel vector *x*. Both *x* and *g* are sent through 1x1 convolutional layers to match the dimension of *g*. They are then added together and passed through a ReLU nonlinearity before passing through a 1 x 1 x 1 convolution to reduce the feature dimension to 1. The output of this is then passed through a sigmoid activation to get the attention weights α that are close to binary. This is upsampled to match the spatial dimension of the pixel vector and multiplied to the input vector to produce the final attention-gated residual connection to be concatenated. This can also be seen visually in Figure 4.

$$q^l_{att} = \psi^T \left(\sigma_1\left(W_x^T x_i^l + W_g^T g_i + b_g\right)\right) + b_\psi$$

$$\alpha_i^l = \sigma_2(q^l_{att}(x_i^l\, , g_i\, ; \Theta_{att})),$$

I chose to use additive attention between the two inputs rather than multiplicative, because this demonstrated better empirical results during training. This adds minimal overhead, with the concept first being introduced in the medical segmentation field by Oktay et al. **Error! Reference source not found.**.

2.3 Training Procedure

The training data was divided into a train and validation set, with the validation being the hold-out of all the videos of one speaker. This was done to ensure that the model is able to generalize across speakers, as each speaker has slightly different physical characteristics. This was done in a cross-validation method while tuning hyperparameters before choosing the speaker who was most out-of-distribution for final test accuracy. During training, data was shuffled into minibatches of 128 input frames across

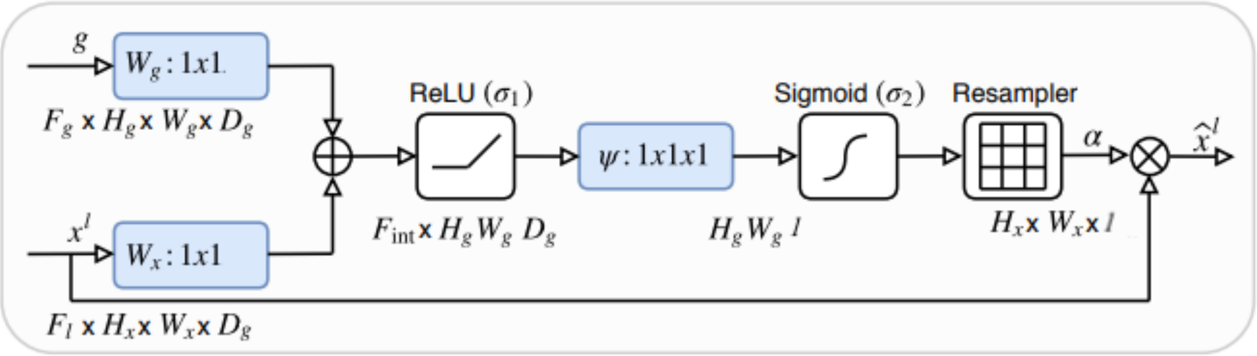

Figure 4: A visual description of my implementation of attention gating, which is slightly modified from the original paper by Oktay et al.

training speakers and trained on a single Nvidia RTX A5000 GPU. Training lasted a total of 8 epochs with each epoch having around 30,000 frames of training data spread across the 6 speakers. After experimenting, the optimal learning rate for both models was determined to be 0.0001.

To evaluate the performance of the model, I chose the Dice similarity coefficient because, similar to precision, it works by both rewarding the positive matches, but also penalizes false positives in the region segmentation task **Error! Reference source not found.**. For two regions *A* and *B*, Dice similarity is calculated by:

$$\frac{2|A \cap B|}{|A| + |B|}$$

However, because this is not differentiable, the proxy training loss is element wise binary cross-entropy loss for the three output segmentation regions, represented by *c*. The following is applied after applying a sigmoid activation to the logits outputted by the network:

$$H_p = -\frac{1}{HW}\sum_{c=1}^{3} w_c \sum_{i,j} y_{c,i,j} \log\left(p(y_{c,i,j})\right) + (1 - y_{c,i,j}) \log\left(1 - p(y_{c,i,j})\right)$$

Here, the weighting term *w* corresponds to weights of each output channel, one corresponding to each specific region. The weights are calculated by summing over the pixels of each class in the training set, normalizing this so that the total sum over classes is 0, and subtracting each weight from 1. As a result, classes with more pixels will not be weighed more than classes with less. Further, the posterior region segmentation has a more undefined boundary as it includes part of the spinal column and is thus arbitrarily bounded on one side. As a result, the weight for this class was additionally reduced by 50%.

3. Conclusion

3.1 Results

The results of training on the 6-speaker training set were evaluated using the dice similarity coefficient on the held-out 'M4' validation speaker. The age of 'M4' and the difference in physical features made the subject a good measure of the model's generalization, as the intention is to use the model to label the larger unlabeled upper-airway rt-MRI available in the TIMIT dataset, and other future data that is being collected, that is audio-aligned. The best results of the two models, dubbed 'FCN' and 'AttentionFCN' are shown below. There is no comparable baseline for this task.

Model	Dice Score
FCN	0.945
AttentionFCN	**0.948**

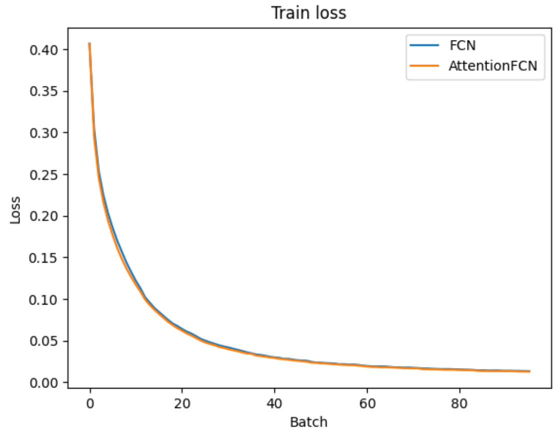

Figure 5: The training loss curves for the two models over 8 epochs. Each batch plotted here is the average loss of approximately 1300 input images.

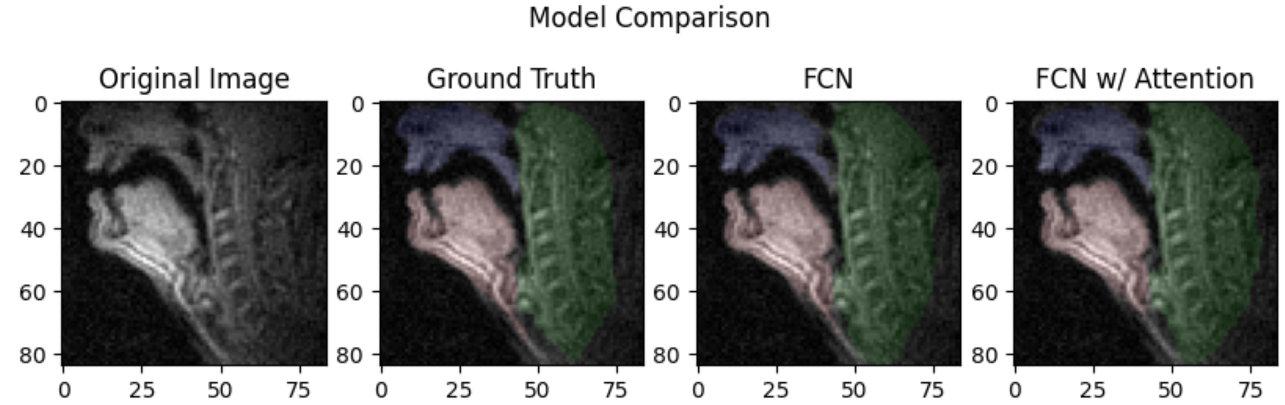

Figure 6: Comparison of the ground truth segmentation to the FCN and AttentionFCN predictions on a frame from the validation set, overlayed on the original frame from the rt-MRI.

While the addition of attention gating is slightly better than without, the difference in performance of the two models is negligible.

The loss curves for the training of the two models can also be seen in Figure 5. Once again, both models train at approximately the same rate, with the slightly better training loss of the model with attention gating being negligible. Both models appear to empirically be equivalent in this case. One hypothesis for why this is may be is that the spatial dimension of the original input is so small that gating is unnecessary, and essentially maintains most of the activations anyway (attention weight map is close to 1 for all indices). It could also be due to the smaller size of the model compared to the standard FCN-8. This could mean that the addition of attention gating is unnecessary in this problem because the information about the relevant activations spatially is not lost. Figure 6 also shows a visualization of the three segmented regions by the two models on a validation image. Both model predictions are very similar, and almost indiscernible from the ground truth segmentation.

3.2 Next Steps

Empirically, both models are performing well enough to use in inference. Due to the smaller size of the model, they run fast enough to be used to label large datasets, allowing these information-dense labels to be used in downstream representation learning tasks. The trained AttentionFCN is currently being applied to the larger 75-speaker dataset. An example of one frame from one of these videos (which were recorded during a different clinical study) is shown in Figure 7 with the entire video available online [6].

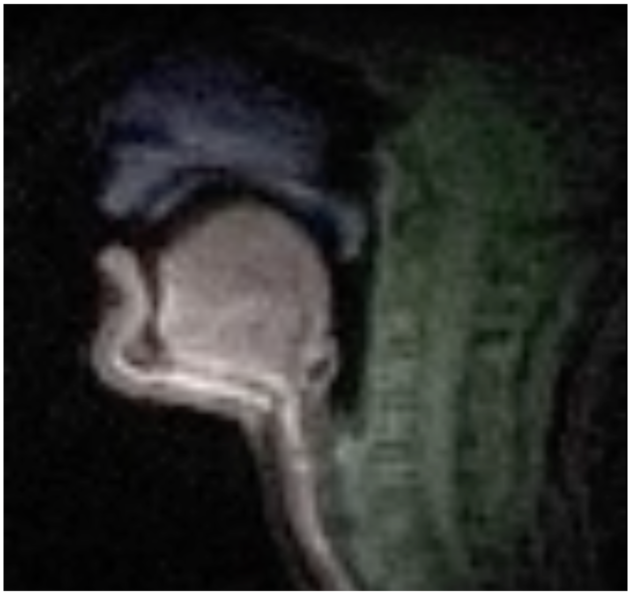

Figure 7: Test segmentation overlayed on a frame from a different rt-MRI study of the upper airway recorded under different conditions. The accuracy of the predictions show that the models generalize well to other data.

I would like to use these segmentations to come up with speaker-independent representations that map to phonemes in the aligned audio. This can then be used to map signals from the brain such as electrocorticographic (ECOG) and cerebral near-infrared spectroscopy (NIRS) signals that are currently being studied. In addition, since the movement of these segmented regions is important to understand speech mechanics, it may be interesting to look at optical flow to see whether motion estimation can provide context to what is being said.

The code for the project is available at: https://github.com/rishiraij/CS280_final_project.git [7].

References

[1] Bresch E, et al. Region Segmentation in the Frequency Domain Applied to Upper Airway Real-Time Magnetic Resonance Images. *PubMed Central*. 2009; PMID: 19244005

[2] Narayanan S, et al. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *J. Acoust. Soc. Am.* 2014;136(3):1307–1311.

[3] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. *Springer* (2015)

[4] Oktay O, et al. Attention U-Net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018).

[5] Dice LR. Measures of the amount of ecologic association between species. *Ecology*. 1945;26:297–302.

[6] https://drive.google.com/file/d/1LgZOyzUedKUx7E5MhPJOymjVayDRxVU2/view?usp=sharing

[7] https://github.com/rishiraij/CS280_final_project.git