Predicting Speech Formants using Vocal Tract MRI Video

Devin Murphy
MIT EECS
devinmur@mit.edu

Abstract

Speech production relies on the coordinated action of the vocal cords and the movement of the vocal tract to generate a variety of sounds. The vocal tract transfer function can be used to characterize the acoustic properties of unvoiced and voiced sounds, but most techniques for estimating this function rely solely on audio signals. In order to better understand the relationship between the physiology of the vocal tract and the acoustic properties of its produced sounds, this paper investigates the feasibility of using computer vision techniques to accurately predict the vocal tract transfer function coefficients from vocal tract MRI images. We construct a dataset by employing linear predictive coding to map 2D Sagittal-view MRI frames to vocal tract transfer function coefficients. Subsequently, we develop and evaluate a CNN-RNN architecture trained on these MRI frame and coefficient vector pairs. While our architecture tends to predict the mean coefficients of our datasets, we demonstrate the potential for added generalization capabilities provided by a combined CNN-RNN architecture, as well as the ability to learn meaningful representations for understanding speech production mechanisms.

1. Introduction

It is well-established that speech can be represented as the output of a time varying source-filter system, where the vocal tract acts as the filter and vibrations produced by the epiglottis serve as the source [5]. The time-varying filter can be characterized by an all-pole transfer function if we approximate the speech signal as a weighted sum of its past values and current values:

$$s[n] = \sum_{k=1}^{p} a_k \cdot s[n-k] + G \cdot u[n] \tag{1}$$

Here, $s[n]$ denotes the speech audio signal, G is the gain, $u[n]$ denotes the current source value , and the a_k's represent the filter coefficients.

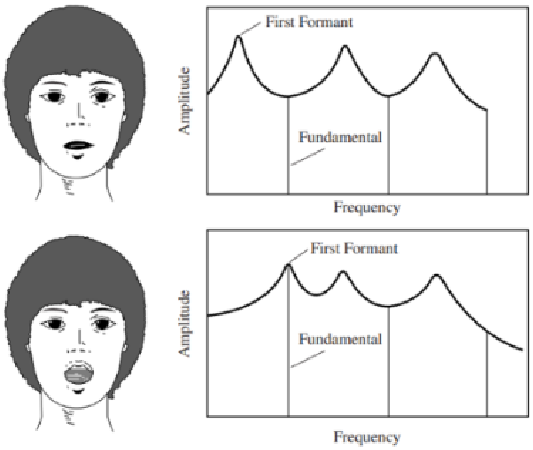

Figure 1. Opera singer changes shape of vocal tract to align first formant with fundamental frequency

The filter coefficients determine the location and magnitude of resonant frequencies of the vocal tract transfer function, known as ”formants”. These formants account for a lot of the variation in speech production, especially for vowels [2], [9]. Traditionally, Linear Predictive Coding is employed to extract these coefficients from segments of an audio signal.

However, if the vocal tract's shape indeed primarily influences the filter as posited, there could be significant benefits in predicting these coefficients directly from vocal tract images. This information could empower opera singers to optimize their mouth shapes for enhanced dynamic range in unamplified settings [1], aid language learners in pinpointing pronunciation nuances, and assist speech therapists in visualizing necessary physiological adjustments for their clients' progress. Our objective is to showcase the capability of our network in accurately predicting coefficients to estimate the first, second, and third formants (F1, F2, and F3) of human speech. Additionally, we aim to demonstrate the potential for reconstructing the audio signal associated with a set of MRI frames using Linear Prediction Vocoder resynthesis techniques and the vocal tract transfer functions predicted by our network. To further elucidate the model's decision-making process, we will employ gradCam visualizations to highlight regions of the vocal tract that the model found pivotal for accurately predicting the vocal tract transfer function. This approach not only enhances interpretability but also provides valuable insights into the anatomical features crucial for speech production.

1.1. Related Works

Previous research has explored the use of neural networks for estimating and tracking formants in speech signals. Notably, DeepFormants employed a feed-forward network architecture to estimate the first three formants from audio input and utilized a Recurrent Neural Network (RNN) to predict sequences of formants [1]. While DeepFormants demonstrated improved performance over traditional formant tracking methods, it relied on direct features of the audio signal for formant retrieval. In contrast, our approach capitalizes on the filtering capabilities of convolutional neural networks (CNNs) to discern patterns between vocal tract images and formant frequencies.

Our proposed method of direct synthesis, wherein spectral coefficients are predicted from images and employed to resynthesize speech using a vocoder, draws inspiration from existing research in Silent Speech Interfaces [6]. These interfaces facilitate speech communication solely through silent articulation, without producing audible sound. For instance, Moliner et. al used ultrasound images of the tongue in a network comprising Convolutional and Bidirectional LSTM layers to resynthesize Hungarian language utterances [4]. While this method effectively synthesized speech from ultrasound images, our approach offers a novel perspective by utilizing 2D sagittal view MRI data, providing a comprehensive view of the entire vocal tract and potentially yielding new insights into the impact of articulators on speech acoustics.

Furthermore, our CNN incorporates transfer learning by adopting the Resnet50 architecture, featuring a significantly higher number of convolutional layers (48 in our model versus 2 in the previous work). This facilitates more nuanced feature extraction, enhancing the model's ability to discern intricate relationships between vocal tract images and formant frequencies. Additionally, while the network in [4] was trained on speech data from a single female subject speaking only in Hungarian, our MRI dataset encompasses data from 75 diverse speakers across various demographics. This broader dataset enables our model to account for greater variation in vocal production across demographic groups, potentially enhancing its generalization capabilities.

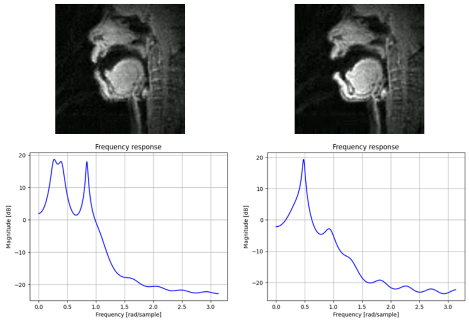

Figure 2. Vocal Tract MRI Input paired with Transfer Function Label

2. Proposed Methods

We implemented a CNN-RNN architecture for a regression task on estimated vocal tract transfer function, leveraging MRI images as inputs. The MRI images and vocal tract transfer functions were obtained through the processing of a multi-speaker dataset of real-time speech production MRI video [3]. Initially, we fine-tuned a ResNet50 architecture to predict the vocal tract transfer function coefficients corresponding to each frame, thereby learning a 256-element embedding of MRI image features. Subsequently, a recurrent neural network with LSTM layers was employed. Its primary objective was to ingest a sequence of MRI frames and capture temporal patterns in the features derived from our pretrained CNN, enabling the prediction of LP coefficients for the last time step in the sequence. For a visual representation of this methodology, refer to figure 3.

2.1. Data Preprocessing

For each MP4 file containing MRI video data from the initial 18 subjects within our dataset, we employed a multi-step process for extracting the input and output tensors for training. Initially, we extracted the audio signal from these files and subjected it to a lowpass filter to eliminate high-frequency noise. Subsequently, we utilized the 'detectSpeech' function from MATLAB's Audio Toolbox to isolate segments of active speech by the user.

To obtain the vocal tract filter coefficients, we used a Linear Predictive Coding (LPC) algorithm with a desired filter order of $p = 18$, a frame length of $fL = 50$ milliseconds, and a hop size h of 25 milliseconds. This LPC analysis yielded a length 20 coefficient vector at a rate of 40 frames per second, with each vector consisting of 18 coefficients plus an additional coefficient for $a_0 = 1$ and one for gain.

However, it's noteworthy that the frame rate of the MRI video is approximately 84 frames per second, leading to more MRI frames than coefficient vectors for a given audio segment. To ensure a 1:1 ratio of input to label in our

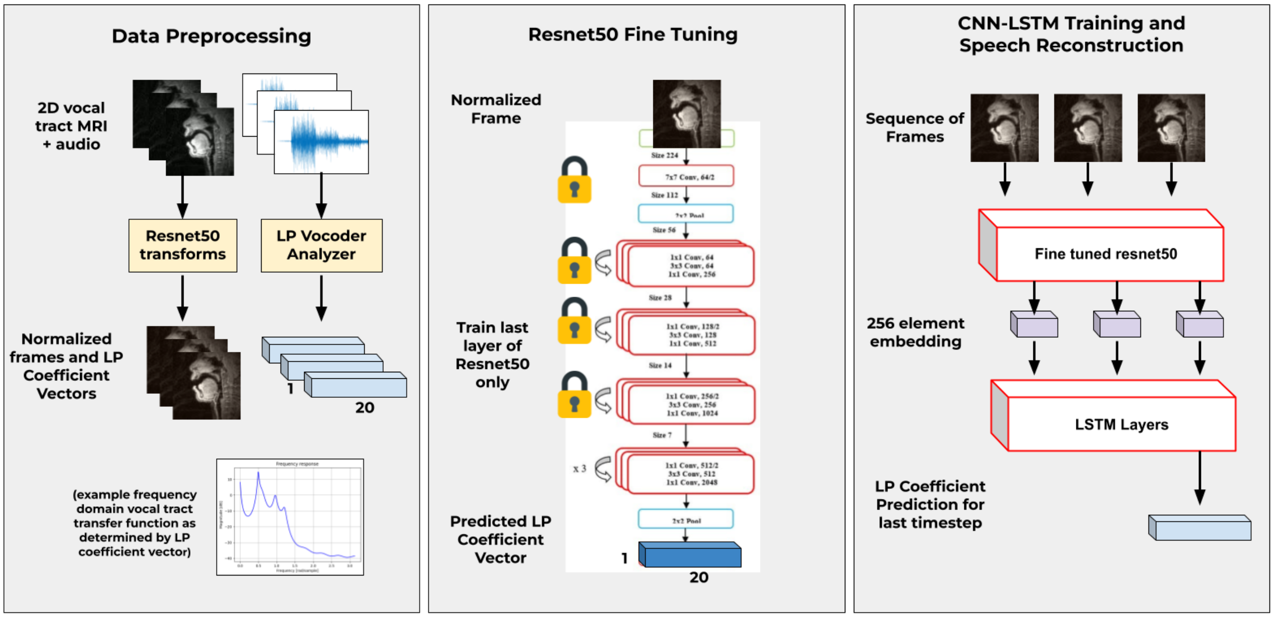

Figure 3. Overview of methods used to predict vocal tract transfer function from 2D vocal tract MRI video. Video frames and audio are separated, and Linear prediction (LP) coefficients are obtained on the audio signal associated with each frame. A Resnet50 CNN is then fine tuned for a linear regression task on the LP coefficients using the normalized MRI video frames as inputs. This tuned CNN is then used to train a CNN-LSTM model for higher accuracy prediction of the LP coefficients, which can then be used to resynthesize the original audio signal associated with a set of MRI image frames.

dataset, we downsampled the video frames within each segment to match the number of coefficient vectors. Examples of MRI frame inputs paired with their vocal tract transfer function label can be seen in figure 2.

Subsequently, we preprocessed the extracted frames using the default weights and transforms provided by the ResNet50 model within the torchvision Python package. In summary, our data preprocessing pipeline resulted in a total of 163,022 image-coefficient pairs. We partitioned these pairs, with 80% allocated for training and the remaining 20% for validation, to facilitate robust model training and evaluation.

2.2. Resnet50 Fine Tuning

In order to harness meaningful features from our MRI video frames for prediction, we fine-tune a pretrained ResNet50 architecture. This has proven successful in classification tasks for MRI images in other contexts [10]. Initially, we trained our network using the default weights tailored for the ImageNet dataset, dedicating approximately a quarter of our training dataset for 10 epochs while utilizing and Adam optimizer with Mean Squared Error (MSE) loss and a learning rate scheduler. However, during this experiment, we encountered overfitting tendencies around epoch 7, as evident in Figure 4.

To mitigate overfitting concerns and expedite training, we opted to freeze the first three layers of the network and retrain it on our entire dataset. This adjustment significantly reduced the number of trainable parameters to around 50,000. The decision to freeze the earlier layers was informed by the understanding that these layers capture more general features, thereby offering a balance between addressing overfitting and preserving the CNN's ability to extract relevant information.

After implementing these adjustments, our model demonstrated convergence, reaching a validation MSE of approximately 0.226. This refined model was then utilized as the feature extractor for our subsequent Recurrent Neural Network (RNN) model, as detailed in the following section. Additionally, we conducted a qualitative evaluation of the learned representations by exploring gradCam activations for layer 4, which are discussed in the experimental results section of our paper.

2.3. CNN-RNN Training

In order to capture the temporal patterns in speech production, we designed a CNN-RNN architecture, as has been done for other applications such as automatic speech recognition [8]. Initially, we replaced the last fully connected layer of a pre-trained ResNet50 with a linear layer, aug-

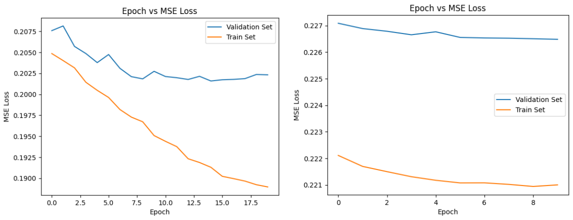

Figure 4. Epoch vs MSE loss in training of Resnet50 without frozen layers (left) and with frozen layers (right). This appears to result in a more stable training paradigm

menting the feature space to 256 dimensions for our LSTM model. Our network comprises two LSTM layers with a hidden size of 128, followed by ReLU activation, batch normalization, and dropout (p = 0.1) to mitigate overfitting. During inference, sequences of 5 MRI frames are fed into ResNet50, generating 256-dimensional embeddings. These embeddings are then sequentially processed by the LSTM layers to predict LP coefficients for the final timestep of each sequence. We trained the network for 5 epochs using an Adam optimizer with MSE loss, updating only the last linear layer of ResNet50 and LSTM parameters. Our model achieved a final MSE loss of 0.2125 on the validation set.

To assess model performance, we use visual comparison of predicted and labeled spectra, focusing on formant peak locations. Additionally, we use MSE loss comparison between the predicted coefficient vectors of the Resnet50 CNN by itself and the CNN-RNN network on our test set.

3. Experimental Results and Evaluation

After training our CNN and CNN-RNN networks for 10 and 5 epochs, respectively, we evaluate them on unseen MRI frames for subject 42 of our dataset. As shown in figure 5, while a lower MSE is seen for the CNN on our training data, the combined CNN and RNN network achieves a lower loss on the validation and test sets. This indicates that this architecture may generalize better to unseen inputs.

In initial visual inspection of the vocal tract transfer function exhibited by predicted and ground truth LP coefficients, we can notice some similar characteristics. For example, the left transfer function in fig 6 shows that for some of our video frames, the location of resonant peaks, or Formants, in the predicted transfer function (blue) are almost identical to those of the actual transfer function.

However, upon further inspection we see that the model has a pattern of predicting three formant frequencies around 581 hz, 1905 hz, and 2993 hz. As our loss function centered around minimizing MSE, we should be skeptical that our model is actually predicting the "mean" of each coefficient in our length 20 vector across all of our data points.

Further analysis of the distribution of the labeled coefficient vectors and predicted coefficient vectors for our test dataset confirms this suspicion. In figures 7, 8, and 9, we show the means and variances for each of our 19 LP coefficients (excluding gain). The figures show that while the model learns to output coefficients in distributions with similar means to our labels, it is often predicting values very close to this mean rather than learning the discriminating capabilities necessary to make more varied predictions.

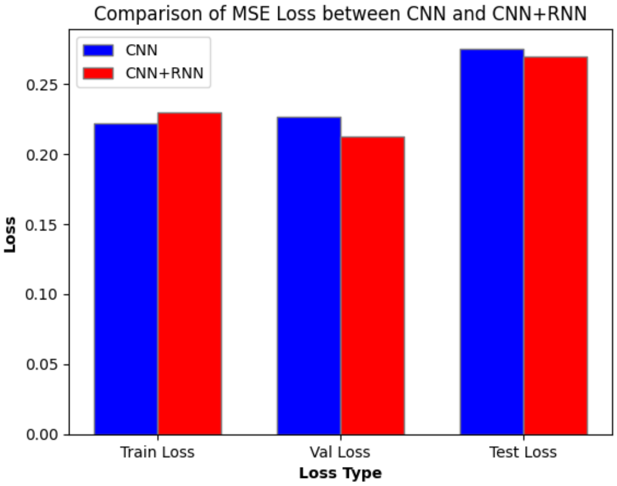

Figure 5. MSE Loss of predicted LP coefficient vectors by network and dataset type

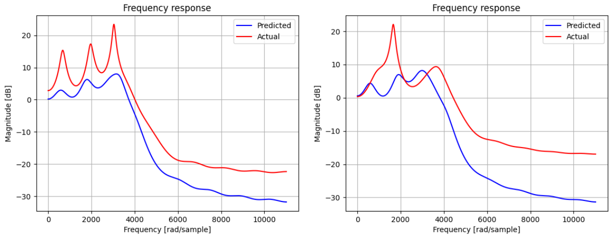

Figure 6. Frequency responses for predicted and ground truth LP coefficient vectors. The plot on the left shows an occasion where our prediction happens to match the label quite well, while the right plot hints at model prediction of the mean

Finally, in order to evaluate what features of the vocal tract our CNN finds most useful for predicting these acoustic properties such as formants, we plot the gradCam activations for the unfrozen layer of our Resnet50 fine-tuned model for specific LP coefficients/MRI image pairs in our train dataset [7].

Figures 10 and 11 demonstrate the potential of Resnet50 to learn meaningful representations for speech understanding, as we can see clear anatomical representations appearing for predicting these coefficients.

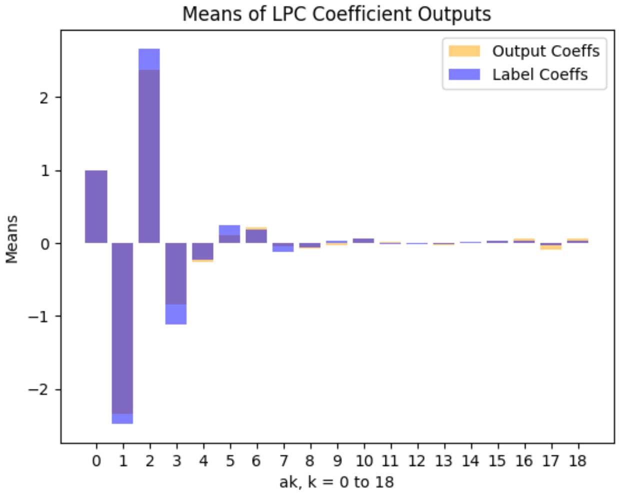

Figure 7. Means for each of the 19 LP coefficients across outputs and labels for our test set

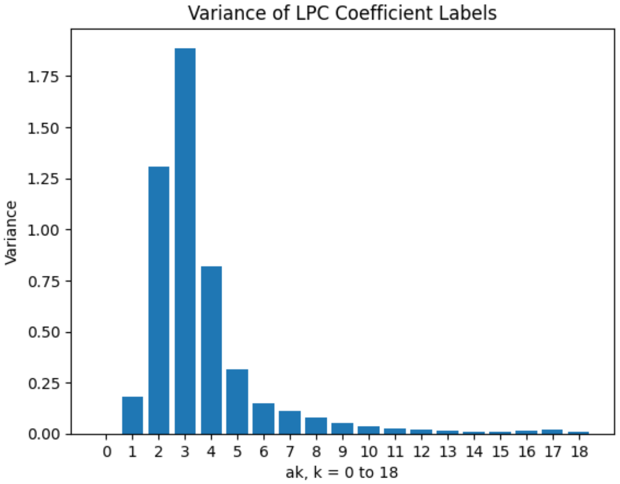

Figure 8. Variances for each of the 19 LP coefficients across labels for our test set

References

[1] Yehoshua Dissen and Joseph Keshet. Formant estimation and tracking using deep learning. pages 958–962, 2016. 2

[2] Matthias Echternach, Johan Sundberg, Thomas Baumann, Michael Markl, and Bernhard Richter. Vocal tract area functions and formant frequencies in opera tenors' modal and falsetto registers. *Journal of the Acoustical Society of America*, 129(6):3955–3963, 2011. 1

[3] Yongwan Lim, Asterios Toutios, Yannick Bliesener, Ye Tian, Sajan Goud Lingala, Colin Vaz, and et al. A multispeaker dataset of raw and reconstructed speech production real-time mri video and 3d volumetric images. https://doi.org/10.6084/m9.figshare.13725546.v1, 2021. 2

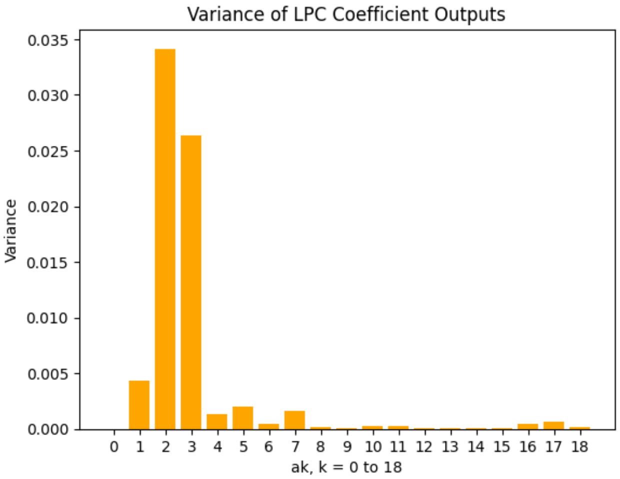

Figure 9. Variances for each of the 19 LP coefficients across model outputs for our test

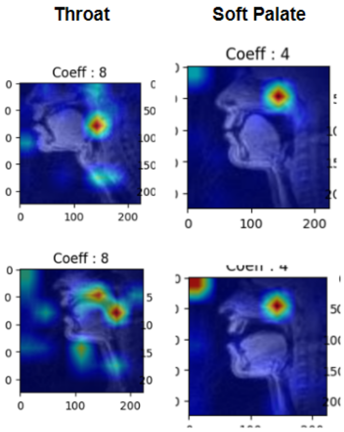

Figure 10. GradCam activations highlighting the throat and soft palate

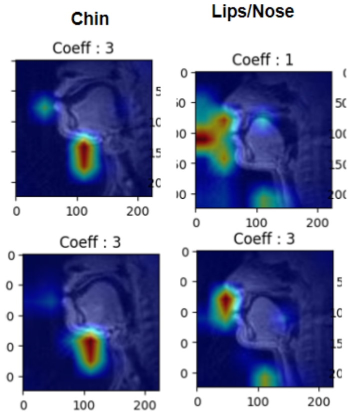

Figure 11. GradCam activations highlighting the nose/lips and chin

[4] Eloi Moliner and Tamás Csapó. Ultrasound-based silent speech interface using convolutional and recurrent neural networks. *Acta Acustica united with Acustica*, 105, 2019. 2

[5] Thomas F. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice Hall, 2002. Accessed May 13, 2024. 1

[6] Pramit Saha, Yadong Liu, Bryan Gick, and Sidney Fels. Ultra2speech – a deep learning framework for formant frequency estimation and tracking from ultrasound tongue images, 2020. 2

[7] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2019. 4

[8] M Soundarya, P R Karthikeyan, and Gunasekar Thangarasu. Automatic speech recognition trained with convolutional neural network and predicted with recurrent neural network. In *2023 9th International Conference on Electrical Energy Systems (ICEES)*, pages 41–45, 2023. 3

[9] Ingo R. Titze, Lewis M. Maxfield, and Melinda C. Walker. A formant range profile for singers. *Journal of Voice*, 31(3): 382.e9–382.e13, 2017. Epub 2016 Oct 28. 1

[10] Yue Zhang, Yuting Lucy Liu, Ke Nie, Jie Zhou, Zhimin Chen, Junting Helen Chen, Xiaohui Wang, Byung Kim, Rajan Parajuli, Rita S Mehta, Meng Wang, and Min-Ying Su. Deep learning-based automatic diagnosis of breast cancer on mri using mask r-cnn for detection followed by resnet50 for classification. *Academic Radiology*, 30(Suppl 2):S161–S171, 2023. 3