Silent Speech and Emotion Recognition from Vocal Tract Shape Dynamics in Real-Time MRI

Laxmi Pandey
Human-Computer Interaction Group
University of California, Merced
California, Merced, USA
lpandey@ucmerced.edu

Ahmed Sabbir Arif
Human-Computer Interaction Group
University of California, Merced
California, Merced, USA
asarif@ucmerced.edu

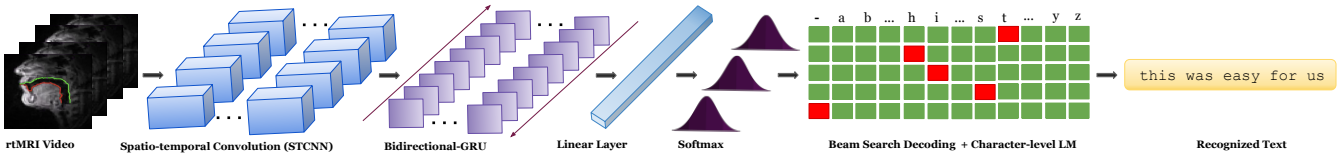

Figure 1: An overview of the proposed model: classification of 2D real-time MRI (rtMRI) of vocal tract shaping into text with an end-to-end deep neural network. A sequence of frames is used as input that is processed by three layers of STCNN for spatiotemporal feature extraction. The extracted features are processed by two Bi-GRUs, followed by a linear layer and a softmax layer. The softmax output is then decoded with prefix beam search with the help of a language model.

ABSTRACT

Speech sounds of spoken language are obtained by varying configuration of the articulators surrounding the vocal tract. They contain abundant information that can be utilized to better understand the underlying mechanism of human speech production. We propose a novel deep neural network-based learning framework that understands acoustic information in the variable-length sequence of vocal tract shaping during speech production, captured by real-time magnetic resonance imaging (rtMRI), and translate it into text. The proposed framework comprises of spatiotemporal convolutions, a recurrent network, and the connectionist temporal classification loss, trained entirely end-to-end. On the USC-TIMIT corpus, the model achieved a 40.6% PER at sentence-level, much better compared to the existing models. To the best of our knowledge, this is the first study that demonstrates the recognition of entire spoken sentence based on an individual's articulatory motions captured by rtMRI video. We also performed an analysis of variations in the geometry of articulation in each sub-regions of the vocal tract (i.e., pharyngeal, velar and dorsal, hard palate, labial constriction region) with respect to different emotions and genders. Results suggest that each sub-regions distortion is affected by both emotion and gender.

KEYWORDS

Speech, silent speech, recognition, neural networks, real-time MRI, vocal tract, accessibility.

SIGGRAPH 2021, August 9–13, 2021, Virtual

1 INTRODUCTION

The vocal tract is the most important component of human speech production that starts at the vocal cords, continues upwards towards the tongue, and ends at the lips [19]. During air expulsion, this tubular passageway changes its position and shape to produce various sounds and their acoustic representations. Estimating and mapping vocal tract configuration to its corresponding acoustic parameters has long been a challenge not only because it is difficult to access the vocal tract but also due to its complex biological structure and rapid movement of its articulators [15]. The speech production process is essentially non-stationary—generally the rapid transition between different articulatory states generates the speech sounds. Hence, extraction of acoustic information embedded in the vocal tract geometry is crucial for recognizing and synthesizing speech.

Recent development in real-time magnetic resonance imaging (rtMRI) makes it possible to acquire complex spatiotemporal visual information about the dynamic shaping of the vocal tract in speech production needed for speech analysis. Unlike X-ray or electromagnetic articulography (EMA), rtMRI does not use potentially hazardous radiation or place extramural devices in the mouth that could interfere with the articulator's movement. This method is also suitable for patients with vocal tract pathology, such as those that have a partial tongue resection or experience pain in the areas responsible for producing speech. The University of Southern California collected a speech production dataset[1] that includes rtMRI data from ten native speakers of general American English [16]. We utilized this dataset to investigate whether it is possible to recognize continuous speech from vocal tract geometry.

For this, we built a deep neural network-based learning framework that can automatically estimate the acoustic information corresponding to a specific vocal tract configuration, called articulatory-to-acoustic mapping, for continuous speech recognition. This, we

[1]USC-TIMIT: A Database of Multimodal Speech Production Data, https://sail.usc.edu/span/usc-timit

believe, is the first end-to-end sentence-level articulatory speech recognition[2] framework for rtMRI data that simultaneously learns spatiotemporal visual features and sequential information. In addition, we performed an extensive analysis on the MR images of emotion-dependent vocal tract movements to compare different emotions (neutral, happy, angry, sad) and genders (female, male) using the data[3] collected in a previous work [11]. An understanding of whether and how emotion affects articulatory movements during speech production is important to reduce ambiguity in recognized sentences. For example, the emotional context of the sentence "I hate you" could inform the system whether it was said sarcastically or literally. The effects of gender on vocal tract movements, in contrast, can increase the accuracy of the recognition system.

We envision numerous applications of this framework. It could be used to input text and communicate with various computer systems using speech or silent speech [17], which are arguably more natural modes of interaction [24]. It can also enable users to interact with public displays and kiosks without contact [18], which is of a particular interest in global spread of infectious diseases, such as the current COVID-19 situation. Most importantly, it could enable people with speech disorder, muteness, and blindness to input text and interact with various computer systems, increasing their access to these technologies.

This article starts with a review of the existing work in the area. It then explains the proposed recognition model, followed by its evaluation and comparison with previous works. It then presents an analysis of variations in the geometry of articulation in each sub-regions of the vocal tract (i.e., pharyngeal, velar dorsal, hard palate, labial constriction region) with respect to different emotion and gender. Finally, it discusses the findings and limitations of the work and concludes with speculations on future extensions.

2 RELATED WORK

The feasibility of using rtMRI and deep learning to recognize speech is not well investigated in the literature. Saha et al. [20] classified vowel-consonant-vowel (VCV) combinations using the same dataset used in this work with an accuracy rate of 42%. Leeuwen et al. [23] classified 27 sustained phonemes from MR images using a convolutional neural network with an accuracy rate of 57%.

Some have also used rtMRI to study articulatory characteristics of emotional speech using vocal tract movement data [11]. Lee et al. [13] analyzed the rtMRI data of emotional speech of one male speaker. They found out that "angry" speech can be characterized by much wider and faster vocal tract shaping and the extra usage of the pharyngeal region than the other examined emotions (neutral, happy, and sad). They also reported that "happy" speech exhibited shorter vocal tract length than the other emotions. Their findings were, however, obtained from a limited dataset collected from only one male speaker. A different work [14] reported the differences in vocal tract behaviors, and between inter-speaker and intra-speaker in different speech production styles, such as different emotion expression. Kim et al. [11] found out that the pharyngeal constriction and releasing are more emphasized for "angry" than "happy", while the palatal constriction and releasing are more emphasized for "happy" than "angry" during the production of one word "five".

[2]Articulatory speech recognition identifies a sequence of characters based on the corresponding sequence of vocal tract shapes.

[3]USC-EMO-MRI: An Emotional Speech Production Database, https://sail.usc.edu/span/usc-emo-mri

3 RECOGNITION MODEL

The aim of our recognition model is to predict the phrase being spoken from a silent video of vocal tract movements during speech production. It uses the LipNet model [4] that has been used in the past to generate text conditioned on lip sequences [17]. However, the decoder was conditioned on vocal tract movement sequences as illustrated in Fig. 1. The proposed recognition model consists of two sub-modules (or sub-networks): a *feature extraction* frontend that takes a sequence of video frames and outputs one feature vector per frame, and a *sequence modeling* module that inputs the sequence of per-frame feature vectors and outputs a sentence character by character. We describe these modules below.

3.1 Feature Extraction

The recognition model takes a sequence of T frames as input to process by 3 layers of spatiotemporal convolutions (STCNN) [9]. It consists of a convolutional layer with 64 3-dimensional (3D) kernels of $5 \times 7 \times 7$ size (time $\times$ width $\times$ height), followed by Batch Normalization (BN) [8] and Rectified Linear Units (ReLU) [2]. Each extracted feature map is passed through a spatiotemporal max-pooling layer, which drops the spatial size of the 3D feature.

3.2 Sequence Modeling

The extracted features are processed by 2-Bidirectional Gated Recurrent Units (Bi-GRUs) [5], where each time-step of the GRU output is processed by a linear layer, followed by a softmax layer over the vocabulary. Then, an end-to-end model is trained with connectionist temporal classification (CTC) loss [7]. Next, the softmax output is decoded with a left-to-right beam search [6] that incorporates prior information from an external language model [25] to recognize the spoken utterances. The model is capable of mapping variable-length video sequences to text sequences. All layers use rectified linear unit (ReLU) activation functions [2]. During inference, we use a 5-gram character-level Language Model (LM), which is a recurrent network with 4 unidirectional layers of 1,024 LSTM cells each. The LM is trained to predict one character at a time.

4 EXPERIMENT

This section describes the dataset preparation, the experiments conducted for parameter selection of the model, and the training protocol used to build the proposed model. We do parameter selection for batch size, number of epochs, and beam width (K). We then compare the performance of proposed articulatory speech recognition with several existing deep learning models to demonstrate that our model performs much better than those.

4.1 Dataset Preparation

To validate the performance of the proposed model, we performed an articulatory speech recognition experiment on the USC-TIMIT dataset[1], which includes 2D rtMRI of vocal tract shaping of ten speakers (5 female and 5 male, M = 28.7 years, SD = 7.2) along with synchronized audio recordings and their time-aligned word-level transcriptions [16]. To prepare the labeled training data for our

model, an alignment between the word-level transcription and the videos frames is needed. Hence, we estimated the number of frames by multiplying duration (second) of each word by video frame rate (23.18 frames/second) and aligned with its word-level transcription. Once we had the labeled data, we divided the total available data into a "training dataset" with 3,680 videos of eight speakers and a "testing dataset" with the remaining 920 videos of two speakers.

4.2 Training

Before feeding the data to the model, we augmented the training dataset by applying a horizontally mirrored transformation on video frames. In total, there were 10,972 samples. We augmented the dataset with simple transformations to reduce overfitting. We trained the model on both regular and horizontally mirrored image sequences. In addition, we varied the parameters of the model one by one keeping all others fixed, and simultaneously conducted evaluations for various combinations to select the set of optimum values. The batch size varied from 16 to 256, then set to the optimum value of 32. Similarly, the number of steps/epoch for the training was changed from 50 to 500 and fixed at 100 since it yielded a better accuracy rate. Table 1 summarizes the hyperparameters of the recognition model, where T denotes the number of frames, H and W denote the height and width, respectively, C denotes channels, F denotes the feature dimension, and V denotes the number of characters in the vocabulary.

Layer	*Dimension*	*Order*
InputLayer	75 x 64 x 64 x 1	T x W x H x C
ZeroPadding3D	77 x 68 x 68 x 1	T x W x H x C
Conv3D	75 x 32 x 32 x 32	T x W x H x C
BatchNorm	75 x 32 x 32 x 32	T x W x H x C
Activation	75 x 32 x 32 x 32	T x W x H x C
Dropout	75 x 32 x 32 x 32	T x W x H x C
MaxPool3D	75 x 16 x 16 x 32	T x W x H x C
ZeroPadding3D	75 x 20 x 20 x 32	T x W x H x C
Conv3D	75 x 16 x 16 x 64	T x W x H x C
BatchNorm	75 x 16 x 16 x 64	T x W x H x C
Activation	75 x 16 x 16 x 64	T x W x H x C
Dropout	75 x 16 x 16 x 64	T x W x H x C
MaxPool3D	75 x 8 x 8 x 64	T x W x H x C
ZeroPadding3D	75 x 10 x 10 x 64	T x W x H x C
Conv3D	75 x 8 x 8 x 96	T x W x H x C
BatchNorm	75 x 8 x 8 x 96	T x W x H x C
Activation	75 x 8 x 8 x 96	T x W x H x C
Dropout	75 x 8 x 8 x 96	T x W x H x C
MaxPool3D	75 x 4 x 4 x 96	T x W x H x C
Bi-GRU	75 x 512	T x F
Bi-GRU	75 x 512	T x F
Linear	75 x 28	T x V
Softmax	75 x 28	T x V

Table 1: Recognition model architecture hyperparameters.

The number of frames was fixed to 75, ~3 seconds. Longer image sequences were truncated and shorter sequences were padded with zeros. We batch-normalized the outputs of each convolution layer. All layers used rectified linear unit (ReLU) activation functions. We applied a dropout [22] of 0.5 after each max-pooling layer. The model was trained end-to-end by the Adam optimizer [12] with a batch size of 32. The learning rate was set to 10^{-3}. The Phoneme Error Rate (PER), Character Error Rate (CER), and Word Error Rate (WER) were computed using the CTC beam search (see Section 5.2). On top of that, we used a character 5-gram binarized language model. The above-described network model was implemented with the Keras deep-learning platform with Tensorflow [1] as the back-end, and an NVIDIA GeForce 1080Ti as the GPU board. Training network with 10,972 samples required approximately 3.5 hours.

5 RESULTS

We evaluated the proposed architecture and training strategies. We also compared the model with previous work on articulatory speech recognition [20, 23] that considered only the simpler case of predicting vowel-consonant vowel (VCV) combinations and phoneme from static MR images using a deep neural network. Note that our model, in contrast, predicts sequences, thus can exploit temporal context to attain higher accuracy. The inference and evaluation procedures used in this work are described below.

5.1 Beam Search

As discussed earlier, our architecture performs character-level prediction on input frames by performing CTC beam search of width 4. At each timestep, the hypotheses in the beam are expanded with every possible character, and only the 4 most probable hypotheses are stored. Fig. 2 illustrates the effect of increasing the beam width, where one can see that there is no observed benefit for increasing the width beyond 4.

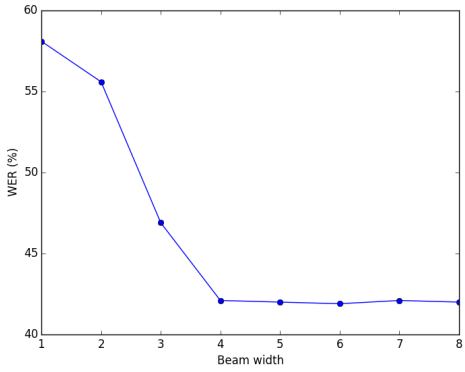

Figure 2: The effect of beam width on Word Error Rate (WER).

5.2 Evaluation Protocol

To measure the performance of the proposed model, we computed the Phoneme Error Rate (PER), Character Error Rate (CER), and Word Error Rate (WER), which are standard metrics for the performance of Automatic Speech Recognition (ASR) models. A predicted

Dictionary	Dataset	PER %	CER %	WER %
Vowel-Consonant-Vowel [20]	Vocal Tract Morphology MRI	58.0	-	-
Phoneme [23]	Vocal Tract Morphology MRI	57.0	-	-
Phrases without LM	**USC-TIMIT**	**44.1**	**41.7**	**45.4**
Phrases with LM	**USC-TIMIT**	**40.6**	**39.4**	**42.1**

Table 2: Performance of the three examined speech recognition models exploiting vocal tract dynamics on unseen data. The last two rows present the performance of the model proposed in this paper. Note that for a fair comparison between the models, we converted the accuracy reported in the respective papers to PER.

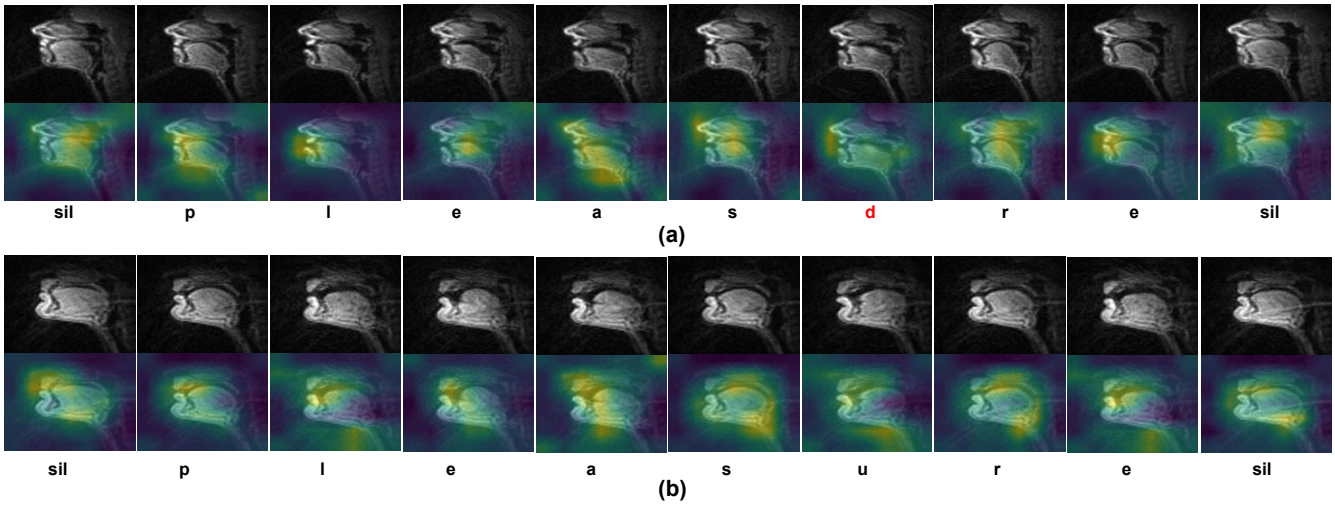

Figure 3: Saliency maps for the word "pleasure": female (a) and male (b) speakers with their corresponding phoneme predictions at the bottom. Red labels indicate incorrect predictions. Yellow shades indicate high sensitivity, that is, small changes in these pixels in the input have a large effect on the predicted class.

word is considered correct when each character of the word is correct. PER is measured by dividing the total number of phoneme errors (the minimum number of phoneme insertions, substitutions, and deletions required to transform the predicted phrase into the ground truth) divided by the total number of phonemes. CER and WER are calculated using the same approach with phoneme replaced with character and word, respectively. Table 2 summarizes the overall PER, CER, and WER on the unseen test data. The mean PER for the proposed method is 40.6%, which is much lower than the existing models. We also conducted an ablation study to analyze the effect of the language model on the overall performance gain. Results revealed that the model with LM exhibited 7.9% reduction in PER, 5.5% reduction in CER, and 7.2% reduction in WER. It demonstrates the feasibility of the model in recognizing phrases from MRI data. The results are particularly encouraging since they suggest that even greater performance can be attained with continued exploration of this interesting and novel problem space. Besides, based on the literature, increasing the amount of training data can significantly improve recognition performance.

5.3 Saliency

We applied saliency visualisation techniques [21] to interpret our model's learned behaviour, showing that it attends to important articulatory regions in the videos. Fig. 3 illustrates analysis of two saliency visualisations for the word "pleasure" for female and male speakers. Notice that the regions where changes in the input have the most impact on the prediction light up. The saliency maps show that the model has learned to focus on the parts of the input frames that represent the crucial articulatory positions needed to distinguish between different phonemes. Most phonemes show a more widespread field between the tongue and palate. As can be seen in Fig. 3, the saliency maps are not similar between the two subjects since vocal tract configurations varies from person to person. Notice that, in Fig. 3 (a), the model incorrectly predicted the phoneme *d* instead of *u* (highlighted in red). This mistake was made when the saliency maps showed places of attention that were not considered to be important for classification.

6 EMOTION AND GENDER ANALYSIS

This section presents the results of an analysis of MR images for emotion-dependent vocal tract movements to relate different emotions, particularly happy, angry, and sad, with a neutral emotion. Following the methodology used in a previous work [10], we first extract the vocal tract airway-tissue boundaries (red line for lower boundary and green line for upper boundary in Fig. 4), then divide

them into four sub-regions: (1) grid lines 1—17 for pharyngeal region, (2) grid lines 18—68 for velar and dorsal constriction region, (3) grid lines 69—79 (alveolar ridge landmark) for the hard palate region, and (4) grid lines 80—86 for labial constriction region.

Then, we compare vocal tract shaping of different emotions by measuring the distortion in the shaping of each sub-region for each emotion (e) relative to neutral emotion (n). This is done by the normalized sum of differences of the cross-distances in the 2D space from the centroid region (mean of all the points on vocal tract airway-tissue boundaries) to each respective landmark (number of points on vocal tract airway-tissue boundaries). The cross-distances are individually computed for lower and upper boundary of each sub-region. To measure this, we developed a new metric, Neutral Emotion Deviation Measure (NEDM), defined as follows.

$$\text{NEDM}_r^b = \sum_l \frac{|d_{n_l} - d_{e_l}|}{d_{n_l}} \tag{1}$$

Where,

r: number of sub-regions in the vocal tract (i.e., 4),

b: lower and upper boundaries,

l: number of landmarks in each sub-region,

d_{n_l}: Euclidean distance between centroid and the landmarks in the particular sub-region for neutral emotion (n),

$$\begin{aligned} d_{n_l} &= \textit{Euclidean-Distance}(p_{centroid}, p_{n_l}) \\ &= \sqrt{\sum_{l=i}^{n} (P_{centroid} - P_{nl})^2} \end{aligned} \tag{2}$$

d_{e_l}: Euclidean distance between centroid and the landmarks in the particular sub-region for different emotions (e) (i.e., happy, angry, sad),

$$\begin{aligned} d_{e_l} &= \text{Euclidean-Distance}(p_{centroid}, p_{e_l}) \\ &= \sqrt{\sum_{l=i}^{n} (P_{centroid} - P_{el})^2} \end{aligned} \tag{3}$$

where,

$P_{centroid}$: x, y coordinate of centroid location of sub-region,

P_{n_l}: x, y coordinate of landmarks location for neutral emotion,

P_{e_l}: x, y coordinate of landmarks location for different emotion.

In order to calculate cross-distance for each sub-region, semi-automatic tissue-airway boundary segmentation is performed using a recently introduced MATLAB software [10]. This software performs (i) tracking of the lips and the larynx, (ii) segmentation of the airway tissue boundary, (iii) pixel sensitivity correction, (iv) noise suppression on the MR image, and (v) computation of the distance function. The processes (i) and (ii) are performed automatically based on the semi-automatically constructed gridlines. This work examined the vocal tract data for the words “clock” and “dock”. A total of 56 productions of each word as a function of emotion spoken by ten speakers (5 male and 5 female) were analyzed (56 productions × 2 words × 3 emotions).

Table 3 presents the identification of the most affected regions in vocal tract airway-tissue upper and lower boundaries for each emotion. The reported values are averages of the distortion measure. On

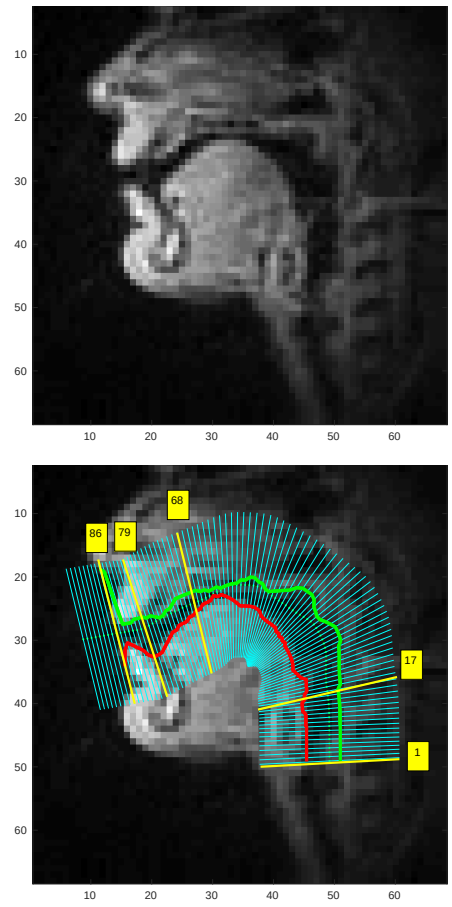

Figure 4: Frame of neutral emotion rtMRI video showing a reference image (top) and corresponding segmentation of lower and upper boundary of vocal tract (bottom).

average, the sub-regions of lower boundary showed a greater deviation from centroid location than upper boundary regions for all emotions. The velar and dorsal constriction region and hard palate region showed more distortion for high arousal emotions (anger and happyness) than low arousal emotion (sadness). In general, the velar and dorsal constriction region was of great importance for all emotions. For low arousal emotions, all regions tended to have less noticeable changes compared to high arousal emotions. The palatal constriction and releasing were more emphasized for happiness than for anger. Results showed that the distortion factor was also affected by gender. For all emotions, female speakers had more noticeable changes in all regions. However, labial constriction region showed very less variation across gender. For anger, female speakers had more geometrical distortion in pharyngeal and velar and dorsal constriction regions than happiness.

7 DISCUSSION

In this work, we demonstrated that end-to-end deep learning framework can automatically map sequences of vocal tract shaping to entire sentences with a phoneme error rate of 40.6%. The proposed

Lower boundary geometrical comparison of each sub-region in the vocal tract with respect to different emotions

Clock (Male)				
Regions/ Emotions	**Pharyngeal**	**Velar and dorsal constriction**	**Hard palate**	**Labial constriction**
Happy	0.67	0.78	0.84	0.62
Angry	0.85	0.91	0.72	0.73
Sad	0.36	0.48	0.41	0.50
Clock (Female)				
Regions/ Emotions	**Pharyngeal**	**Velar and dorsal constriction**	**Hard palate**	**Labial constriction**
Happy	0.71	0.80	0.93	0.64
Angry	0.89	1.00	0.86	0.74
Sad	0.41	0.54	0.49	0.53
Dock (Male)				
Regions/ Emotions	**Pharyngeal**	**Velar and dorsal constriction**	**Hard palate**	**Labial constriction**
Happy	0.68	0.74	0.83	0.62
Angry	0.83	0.94	0.70	0.72
Sad	0.32	0.43	0.38	0.53
Dock (Female)				
Regions/ Emotions	**Pharyngeal**	**Velar and dorsal constriction**	**Hard palate**	**Labial constriction**
Happy	0.75	0.80	0.94	0.61
Angry	0.91	0.98	0.87	0.69
Sad	0.43	0.50	0.49	0.48

Upper boundary geometrical comparison of each sub-region in the vocal tract with respect to different emotions

Clock (Male)				
Regions/ Emotions	**Pharyngeal**	**Velar and dorsal constriction**	**Hard palate**	**Labial constriction**
Happy	0.42	0.56	0.34	0.48
Angry	0.37	0.44	0.48	0.45
Sad	0.30	0.41	0.33	0.35
Clock (Female)				
Regions/ Emotions	**Pharyngeal**	**Velar and dorsal constriction**	**Hard palate**	**Labial constriction**
Happy	0.44	0.48	0.53	0.49
Angry	0.61	0.54	0.47	0.41
Sad	0.33	0.42	0.49	0.39
Dock (Male)				
Regions/ Emotions	**Pharyngeal**	**Velar and dorsal constriction**	**Hard palate**	**Labial constriction**
Happy	0.40	0.43	0.39	0.45
Angry	0.37	0.54	0.59	0.45
Sad	0.31	0.44	0.32	0.28
Dock (Female)				
Regions/ Emotions	**Pharyngeal**	**Velar and dorsal constriction**	**Hard palate**	**Labial constriction**
Happy	0.48	0.49	0.41	0.38
Angry	0.57	0.52	0.44	0.39
Sad	0.36	0.48	0.40	0.34

Table 3: Average Neutral Emotion Deviation Measure (NEDM) indicating the relation between each subregion and each emotion across gender. The displacements are calculated using centroid position for subregion POI.

model performed much better than the existing models [20, 23] that either consider only a simpler case of predicting vowel-consonant-vowel (VCV) combinations with an error rate of 58% or phonemes with an error rate of 57% rather than phrases. The findings suggest that deep learning represents a viable tool for continuous speech recognition from rtMRI. Most importantly, our proposed model does not rely on hand-engineered spatiotemporal visual features or a separately-trained sequence model. The proposed end-to-end model also eliminates the need for segmenting videos into words before predicting a sentence. Furthermore, saliency visualisations revealed that the proposed model learns to attend phonologically important regions of the vocal tract. It provides an insight into the vocal tract regions that are most important for phoneme classification. Further analysis revealed that mistakes were more frequent when the saliency maps showed places of sensitivity that were not expected to be important for classification.

Analysis of the MR images for emotion-dependent vocal tract movements to relate three different emotions, namely happy, angry, sad, with a neutral emotion provided interesting insights into the most affected regions of vocal tract during emotional speech production. We found out that the sub-regions of vocal tract lower boundary tended to have a more noticeable change than upper boundary regions for all emotions. It also showed that variation in each sub-region was affected by gender variability. Female speakers had more geometrical distortion in pharyngeal and velar and dorsal constriction regions for negative emotions (anger) than positive emotions (happiness). Overall, for all emotions, female speaker had more noticeable changes in all regions.

The proposed rtMRI-based speech recognition system could potentially be used as medium for input and interaction with various computer systems, incorporated in day-to-day usage. This approach could also enable people with speech disorder, muteness, and blindness to input and interact with computer systems, increasing their access to these technologies. Although we do not have the technology to achieve these just yet, the findings of this work shows its potential. Furthermore, one could apply the proposed methodology to the data from people with speech disorder and compare it to the speech of people without speech disorder to find out which articulators are involved in the impairment of phoneme production. The acoustic information extracted from vocal tract's dynamics might reveal how different phonemes production mechanism are related to each other. Also, an analysis of emotion-dependent vocal tract geometry of people with speech disorder could provide new insights into variations in emotions of their speech.

8 CONCLUSION

We proposed a deep learning framework that can decode text using the cues provided by the movements of the vocal tract. On the USC-TIMIT corpus, the proposed model achieves a 40.6% PER at sentence-level, which is much lower than the existing models. Literature on deep speech recognition suggests that this performance is likely to further improve with addtional data [3].Furthermore, we conducted an analysis of variations in the geometry of articulation in each sub-regions of the vocal tract with respect to different emotions and genders. Results revealed that each sub-regions distortion was affected by both gender and emotion. In the future, we will extend this work to people with various speech disorders. We will also explore different learning models and compare their performance in the defined context.

REFERENCES

[1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A System for Large-Scale Machine Learning. 265–283. https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

[2] Abien Fred Agarap. 2019. Deep Learning using Rectified Linear Units (ReLU). (2019). http://arxiv.org/abs/1803.08375

[3] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* (New York, NY, USA, 2016-06-19) *(ICML'16)*. JMLR.org, 173–182. http://arxiv.org/abs/1512.02595

[4] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. 2016. LipNet: End-to-End Sentence-level Lipreading. (2016). https://arxiv.org/abs/1611.01599v2

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. (2014). http://arxiv.org/abs/1412.3555

[6] Ronan Collobert, Awni Hannun, and Gabriel Synnaeve. 2019. A Fully Differentiable Beam Search Decoder. (2019). arXiv:1902.06022 http://arxiv.org/abs/1902.06022

[7] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd international conference on Machine learning* (Pittsburgh, Pennsylvania, USA, 2006-06-25) *(ICML '06)*. Association for Computing Machinery, 369–376. https://doi.org/10.1145/1143844.1143891

[8] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37* (Lille, France, 2015-07-06) *(ICML'15)*. JMLR.org, 448–456.

[9] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. 3D Convolutional Neural Networks for Human Action Recognition. 35, 1 (2013), 221–231. https://doi.org/10.1109/TPAMI.2012.59 Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[10] Jangwon Kim, Naveen Kumar, Sungbok Lee, and Shrikanth S. Narayanan. 2014. Enhanced Airway-tissue Boundary Segmentation for Real-time Magnetic Resonance Imaging Data.

[11] Jangwon Kim, Asterios Toutios, Yoon-Chul Kim, Yinghua Zhu, Sungbok Lee, and Shrikanth Narayanan. 2014. USC-EMO-MRI corpus: An Emotional Speech Production Database Recorded by Real-time Magnetic Resonance Imaging.

[12] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. (2017). arXiv:1412.6980 http://arxiv.org/abs/1412.6980

[13] Sungbok Lee, Erik Bresch, Jason Adams, Abe Kazemzadeh, and Shrikanth Narayanan. 2006. A Study of Emotional Speech Articulation Using a Fast Magnetic Resonance Imaging Technique, Vol. 5.

[14] Sungbok Lee and Shrikanth S. Narayanan. 2010. Vocal Tract Contour Analysis of Emotional Speech by the Functional Data Curve Representation. In *INTERSPEECH* (2010).

[15] Vikramjit Mitra, Ganesh Sivaraman, Chris Bartels, Hosung Nam, Wen Wang, Carol Espy-Wilson, Dimitra Vergyri, and Horacio Franco. 2017. Joint Modeling of Articulatory and Acoustic Spaces for Continuous Speech Recognition Tasks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017-03). 5205–5209. https://doi.org/10.1109/ICASSP.2017.7953149 ISSN: 2379-190X.

[16] Shrikanth Narayanan, Asterios Toutios, Vikram Ramanarayanan, Adam Lammert, Jangwon Kim, Sungbok Lee, Krishna Nayak, Yoon-Chul Kim, Yinghua Zhu, Louis Goldstein, Dani Byrd, Erik Bresch, Prasanta Ghosh, Athanasios Katsamanis, and Michael Proctor. 2014. Real-time Magnetic Resonance Imaging and Electromagnetic Articulography Database for Speech Production Research (TC). 136 (2014), 1307. https://doi.org/10.1121/1.4890284

[17] Laxmi Pandey and Ahmed Sabbir Arif. 2021. LipType: A Silent Speech Recognizer Augmented with an Independent Repair Model. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 1, 19 pages. https://doi.org/10.1145/3411764.3445565

[18] Laxmi Pandey, Khalad Hasan, and Ahmed Sabbir Arif. 2021. Acceptability of Speech and Silent Speech Input Methods in Private and Public. In *Proceedings*

of the 2021 CHI Conference on Human Factors in Computing Systems (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 251, 13 pages. https://doi.org/10.1145/3411764.3445430

[19] Veena S., Nilashree Wankhede, and Milind Shah. 2016. Study of Vocal Tract Shape Estimation Techniques for Children. 79 (2016), 270–277. https://doi.org/10.1016/j.procs.2016.03.035

[20] Pramit Saha, Praneeth Srungarapu, and Sidney Fels. 2018. Towards Automatic Speech Identification from Vocal Tract Shape Dynamics in Real-time MRI. 1249–1253. https://doi.org/10.21437/Interspeech.2018-2537

[21] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. (2014).

[22] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. 15, 1 (2014), 1929–1958.

[23] Kicky van Leeuwen, P. Bos, S. Trebeschi, Maarten Alphen, Luuk Voskuilen, L.E. Smeele, F. Van der Heijden, and Rob van Son. 2019. CNN-Based Phoneme Classifier from Vocal Tract MRI Learns Embedding Consistent with Articulatory Topology. 909–913. https://doi.org/10.21437/Interspeech.2019-1173

[24] Daniel Wigdor and Dennis Wixon. 2011. *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture* (1st ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[25] Ian Williams, Anjuli Kannan, Petar Aleksic, David Rybach, and Tara Sainath. 2018. Contextual Speech Recognition in End-to-end Neural Network Systems Using Beam Search. 2227–2231. https://doi.org/10.21437/Interspeech.2018-2416