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

Laxmi Pandey
University of California, Merced
Merced, California, United States
lpandey@ucmerced.edu

Ahmed Sabbir Arif
University of California, Merced
Merced, California, United States
asarif@ucmerced.edu

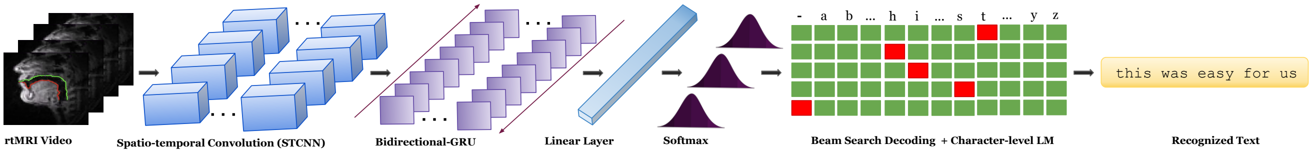
rtMRI Video
Spatio-temporal Convolution (STCNN)
Bidirectional-GRU
Linear Layer
Softmax
- a b ... h i ... s t ... y z
Beam Search Decoding + Character-level LM
this was easy for us
Recognized Text

Figure 1: The model classifies 2D real-time MRI (rtMRI) of vocal tract shaping into text with an end-to-end deep neural network. A sequence of input frames is processed by 3 layers of STCNN to extract spatiotemporal features, which are processed by 2 Bi-GRUs, and a linear and a softmax layer. Then, the output is decoded with prefix beam search with a language model.

ABSTRACT

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. In an experiment, it achieved a 40.6% PER at sentence-level, much better compared to the existing models. We also performed 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 suggest that each sub-regions distortion is affected by both emotion and gender.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility technologies**.

KEYWORDS

MRI, speech, silent speech, neural network, vocal tract, accessibility

ACM Reference Format:
Laxmi Pandey and Ahmed Sabbir Arif. 2021. Silent Speech and Emotion Recognition from Vocal Tract Shape Dynamics in Real-Time MRI. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters (SIGGRAPH '21 Posters), August 09-13, 2021.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3450618.3469176

1 INTRODUCTION

Speech sounds of spoken language are obtained by varying configuration of the vocal tract articulators. They contain abundant information that can be used to better understand the underlying mechanism of speech production. We propose a model that maps a variable-length sequence of rtMRI video frames to text using spatiotemporal convolutions, a recurrent network, and the connectionist temporal classification loss, which we believe is the first end-to-end sentence-level articulatory speech recognition model[1]. We also performed an analysis on the MR images of emotion-dependent vocal tract movements to compare different emotions and genders using an existing dataset [Kim and et al. 2014]. An understanding of how emotion affects articulatory movements during speech production can reduce emotional ambiguity in recognized sentences (e.g., "I hate you" could be said said either sarcastically or literally). The effects of gender on vocal tract movements, in contrast, can increase the accuracy of the recognition system.

2 RECOGNITION MODEL

The aim of the proposed model is to predict the spoken phrases from silent videos of vocal tract movements during speech production (Fig. 1). It consists of 2 sub-modules: a *feature extraction* frontend that takes a sequence of video frames to output one feature vector per frame using 3 layers of spatiotemporal convolutions (STCNN), and a *sequence modeling* module that inputs the sequence of per-frame feature vectors to process them by 2-Bidirectional Gated Recurrent Units (Bi-GRUs), where each time-step of the 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 and the softmax output is decoded with a left-to-right beam search using Stanford-CTC's decoder and 5-gram character language model to recognize the spoken phrases.

3 PERFORMANCE EVALUATION

To validate the model's performance, we performed an articulatory speech recognition experiment on the Narayanan and et al. [2014] dataset that includes 2D rtMRI of vocal tract shaping of 10 speakers and their time-aligned word-level transcriptions. We divided the data into a *training* set with 3,680 videos of 8 speakers and a *testing*

SIGGRAPH '21 Posters, August 09-13, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8371-4/21/08.
https://doi.org/10.1145/3450618.3469176

[1] An extended version of this work is available here: http://arxiv.org/abs/2106.08706

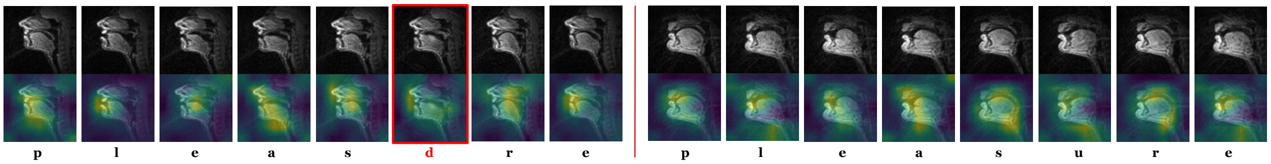

Figure 2: Saliency maps for the word "pleasure": female (left) and male (right) speakers with corresponding phoneme predictions at the bottom. Red labels indicate incorrect predictions. Yellow shades indicate high sensitivity, that is, small changes in these pixels in the input have a large effect on the predicted class.

Table 1: Average Neutral Emotion Deviation Measure (NEDM) indicating the relationship between each subregion and emotion across gender (M, F) for lower and upper boundaries (respectively) of vocal tract.

Word	Emotion	Pharyngeal	Velar & dorsal constriction	Hard palate	Labial constriction
Clock	**Happy**	M: 0.67, 0.42 | F: 0.71, 0.44	M: 0.78, 0.56 | F: 0.80, 0.48	M: 0.84, 0.34 | F: 0.93, 0.53	M: 0.62, 0.48 | F: 0.64, 0.49
	Angry	M: 0.85, 0.37 | F: 0.89, 0.61	M: 0.91, 0.44 | F: 1.00, 0.54	M: 0.72, 0.48 | F: 0.86, 0.47	M: 0.73, 0.45 | F: 0.74, 0.41
	Sad	M: 0.36, 0.30 | F: 0.41, 0.33	M: 0.48, 0.41 | F: 0.54, 0.42	M: 0.41, 0.33 | F: 0.49, 0.49	M: 0.50, 0.35 | F: 0.53, 0.39
Dock	**Happy**	M: 0.68, 0.40 | F: 0.75, 0.48	M: 0.74, 0.43 | F: 0.80, 0.49	M: 0.83, 0.39 | F: 0.94, 0.41	M: 0.62, 0.45 | F: 0.61, 0.38
	Angry	M: 0.83, 0.37 | F: 0.91, 0.57	M: 0.94, 0.54 | F: 0.98, 0.52	M: 0.70, 0.59 | F: 0.87, 0.44	M: 0.72, 0.45 | F: 0.69, 0.39
	Sad	M: 0.32, 0.31 | F: 0.43, 0.36	M: 0.43, 0.44 | F: 0.50, 0.48	M: 0.38, 0.32 | F: 0.49, 0.40	M: 0.53, 0.28 | F: 0.48, 0.34

set with the remaining 920 videos of 2 speakers. For training, we augmented the data by applying a horizontally mirrored transformation on the video frames, resulting in 10,972 samples. The model was trained end-to-end by Adam optimizer with a batch size of 32.

Our model yielded 40.6% PER compared to 58% and 57% PER of existing models that consider only the simpler case of predicting vowel-consonant-vowel (VCV) [Saha and et al. 2018] and phoneme from static MR images using a deep neural network [van Leeuwen and et al. 2019], respectively. Fig. 2 illustrates two saliency visualisations for the word "pleasure" for female and male speakers. 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. The saliency maps for female and male speakers are different as the vocal tract configurations varies from person to person. The phoneme *u* was incorrectly predicted as *d* (highlighted in red) when the model payed attention to parts that are not crucial in distinguish between the two (see the saliency maps).

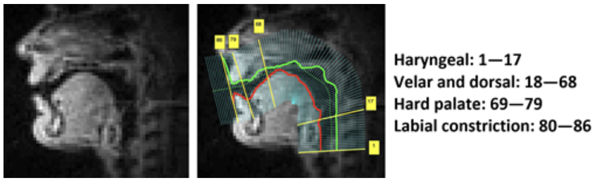

Figure 3: Segmentation of lower and upper boundary of vocal tract in four sub-regions.

4 EMOTION AND GENDER ANALYSIS

We compared vocal tract shaping of different emotions by measuring the distortion in the shaping of each sub-region (r) for each emotion (e) relative to neutral emotion (n) (Fig. 3). 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. For this, we developed a new Neutral Emotion Deviation Measure (NEDM): $\text{NEDM}_r^b = \sum_l \frac{|d_{n_l} - d_{e_l}|}{d_{n_l}}$, where b: lower and upper boundaries, l: number of landmarks in each sub-region, and d: Euclidean distance between centroid and the landmarks.

We examined 2 words (clock, dock) × 3 emotions (happy, angry, sad) × 56 productions from the Kim and et al. [2014] dataset. Table 1 shows the most affected regions in vocal tract airway-tissue upper and lower boundaries for all emotions. On average, the sub-regions of lower boundary showed a greater deviation from centroid location than upper boundary for all emotions. The velar & dorsal constriction and the hard palate regions showed more distortion for high arousal emotions (anger, happiness) than low arousal (sadness). The velar & dorsal constriction region was important for all emotions. The palatal constriction and releasing were more emphasized for happiness than anger. The distortion factor was affected by gender. For all emotions, female speakers had more noticeable changes in all regions. Labial constriction region showed a small variation across gender. For anger, female speakers had more geometrical distortion in pharyngeal and velar & dorsal constriction regions than happiness.

REFERENCES

Jangwon Kim and et al. 2014. USC-EMO-MRI corpus: An Emotional Speech Production Database Recorded by Real-time Magnetic Resonance Imaging.

Shrikanth Narayanan and et al. 2014. Real-time Magnetic Resonance Imaging and Electromagnetic Articulography Database for Speech Production Research (TC). 136 (2014), 1307.

Pramit Saha and et al. 2018. Towards Automatic Speech Identification from Vocal Tract Shape Dynamics in Real-time MRI. 1249–1253.

Kicky van Leeuwen and et al. 2019. CNN-Based Phoneme Classifier from Vocal Tract MRI Learns Embedding Consistent with Articulatory Topology. 909–913.