Interspeech 2022
18-22 September 2022, Incheon, Korea

Watch Me Speak: 2D Visualization of Human Mouth during Speech

C Siddarth[1], *Sathvik Udupa*[2], *Prasanta Kumar Ghosh*[2]

[1]Indian Institute of Information Technology, Design and Manufacturing, Kancheepuram, India
[2]Indian Institute of Science (IISc), Bangalore-560012, India

siddarthc2000@gmail.com, sathvikudupa66@gmail.com, prasantag@gmail.com

Abstract

We present a web interface to visualize the midsagittal plane of the human mouth during speech. Given an articulated sentence, we estimate the corresponding articulatory trajectories and visualize the same. This web interface provides a comprehensive view of the articulators' trajectories and could serve as an important tool for speech training.

Index Terms: Electromagnetic articulograph, Acoustic to articulatory inversion, Midsaggital plane visualization

1. Introduction

Production of a speech involves a complex and precise combination of multiple articulators in the vocal tract [1]. To understand this complex procedure better, it is crucial to understand the relationship between the articulators' movement and the sound produced as a result. The process of estimating these articulators' trajectories from the acoustic speech signal is known as Acoustic to Articulatory Inversion (AAI). The initial approaches involved code-book based procedures [2], followed by statistical modelling [3]. Later approaches adopted neural networks based modelling [4, 5, 6] and currently, the state-of-the-art results are achieved by Transformer-based model [7].

Applications : But more than just a theoretical problem that is to be solved, AAI has numerous practical applications in the real world. With the advent of the internet, the number of animated movies and games has been steadily increasing. Graphic animators could benefit greatly from an automated system that could animate the oral movements of animated characters based solely on input speech or text. Further, real-time AAI could assist individuals in speech training by providing them with visual feedback of the articulators. This provides more details on the trajectories of the articulators present deeper in the oral cavity, which otherwise is very difficult to visualize.

Furthermore, when compared to acoustic signals, certain articulatory movements have been shown to be speaker-invariant [8]. As a result, the inclusion of articulatory features is shown to aid ASR when compared to using only acoustic features [9, 10].

Past attempts : Given the multitude of applications, there have been past attempts to visualize the articulators from either the acoustic or articulatory data. Though there is significant literature that models the 3D tongue dynamically [11, 12, 13, 14], there are very few works that reconstructs the complete midsagittal plane, that animates more than just the tongue [15, 16]. In fact, a simple 2D model that showcases both the tongue and lips is sufficient, many times better than a complex 3D model, to understand the data at hand better [15]. Such a simple animation could be found on Google [1] but is very naive and restricted to a single word.

Contributions : In this paper, we present a web interface that recreates the complete 2D midsagittal plane with the ability to visualize sentences of any length. We recreate the main articulators, namely the uvula, tongue, teeth, lips and jaw. We employ an AAI model to estimate the articulatory trajectories that produce a given acoustic signal. Further, the estimated trajectories are utilized to recreate the complete midsagittal plane of the human mouth. The uvular positions are approximated based on the nasality of the articulated phoneme. Unlike previous attempts, we reconstruct the complete midsagittal plane of the human mouth that works on sentences of any length.

[1]https://www.google.com/search?q=acoustics+pronunciation

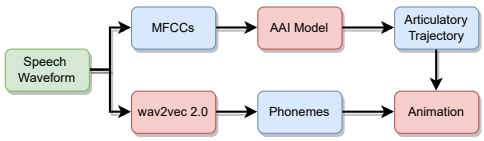

Figure 1: *Overview of the web interface*

2. Dataset

A set of 460 phonetically balanced English sentences are considered from the MOCHA-TIMIT corpus as the stimuli for data collection from 38 subjects. Six sensors were glued on different articulators, viz. Upper Lip (UL), Lower Lip (LP), Jaw, Tongue Tip (TT), Tongue Body (TB), and Tongue Dorsum (TD). We simultaneously recorded audio signals using a microphone and articulatory movement data using Electromagnetic Articulograph (EMA) AG501[17]. In this work, only the movements in the midsagittal plane are considered, corresponding to horizontal and vertical directions. Thus, we have twelve articulatory trajectories denoted by ULx, ULy, LLx, LLy, Jawx, Jawy, TTx, TTy, TBx, TBy, TDx, TDy. More details on the dataset can be found in [7].

3. Overview

The proposed framework consists of 2 parts, estimation of articulatory trajectories from the acoustic signal, visualizing the estimated trajectories (Fig 1).

AAI model : The input acoustic signal is represented as 13-dimensional MFCCs features. These are in turn processed by a neural network-based AAI model to estimate the articulatory movements.

wav2vec 2.0 : The predicted articulators give control over regions of tongue, lips and Jaw. To estimate the movement of uvula, we use a phoneme estimator to know if the sound in a frame is nasal or non-nasal. We employ a wav2vec 2.0 Large model [18] pre-trained on LibriVox (LV-60k) and fine-tuned on CommonVoice to recognize phonetic labels. With this information, we control the position of the uvula.

Visualization using Bezier Curves : The estimated points are then utilized to construct multiple bezier curves, which

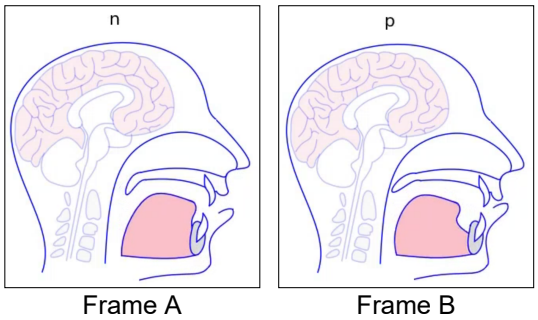

Figure 2: *A sample output snippet. The estimated phoneme is printed above the animation. Note that in Frame A, where the articulated phoneme is nasal (n), the uvula is open. In Frame B, note that the lips touch each other to articulate the phoneme p.*

make up the animation. A bezier curve is a parametric curve, defined by control points that yields a smooth and continuous curve. They are commonly used in computer graphics and related areas. The 12 estimated points (corresponding to 6 articulators per frame), acting as the control points of the bezier curves, determine the position and structure of the articulators. We have tested the visualization on multiple subjects with several sentences and observed that it has generalized well (Fig 2).

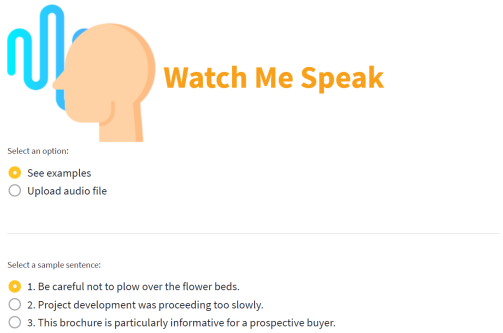

Figure 3: *Sample snippet of the web page*

4. Functionality

The user is given 2 options (Fig 3), either to play an existing example or upload an audio file and visualize the same. If a file is uploaded, the audio is resampled to 16kHz before extracting the MFCCs. The MFCCs are utilized by a pre-trained AAI model to estimate the articulatory trajectories at 100Hz, which is downsampled for the animation. Also, the corresponding phonemes are estimated by a pre-trained wav2vec model from the raw waveform and are downsampled to align with articulatory trajectories. Finally, the trajectories are visualized at 24 FPS. We use Librosa for preprocessing, PyTorch for model inference and StreamLit for frontend. More resources on the project can be found at https://spire.ee.iisc.ac.in/spire/aaimpvis.php.

5. Discussion

The web interface allows users to visualize the midsaggital plane for any articulated sentence. We employ an AAI model to estimate the articulatory trajectories from the acoustic space and visualize the estimated values with beizer curves, as an animation. In the future, we plan to improve the anatomical correctness of the articulators and their movements.

6. References

[1] F. H. Guenther and G. Hickok, "Role of the auditory system in speech production." *Handbook of clinical neurology*, vol. 129, pp. 161–75, 2015.

[2] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion." *The Journal of the Acoustical Society of America*, vol. 118 1, pp. 444–60, 2005.

[3] L. Zhang and S. Renals, "Acoustic-articulatory modeling with the trajectory hmm," *IEEE Signal Processing Letters*, vol. 15, pp. 245–248, 2008.

[4] A. Illa and P. K. Ghosh, "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory," in *INTERSPEECH*, 2018.

[5] N. Bozorg and M. T. Johnson, "Acoustic-to-articulatory inversion with deep autoregressive articulatory-wavenet," in *INTERSPEECH*, 2020.

[6] A. Illa and P. K. Ghosh, "Representation learning using convolution neural network for acoustic-to-articulatory inversion," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5931–5935.

[7] S. Udupa, A. Roy, A. Singh, A. Illa, and P. K. Ghosh, "Estimating articulatory movements in speech production with transformer networks," in *Interspeech*, 2021.

[8] O. Fujimura, "Relative invariance of articulatory movements: An iceberg model," in *Invariance and variability in speech processes*, J. S. Perkell and D. H. Klatt, Eds., 1986, pp. 226–242.

[9] P. K. Ghosh and S. S. Narayanan, "Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion." *The Journal of the Acoustical Society of America*, vol. 130 4, pp. EL251–7, 2011.

[10] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition." *The Journal of the Acoustical Society of America*, vol. 121 2, pp. 723–42, 2007.

[11] J. E. Lloyd, I. Stavness, and S. S. Fels, "Artisynth: A fast interactive biomechanical modeling toolkit combining multibody and finite element simulation," 2012.

[12] I. Stavness, J. E. Lloyd, and S. S. Fels, "Automatic prediction of tongue muscle activations using a finite element model." *Journal of biomechanics*, vol. 45 16, pp. 2841–8, 2012.

[13] K. Xu, Y. Yang, A. Jaumard-Hakoun, C. Leboullenger, G. Dreyfus, P. Roussel-Ragot, M. L. Stone, and B. Denby, "Development of a 3d tongue motion visualization platform based on ultrasound image sequences," *ArXiv*, vol. abs/1605.06106, 2015.

[14] W. F. Katz, T. F. Campbell, J. Wang, E. Farrar, J. C. Eubanks, A. Balasubramanian, B. Prabhakaran, and R. Rennaker, "Optispeech: a real-time, 3d visual feedback system for speech training," in *INTERSPEECH*, 2014.

[15] K. James and M. Wieling, "Read my points: Effect of animation type when speech-reading from ema data," in *SIGMORPHON*, 2016.

[16] A. Suemitsu, T. Ito, and M. K. Tiede, "An electromagnetic articulography-based articulatory feedback approach to facilitate second language speech production learning," *Journal of the Acoustical Society of America*, vol. 19, pp. 060 063–060 063, 2013.

[17] "3d electromagnetic articulograph," available online: http://www.articulograph.de/, last accessed: 4/2/2020.

[18] A. Baevski, H. Zhou, A. rahman Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *ArXiv*, vol. abs/2006.11477, 2020.