INTERSPEECH 2021: Show & Tell Contribution
30 August – 3 September, 2021, Brno, Czechia

Web Interface for estimating articulatory movements in speech production from acoustics and text

Sathvik Udupa, Anwesha Roy, Abhayjeet Singh, Aravind Illa, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India

Abstract

We release a web interface to visualise estimated articulatory movements in speech production from different modalities - acoustics and text. We allow the use of various trained models for this purpose. This tool also serves the purpose of comparing the predicted articulatory movements from different modalities and visually understanding the effect of noise in speech.

Index Terms: Electromagnetic articulograph, Acoustic to articulatory inversion, Phoneme to articulatory estimation, Visualisation

1. Introduction

Knowledge about the articulatory position along with the acoustics is useful in applications like automatic speech recognition [1, 2], language learning [3, 4] and speech synthesis [5, 6]. In practice, in the absence of direct articulatory movements, they are typically estimated from different modalities like acoustic features (Mel-Frequency Cepstral Coefficients (MFCC)) and text (phoneme sequence).

Gaussian Mixture Model (GMM) [7], Hidden Markov Model (HMM) [8], neural networks [9, 10] and bidirectional long short-term memory (BiLSTM) networks have been used for estimation of articulatory movements from acoustic features of speech, known as acoustic-to-articulatory inversion (AAI) [11, 12]. The state-of-art performance is achieved by transformer networks [13].

On the other hand, for phoneme to articulatory (PTA) mapping, we have deployed the Tacotron2 [14] and FastSpeech [15] used in our previous works [13, 16].

The webpage interface allows users to explore all the aforementioned approaches of estimation of articulatory movements. A user can provide both speech and text inputs and observe the corresponding articulatory estimations by selecting various models. Also, an user can compare the predictions across modalities for the same input(audio and corresponding text). The interface can be accessed at https://spire.ee.iisc.ac.in/spire/aaidemo.php.

2. Dataset

In order to build various AAI models, a set of 460 phonetically balanced English sentences are considered from the MOCHA-TIMIT corpus as the stimuli for data collection from 38 subjects with ages in the range of 20-28 years. For each sentence, we simultaneously recorded audio signals using a microphone [17] and articulatory movement data using Electromagnetic Articulograph (EMA) AG501 [18]. We consider twelve articulatory trajectories of EMA data, denoted by UL$_x$, UL$_y$, LL$_x$, LL$_y$, Jaw$_x$, Jaw$_y$, TT$_x$, TT$_y$, TB$_x$, TB$_y$, TD$_x$, TD$_y$. Further details about the data collection procedure can be found in [11].

3. Models

As described in our previous works [11, 19, 13, 16], we employ trained models for articulatory movement estimation. Different models are presented for AAI and PTA tasks. In this section, we list the model architectures involved in both tasks.

3.1. AAI

For the AAI task, Gaussian Mixture Model (GMM) [7], Deep Neural Net (DNN) [9, 10], Bidirectional Long Short Term Memory (LSTM) [11, 12], Convolutional Neural Network (CNN) [19] and Transformers [13] are present.

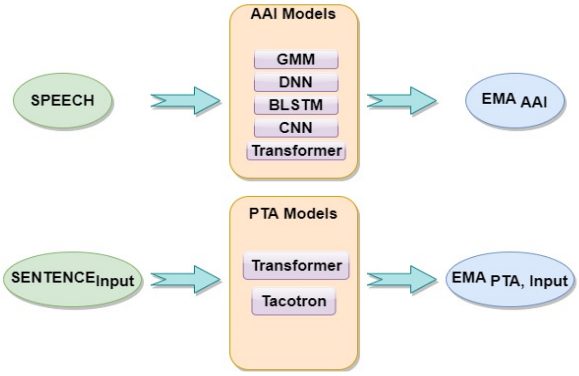


Figure 1: *Webpage workflow*

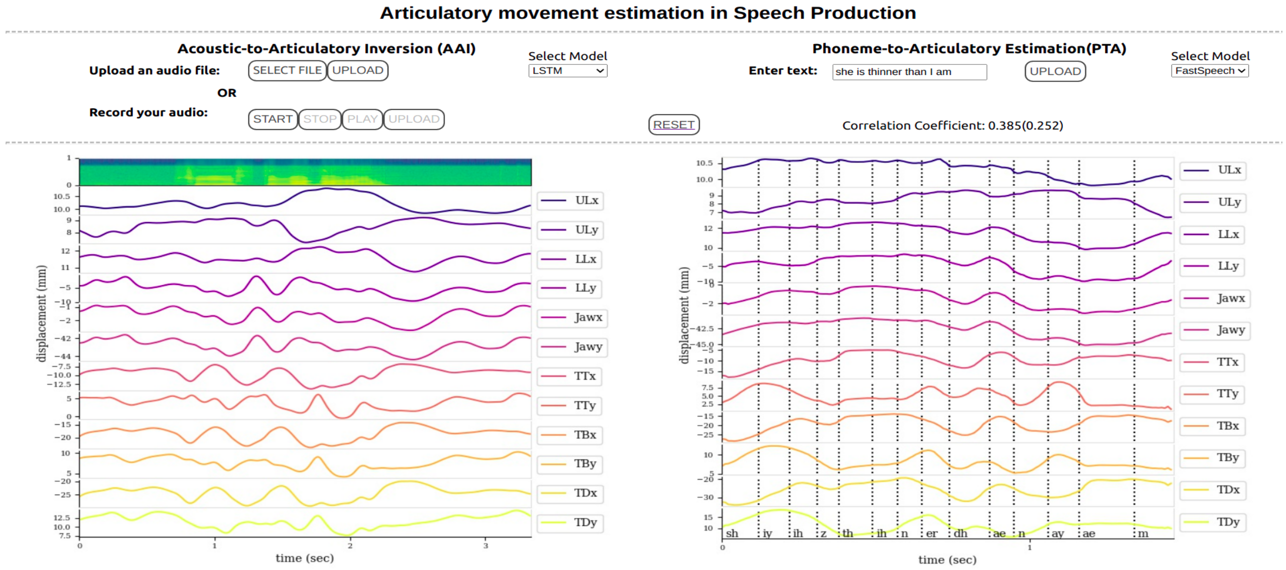
3.2. PTA

For the PTA task, we use FastSpeech [13] and Tacotron2 [16] model architectures trained to estimate articulatory movements.

4. Functionality

As shown in Fig. 1, the end-user can visualise estimated articulators from two modalities - audio (with AAI models) and text (with PTA models). The text input is converted to a sequence of phonemes using a grapheme to phoneme converter [20]. The audio can be inserted from a file or through the live recording which is then converted into MFCC features. Articulatory movements are then predicted using the selected model and visualised on the webpage. The user can also provide text for the audio uttered and obtain a Correlation Coefficient score on the articulatory movements predicted between the two modalities. Towards this, DTW alignment is performed using the Euclidean distance metric between the predicted articulatory movements between the two modalities, as utilised in [16]. A screenshot of the webpage is shown in Fig. 2.

We use Pytorch and Keras for different model inference in the backend which is hosted by a python FASTApi server. We design the front end with reactJS.



Articulatory movement estimation in Speech Production

Acoustic-to-Articulatory Inversion (AAI)

Upload an audio file: SELECT FILE UPLOAD

OR

Record your audio: START STOP PLAY UPLOAD

Select Model: LSTM

Phoneme-to-Articulatory Estimation(PTA)

Enter text: she is thinner than I am UPLOAD

Select Model: FastSpeech

RESET

Correlation Coefficient: 0.385(0.252)

Figure 2: *The screenshot of the webpage is shown in the figure. The audio needs to be uploaded in the left half of the page while the text is required to be entered in the right half. The user can select the model to be used for inference. The predicted articulatory movements plot is displayed below the corresponding modality. The correlation score is displayed once both inputs are entered.*

5. Discussion

The webpage interface allows the end-user to visualise estimated articulators for audio and text input with different trained models. PTA inference is considered noise-free since the input is text. On the other hand, AAI is trained on audio collected in a noise-free environment while the inference could be on a noisy audio. Due to this, for the same input pair (audio and corresponding text) we consider PTA estimation to be clean and AAI inference to be conditioned on environmental noise. We present the correlation between the two predictions to provide insight into the effect of noise. In future, we intend to add Automatic Speech Recognition to the interface to avoid user text input.

6. References

[1] J. Frankel, K. Richmond, S. King, and P. Taylor, "An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces," *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, 2000.

[2] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, University of Bielefeld, 1999.

[3] C. S, C. Yarra, R. Aggarwal, S. K. Mittal, K. N K, R. K T, A. Singh, and P. K. Ghosh, "Automatic visual augmentation for concatenation based synthesized articulatory videos from real-time MRI data for spoken language training," in *Proc. Interspeech*, 2018, pp. 3127–3131.

[4] U. Desai, C. Yarra, and P. K. Ghosh, "Concatenative articulatory video synthesis using real-time MRI data for spoken language training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4999–5003.

[5] A. Illa and P. K. Ghosh, "An investigation on speaker specific articulatory synthesis with speaker independent articulatory inversion," in *Proc. Interspeech*, 2019, pp. 121–125.

[6] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, 2009.

[7] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, 2008.

[8] L. Zhang and S. Renals, "Acoustic-articulatory modeling with the trajectory HMM," *IEEE Signal Processing Letters*, 2008.

[9] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping." in *Proceedings of the ICSLP, Pittsburgh*, 2006, pp. 577–580.

[10] Z. Wu, K. Zhao, X. Wu, X. Lan, and H. Meng, "Acoustic to articulatory mapping with deep neural network," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9889–9907, 2015.

[11] A. Illa and P. K. Ghosh, "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory," in *Proc. Interspeech*, 2018, pp. 3122–3126.

[12] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4450–4454.

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

[14] J. Shen *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[15] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems*, 2019.

[16] A. Singh, A. Illa, and P. K. Ghosh, "A comparative study of estimating articulatory movements from phoneme sequences and acoustic features," 2020.

[17] "EM9600 shotgun microphone," avaliable online: http://www.tbone-mics.com/en/product/information/details/the-tbone-em-9600-richtrohr-mikrofon/, last accessed:4/2/2020.

[18] "3d electromagnetic articulograph," available online: http://www.articulograph.de/, last accessed: 4/2/2020.

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

[20] K. Park and J. Kim, "g2pe," https://github.com/Kyubyong/g2p, 2019.