Ultra-Arti-Synth - Articulatory Vowel Synthesis from Ultrasound Tongue

Pramit Saha [1]*, Yadong Liu*[2]*, Bryan Gick*[2,3]*, Sidney Fels*[1]

[1] Dept. of Electrical Computer Engineering, University of British Columbia, Vancouver, Canada
[2] Dept. of Linguistics, University of British Columbia, Vancouver, Canada
[3] Haskins Laboratories, New Haven, Connecticut, USA

pramit@ece.ubc.ca

Abstract

In this work, we develop an Ultrasound based Silent Speech Interface (SSI) for generating vowel sounds utilizing articulatory speech synthesis. For this, we collect Ultrasound scans of tongue movements from a single participant uttering continuous vowel sounds. We further trace the palate as well as the tongue contour from the ultrasound images to compute the 1D vocal tract area functions between the tongue surface and palate. We feed these area function values to the articulatory speech synthesizer JASS to get the synthesized continuous vowel sounds as the outputs. We perform a formant-based analysis to evaluate the similarity of the reconstructed vowel sounds with the original utterances. In order to further improve vowel outputs from Ultra-Arti-Synth, the user can use the output as feedback to manipulate tongue positions in different ways to achieve the desired vowel sound targets.

1. Introduction

Silent Speech Interfaces (SSI) [1] are used to recognize or generate speech in the absence of any acoustic signal recording. Such recognition or reconstruction of speech sounds can be done from other bio-signals like EEG [2–4], EMG [1], EMA [1], imaging modalities like MRI [5], Ultrasound [1], CT as well as different kinematic sensors, optical imaging [1] and video acquisition systems. Among these methods, most reliable are those which capture more relevant information corresponding to desired speech motor tasks by directly recording the soundless articulatory movements.

Motivated by [6], this paper particularly explores the use of ultrasound tongue images as the input signal for directly synthesizing continuous vowel sounds, without the necessity of any intermediate vowel recognition steps. Rather than identifying discrete vowel sounds or synthesising discrete vowels with indirect methods like Mel-generalized cepstral coefficients, Linear Predictive Coding *etc.*, we follow bio-inspired articulatory speech synthesis technique for continuous vowel production. To the best of our knowledge, this is the first work investigating area function-based articulatory speech synthesis for continuous vowel sounds from ultrasound tongue movements.

2. Data collection

We collect midsagittal ultrasound video of a single male participant for a total time duration of 2 hours. Throughout the data collection procedure, the participant was seated with his head stabilized against a headset and was asked to make continuous open vocal tract sounds (vowel sounds) for each trial. For imaging the tongue, the ultrasound transducer was placed beneath the chin. The ultrasound beam travels upward through the tongue body and reflects back from the upper tongue surface because the air above the tongue has a different acoustic impedance than the tissue. This reflection results in a bright curved line in the ultrasound image which changes dynamically as the tongue moves.

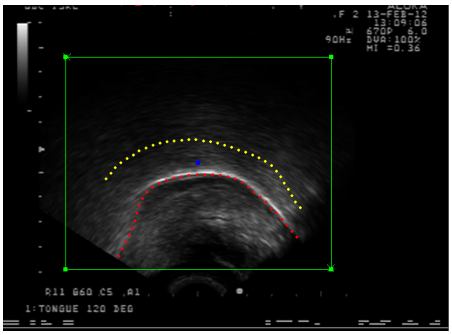

Figure 1: *The palate (indicated by yellow dotted line) and tongue contour (indicated by red dotted line) in an ultrasound image*

3. Palatal Tracing and Tongue Tracking

The air between the tongue surface and the upper palate reflects the ultrasound beam back to the transducer and hence Ultrasound Imaging cannot clearly capture the palatal structure [7]. However, during swallowing, the tongue touches the palate and the ultrasound beam can be transmitted through the soft tissue of the tongue till it gets reflected by the palatine bone. We observe such instants of dry and wet swallowing in the ultrasound image sequences and manually trace the palatal contour. Since the participant's head is kept fixed throughout the process, it helps in keeping the palatal contour static at all instants. This is further confirmed by comparing the traced palatal contours at different intermediate swallowing processes in between the utterances.

In order to semi-automatically delineate and track the tongue contour in the ultrasound image sequences, we use SLURP [8]. This method of tongue tracking utilizes simple tongue shape and combines it with motion models possessing highly flexible active contour (snake) representation. It also has the additional option of correcting the tongue contour via a particle filtering algorithm. To achieve the tongue contour segmentation and tracking, one needs to select 6 anchor points just below the tongue surface in the initial ultrasound image frame and fit the active contour onto it. For tracking the contour across the image sequences, we use the particle filter based correction option in the SLURP Menu.

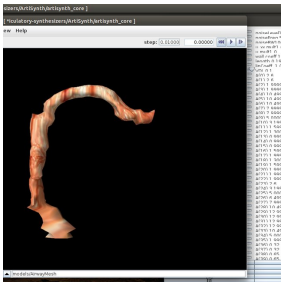

Figure 2: *Vocal Tract visualized in Artisynth (JASS based synthesis)*

4. Area function computation

At regular spatial intervals *i.e.* 30 control points (i) along the length of vocal tract, we compute the vertical distance (d_i) between the pixel coordinates of the palate and tongue surface in ultrasound images, as indicated in Fig 1, and using these values as the diameters, we derive the corresponding vocal tract area functional values (A_i) [9] which will be utilized for the next step.

5. Articulatory speech synthesis engine

Following our previous work [10], we approximately model the vocal tract sound propagation by using a wave-guide model in a 1D acoustical tube. For sound propagation in the vocal tract, a well known physical model is Kelly-Lochbaum (KL) which employs a 1D acoustical tube structure characterized by an area function. The underlying idea of KL model is that a 1D plane wave that surfaces from the far end of the vocal tract (glottis) travels through a line of concentric cylinder segments with varying cross-sectional areas defined by area function to the open mouth end.

In this work, we implement a method described in [11] which eliminates the drawbacks of the KL model. The vocal tract is modeled as an acoustic tube with its shape changing accordingly with the area functions received from the previous module. Glottal excitation pulse is generated according to Rosenberg's model [12]. This vocal fold model is coupled to discretized acoustic equations in the vocal tract. The acoustic wave propagation is simulated by numerically integrating the linearized 1D Navier-Stokes pressure-velocity Partial Differential Equations (PDE) in time and space on a non-uniform grid. The synthesis mechanism involves excitations acting as source placed in the tube and sound propagation being simulated by approximating the pressure-velocity wave equations. The models are implemented an articulatory speech synthesizer named Java Audio Synthesis System (JASS) [13] written in pure Java as visualized in Fig 2.

6. Discussion and Conclusion

We intend to have two modes of operation in 'Ultra-Arti-Synth' - both offline and online. The main objective of the offline mode is to see if the captured area function results in the same formants as the persons' actual acoustics, where as the online mode is targeted to provide acoustic and visual feedback to the user to see if they can adapt their tongue shape for effective vowel control. For the offline one, we are performing qualitative study as well as statistical analysis of the output vowels in the formant space and quantitatively comparing the formant trajectories of the estimated output with the desired sound output from actual utterances. This will demonstrate how area function-based sound reconstruction varies from the actual sound synthesis. We also endeavour to make the process real-time so as to enable the user to learn to manipulate tongue posture to produce desired vowel sounds through the 'Ultra-Arti-Synth' interface in online mode. This will ensure better adaptability of the user to our interface through instantaneous audio-visual feedback.

We will continue to acquire more tongue ultrasound data and data from more participants to test the robustness and inter-subject variability of the sound output. Furthermore, we continue to improve the tongue tracking and achieve a more accurate area-functional computation in order to synthesize vowel sounds closer to the actual vowel utterances.

7. Acknowledgements

This work was funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada and Canadian Institute for Health Research (CIHR).

8. References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[2] P. Saha, S. Fels, and M. Abdul-Mageed, "Deep learning the eeg manifold for phonological categorization from active thoughts," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2762–2766.

[3] P. Saha and S. Fels, "Hierarchical deep feature learning for decoding imagined speech from eeg," *arXiv preprint arXiv:1904.04352*, 2019.

[4] P. Saha, M. Abdul-Mageed, and S. Fels, "Speak your mind! towards imagined speech recognition with hierarchical deep learning," *arXiv preprint arXiv:1904.05746*, 2019.

[5] P. Saha, P. Srungarapu, and S. Fels, "Towards automatic speech identification from vocal tract shape dynamics in real-time mri," *Proc. Interspeech 2018*, pp. 1249–1253, 2018.

[6] F. Vogt, G. McCaig, M. A. Ali, and S. S. Fels, "Tongue'n'groove: An ultrasound based music controller." in *NIME*, 2002, pp. 60–64.

[7] M. A. Epstein and M. Stone, "The tongue stops here: Ultrasound imaging of the palate," *The Journal of the Acoustical Society of America*, vol. 118, no. 4, pp. 2128–2131, 2005.

[8] C. Laporte and L. Ménard, "Multi-hypothesis tracking of the tongue surface in ultrasound video recordings of normal and impaired speech," *Medical image analysis*, vol. 44, pp. 98–114, 2018.

[9] S. Mathur and B. H. Story, "Vocal tract modeling: Implementation of continuous length variations in a half-sample delay kelly-lochbaum model," in *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No. 03EX795)*. IEEE, 2003, pp. 753–756.

[10] P. Saha, D. R. Mohapatra, P. SV, and S. Fels, "Sound-stream ii: Towards real-time gesture controlled articulatory sound synthesis," *arXiv preprint arXiv:1811.08029*, 2018.

[11] K. van den Doel and U. M. Ascher, "Real-time numerical solution of webster's equation on a nonuniform grid," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 6, pp. 1163–1172, 2008.

[12] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 583–590, 1971.

[13] K. van den Doel and D. K. Pai, "Jass: a java audio synthesis system for programmers." Georgia Institute of Technology, 2001.