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

ViSTAFAE: A Visual Speech-Training Aid with Feedback of Articulatory Efforts

Pramod H. Kachare [1, 2], Prem C. Pandey [1], Vishal Mane [3], Hirak Dasgupta [1], K. S. Nataraj [1], Akshada Rathod [3], Sheetal K. Pathak [3]

[1]Electrical Engineering Dept., Indian Institute of Technology Bombay, India
[2]Electronics & Telecommunication Engineering Dept., Ramrao Adik Institute of Technology, India
[3]Digital India Corporation, India

kachare.pramod1991@gmail.com, pcpandey@ee.iitb.ac.in, vishal.mane@digitalindia.gov.in, hirakdgpt@ee.iitb.ac.in, natarajks@ee.iitb.ac.in, akshadarathod2@gmail.com, sheetalkumar.pathak66@gmail.com

Abstract

An app is presented as a speech-training aid for providing visual feedback of articulatory efforts using information obtained from the utterances' audiovisual recording. It has two panels to enable comparison between the articulatory efforts of the learner and the teacher or a pre-recorded reference speaker. The visual feedback consists of a slow-motion animation of lateral vocal tract shape, level, and pitch, and time-aligned display of the frontal view of the speaker's face along with playback of the time-scaled speech signal. The app comprises a graphical user interface and modules for signal acquisition, analysis, and animation. It is developed using Python as a Windows-based app and may be accessed remotely through a web browser.

Index Terms: articulatory feedback, speech-training aid, vocal tract animation

1. Introduction

In hearing-impaired children, lack of auditory feedback hinders the acquisition of speech production despite functional articulatory organs. Speech training involves repeated utterances of syllables, words, word clusters, and sentences by the learner and corrective feedback by the teacher. A mirror is often used for visual feedback. It shows movement of the external articulators (lips, jaw) but not that of the internal articulators (tongue, vocal tract shape, glottis), and it does not provide information on level and pitch. Several computer-based aids with a dynamic display of acoustic parameters (level, voicing, pitch, spectral features, etc.) and articulatory parameters (movements of lips, jaw, tongue, etc.) have been reported for speech training [1]–[6]. However, most speech therapists and special education teachers prefer to use gestures with repeated and extended articulations. It has been suggested that the usefulness of the aids can be improved by minimizing the level of details on the display, highlighting the key articulatory efforts, adapting the feedback to the learning level, and enhancing the feedback with complementary information [7].

After interactions with the therapists and teachers to understand the difficulties experienced by them in using speech-training aids, an aid was developed with a two-panel display for a slow-motion animation of the lateral vocal tract shape with indicators for the level and the pitch [8]. The two-panel design enables the teacher and learner to visually compare the articulatory efforts, avoids a need for repetitive articulation by the teacher, and facilitates distant learning with pre-recorded utterances of a reference speaker. It was suggested during further interactions that the aid should also show lip movements by displaying the speaker's face and that a time-scaled speech playback could help children with hearing aids or cochlear implants [9]. An aid with these features may also be helpful for improving the pronunciation of second-language learners and speech rehabilitation therapy.

We present ViSTAFAE (*Vi*sual *S*peech-*T*raining *A*id with *F*eedback of *A*rticulatory *E*fforts) developed as a two-panel aid with slow-motion animation of the lateral vocal tract shape, level and pitch indicators, time-aligned display of the frontal view of the speaker's face, and playback of time-scaled speech. It is a Windows-based app, comprising signal acquisition, analysis, and animation modules integrated with a graphical user interface (GUI). The second section describes the app features, followed by the app implementation in the third section and the conclusion in the last section.

2. App Features

The app has two panels: (i) for the utterances of the learner and (ii) for those of the teacher or a pre-recorded reference speaker. It provides animation of the lateral vocal tract shape, level, and pitch. It also provides a time-aligned display of the frontal view of the speaker's face so that the learner can relate them together. For facilitating the learning process, the rate of slow-motion animation can be changed with the learning level. It also provides playback of time-scaled speech that may be helpful for children with hearing aids or cochlear implants. The app is developed using Python to facilitate its distribution for machines with different operating systems. It works using the machine's audio and video peripherals without additional hardware. It is usable as a standalone app or as a server-based solution accessed remotely using a web browser.

3. App Implementation

The app is implemented as a GUI integrated with three modules: (i) audiovisual signal acquisition (Signal), (ii) signal analysis for obtaining the information related to articulatory efforts (Analysis), and (iii) slow-motion animation of invisible articulatory efforts with time-aligned video of the speaker's face and speech playback (Animation). The screen is split into two panels with two vertical control bars for module selection and a bottom horizontal toolbar along with a central vertical toolbar for controlling the display features. Colors for all sections of the screen are user-settable.

The Signal module is used for recording an utterance or accessing a pre-recorded utterance. The recording duration can be up to 10 s, for flexible utterance timing, repeated utterances to adjust the level, etc. The audio signal is recorded at a sampling frequency of 10 kHz with the waveform displayed

using color bands indicating the volume as low, acceptable, and high. The video of the speaker's face is simultaneously recorded at 10 frames/s with markings in the frame for positioning the face. Each panel has graphical controls to record and playback the audiovisual signal, open and save the signal file, and select a segment for analysis and animation.

The Analysis module analyzes the selected signal segment to obtain the information related to articulatory efforts. The pitch estimation uses glottal epoch detection with the Hilbert envelope for excitation enhancement as in [10]. The lateral vocal tract shape estimation uses Wakita's LP-based inverse filtering [11] with analysis window positioning for improved estimation consistency [12]. The shape is obtained as a set of area values for 20-ms frames and 5-ms frameshift. Cubic B-spline interpolation is used to obtain twenty area values uniformly placed along the oral cavity length. Time-scaled speech signals are obtained using the synchronous overlap-add with fixed synthesis (SOLAFS) method as in [13]. The video frames are processed for face detection using Viola-Jones algorithm [14] and resized to emphasize lip movement. The module displays the speech signal, level, pitch, spectrogram, and 2D plot of time-varying vocal tract area function as an 'areagram' [15]. The spectrum and area values of either panel can be read by moving the cursor.

The Animation module provides a variable-rate animation of the lateral vocal tract shape, level and pitch indicators, and time-aligned playback of video of the speaker's face and speech signal. A mid-sagittal view of the head, as in [2], with area between the fixed upper curve (comprising upper lip, upper teeth, and palate) and moving lower curve (comprising lower lip, lower teeth, and tongue) is used for vocal tract animation. The maximum constriction can be optionally marked. The display at the selected animation rate is obtained by downsampling the analysis results available at 200 frames/s and upsampling the video frames available at 10 frames/s. Time-scaled speech signal is used for time-aligned audio playback. Specific graphical controls are provided for customizing the display of level and pitch bars (horizontal, vertical, off), maximum constriction marker (on, off), the animation face (man, woman, boy, girl), the face orientation (left, right), video (on, off), and slowdown factor for the animation rate (1, 2, 5, 10, 20). Two cursors control the start and stop positions of the animation. There is a provision for simultaneous animation in the two panels. A screenshot of the app during animation is shown in Fig. 1, and a demonstration of the app is available in [16].

4. Conclusions

ViSTAFAE has been developed as an app for visual feedback of key articulatory efforts based on the information obtained from audiovisual recordings. It has been tested for its functionalities and user interface. The approximate processing time for generating the animation is twice the segment duration, and the implementation should be revised to reduce this time. The app has to be evaluated for speech training of hearing-impaired children, improving the pronunciation of unfamiliar sounds by second-language learners, and speech rehabilitation therapy for level and intonation control.

5. Acknowledgments

The work is supported by the project "Visual Speech Training System Phase-2", MEITY, Government of India. The authors are grateful to Dr. Niranjan D. Khambete (Deenanath Mangeshkar Hospital & Research Center, Pune) and Dr. Milind S. Shah (Fr. C. Rodrigues Institute of Technology, Navi Mumbai) for insightful suggestions and discussions.

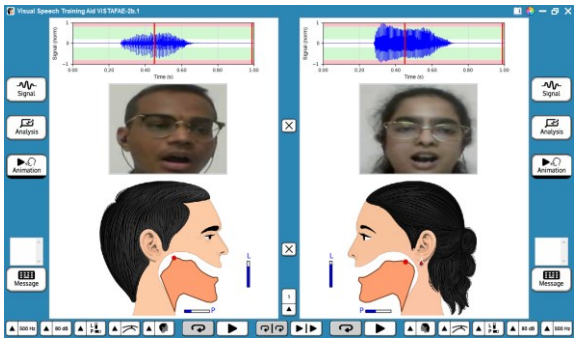

Figure 1: *Screenshot of the app during animation.*

6. References

[1] F. R. Adams, H. Crepy, D. Jameson, and J. Thatcher, "IBM products for persons with disabilities," in *Proc. IEEE Global Telecommun. Conf. Exhibition 'Commun. Technol. 1990s and Beyond*, Dallas, TX, USA, 1989, pp. 980–984.

[2] S. H. Park, D. J. Kim, J. H. Lee, and T. S. Yoon, "Integrated speech training system for hearing impaired," *IEEE Trans. Rehab. Eng.*, vol. 2, pp. 189–196, 1994.

[3] D. Xu, Z. Ma, Z. Jian, L. Shi, L. Wang, and J. Gao, "Speech rehabilitation system for hearing impaired children on virtual reality technology," in *Proc. Int. Conf. Virtual Reality Visualization*, Hong Kong, China, pp. 211–214, 2020.

[4] Tiger DRS, Dr. Speech, Accessed: Feb, 2021. Available: www.drspeech.com/platform

[5] Micro Video Corp., Video Voice Speech Training System, Accessed: Feb, 2021. Available: www.videovoice. com

[6] K. Vicsi, Box of Tricks, Accessed: Feb. 2021. Available: www.enl.auth.gr/phonlab/box_of_tricks.html

[7] E. Eriksson, O. Engwall, O. Bälter, A. Öster, and H. Kjellström, "Design recommendations for a computer-based speech training system based on end-user interviews," in *Proc. SPECOM*, Patras, Greece, 2005, pp. 483–486.

[8] R. Jain, K. S. Nataraj, and P. C. Pandey, "Dynamic display of vocal tract shape for speech training," in *Proc. NCC 2016,* Guwahati, India, paper no. 1570220186

[9] L. Czap, "Automated speech production assessment of hard of hearing children," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 380–389, 2020.

[10] H. Dasgupta, P. C. Pandey, and K. S. Nataraj, "Epoch detection using Hilbert envelope for glottal excitation enhancement and maximum-sum subarray for epoch marking," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 461–471, 2020.

[11] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.*, vol. 21, no. 5, pp. 417–427, 1973.

[12] K. S. Nataraj, Jagbandhu, P. C. Pandey, and M. S. Shah, "Improving the consistency of vocal tract shape estimation," in *Proc. NCC 2011*, Bangalore, India, paper SpPrII.4.

[13] D. Hejna, B. R. Musicus, "The SOLAFS time-scale modification algorithm,". Bolt, Beranek and Newman (BBN) Technical Report, University of Cambridge, UK, 1991.

[14] P. Viola, and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, 137–154, 2004.

[15] P. C. Pandey and M. S. Shah, "Estimation of place of articulation during stop closures of vowel-consonant-vowel utterances," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 277–286, 2009.

[16] P. H. Kachare, "Demo Video of ViSTAFAE (Visual Speech-Training Aid with Feedback of Articulatory Efforts), Show-and-tell session, Interspeech 2021," Available: https://www.ee.iitb.ac.in/~spilab/material/pramod/is2021