Speech-Training Aid with Time-Scaled Audiovisual Feedback of Articulatory Efforts

Pramod H. Kachare [1, 2], Prem C. Pandey [1], Vishal Mane [3], Hirak Dasgupta [1], K. S. Nataraj [1]
[1] *Dept. of Electrical Engg., Indian Institute of Technology Bombay, Mumbai, India*
[2]*Dept. of Electronics & Telecom. Engg., Ramrao Adik Institute of Technology, Navi Mumbai, India*
[3] *Digital India Corporation, New Delhi, 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

Abstract— **Hearing-impaired children lack auditory feedback and experience difficulty in acquiring speech production. They can benefit from speech training aids providing visual feedback of key articulatory efforts. Requirements for such aid are developed through extended interaction with speech therapists and special education teachers. The aid is developed as a PC-based app for ease of distribution and use. It has two panels to enable comparison between the articulatory efforts of the learner and a teacher or a pre-recorded reference speaker. The visual feedback for an utterance is based on the information obtained from its audiovisual recording. The speech signal is processed to obtain time-varying vocal tract shape, level, and pitch. The vocal tract shape estimation uses LP-based inverse filtering, and the pitch estimation uses glottal epoch detection using Hilbert envelope for excitation enhancement. Visual feedback comprises a variable-rate animation of the lateral vocal tract shape, level, and pitch, and time-aligned display of the frontal view of the speaker's face along with playback of time-scaled speech signal. The graphical user interface and modules for signal acquisition, speech analysis, and time-scaled animation are developed and integrated using Python. The app has been tested for its functionalities and user interface and needs to be evaluated for speech training of hearing-impaired children. It may also be useful to second-language learners in improving the pronunciation of unfamiliar sounds.**

Keywords— Articulatory feedback, hearing impairment; speech-training aid; vocal tract animation

I. INTRODUCTION

Children with hearing impairment lack auditory feedback during speech production. They consequently experience difficulty in the acquisition of speech production despite functional articulatory organs. Speech therapy often uses a mirror for a visual feedback on movements of the external articulators (lips, jaw), but movements of the internal articulators (tongue, vocal tract shape, glottis) remain hidden. Computer-based speech-training aids providing a dynamic display of acoustic parameters (speech level, voicing, pitch, spectral features, etc.) and articulatory parameters (movements of lips, jaw, tongue, etc.) have been found to be useful for speech training [1]–[15]. These aids include 'Speech Viewer' [9], 'Video Voice Training System' [10], 'Dr. Speech' [11], and 'Box-of-Tricks' [12], concatenative articulatory video synthesis [13], 3-D virtual talking head [14], speech rehabilitation system [15], system based on automatic lip-reading recognition [16]. These aids display acoustic parameters to help persons with hearing impairment to improve the articulation. Several mobile-based apps have been developed, including 'Voice Analyst' [17], 'Voice Tools' [18], and 'Voice Pitch Analyzer' [19], to display speech level and pitch information for use in speech therapy. The usefulness of most of these aids is limited as they use language-dependent processing and machine learning for generating corrective feedback.

Despite the development of several speech-training aids, most speech therapists and special education teachers in the schools for hearing-impaired children find it more convenient to use mirrors and improvised methods with gestures, repeated and extended articulations, etc. It indicates a need to co-design the aid with the speech therapists and teachers, by interacting with them to understand their difficulties with the available aids and getting their assessment of the features in the new aid. It has been reported that optimizing the level of details on the display, emphasizing the key articulatory efforts, adapting the feedback to the learner's level, and augmenting with complementary information can improve the usefulness of the aids [20]. Interaction with the users showed the usefulness of a two-panel aid with a variable-rate display. It enables visual comparison of the teacher's and learner's articulatory efforts at a rate suitable for the learning level, avoids the need for repetitive articulation by the teacher, and facilitates remote learning with pre-recorded utterances of a reference speaker.

An aid in the form of a PC-based app with two-panel display and a reconfigurable graphical user interface (GUI) was developed to display level, pitch, and lateral vocal tract shape estimated from the speech signal, with a variable-rate animation of the articulatory efforts [8]. The app was demonstrated and installed for extended use by the speech therapists in three speech therapy clinics and the special education teachers in four schools for hearing-impaired children. Responses were received from 35 users. They suggested improving the GUI layout, extending the signal acquisition duration, and introducing a selection of faces (male/female, young/adult) for vocal tract shape animation. They also suggested a facility for playback of the speech signal time-aligned with the variable-rate animation of the articulatory efforts to improve speech acquisition in children using hearing aid or cochlear implants. As in [21], some respondents suggested displaying the frontal view of the speaker's face to serve the mirror function for feedback on the lip movements. Based on these inputs, a speech-training aid with two-panel display for variable-rate audiovisual feedback of the articulatory efforts is developed.

The main features of the speech-training aid include (i) variable-rate animation of the lateral vocal tract shape, (ii) display of the level and the pitch, and (iii) a time-aligned audiovisual playback comprising the time-scaled speech signal and the time-scaled video of frontal view of the speaker's face. As suggested in [22], the aid uses signal processing for generating the feedback of articulatory efforts and enables a language-independent speech visualization.

978-1-6654-4177-3/21/$31.00 ©2021 IEEE

The system implementation of the aid is described in the next section, followed by the test results in the third section and the conclusion in the last section.

II. System Description

Speech training involves repeated utterances of syllables, words, word clusters, and sentences by the learner and feedback for correct production by the teacher. The speech-training aid is developed to assist this learning process by providing information not visible on the speaker's face. The feedback is in the form of slow-motion animation of the vocal tract shape, pitch, and signal level, obtained by processing the speech signal. It is enhanced by complementary information on lip movements by a time-aligned display of the frontal view of the speaker's face so that the learner can relate the various articulatory efforts together. For facilitating the learning process, the rate of audiovisual feedback can be altered with the learning level. The aid also includes playback of the time-scaled speech signal for children with hearing aids or cochlear implants.

The aid is developed as a Windows-based app using Python to integrate the information processing modules and GUI and to ease its distribution for machines with other operating systems. For a large-scale deployment in schools for children with hearing impairment, the app is developed to work using the machine's audio and video peripherals without additional hardware or software.

The app is developed with modules for signal acquisition, speech analysis, and time-scaled animation of the key articulatory efforts, along with a GUI for user inputs and information display. The signal acquisition comprises recording the speaker's utterance with the speech signal as an audio signal and frontal view of the speaker's face as a video signal. A segment of the audio recording and the corresponding video recording are selected from the acquired audiovisual signal for analysis and animation. The speech analysis comprises analyzing the selected audio segment to obtain the time-varying level, spectrum, pitch, and vocal tract shape, and generating time-scaled audio with different scaling factors. The display comprises animation of the time-varying lateral vocal tract shape and indicators for pitch, level, and place of articulation, with settable display rate. The animation may be accompanied by playback of a time-scaled video of the speaker's face with speech signal. The three information processing modules and the GUI are described in the following subsections.

A. Signal Acquisition

The signal acquisition module is used for recording the speaker's utterance with the speech signal as an audio signal and the frontal view of the speaker's face as a video signal. The audio signal acquisition and playback use the machine's audio peripheral. The audio signal is displayed in real-time during signal acquisition, with three color bands to indicate the utterance volume as low, acceptable, and high. The audio signal acquisition is carried out at 10 kHz as it is considered adequate for the pitch and vocal-tract shape estimation [23]. A video of the speaker's face is recorded and displayed in real-time using the machine's video peripheral at the frame rate of ten frames/s. The low frame rate is selected to reduce memory requirement and execution time on low-resource machines. The video is displayed in a frame, with its center marked by vertical and horizontal lines to assist the speaker in adjusting the face position. A previously recorded audiovisual signal can also be used for analysis and display. The utterances used for speech training are usually shorter than 2 s. The signal acquisition module permits a recording of up to 10 s, allowing flexibility for the timing of the utterance and multiple utterances to adjust the level, etc. The maximum duration is selected for real-time display of the audio signal with a fixed time axis, and it can be relaxed by providing a scrolling display. A segment of the audio signal is selected for analysis and animation, and the time-aligned video segment is selected automatically for simultaneous display of articulatory efforts visible on the speaker's face.

B. Speech Analysis

This module is used for obtaining the information for speech therapy and providing feedback for speech training. It comprises analysis of the selected audio segment to obtain the time-varying level, spectrum, pitch, and vocal tract shape, and generation of time-scaled audio with different time-scaling factors.

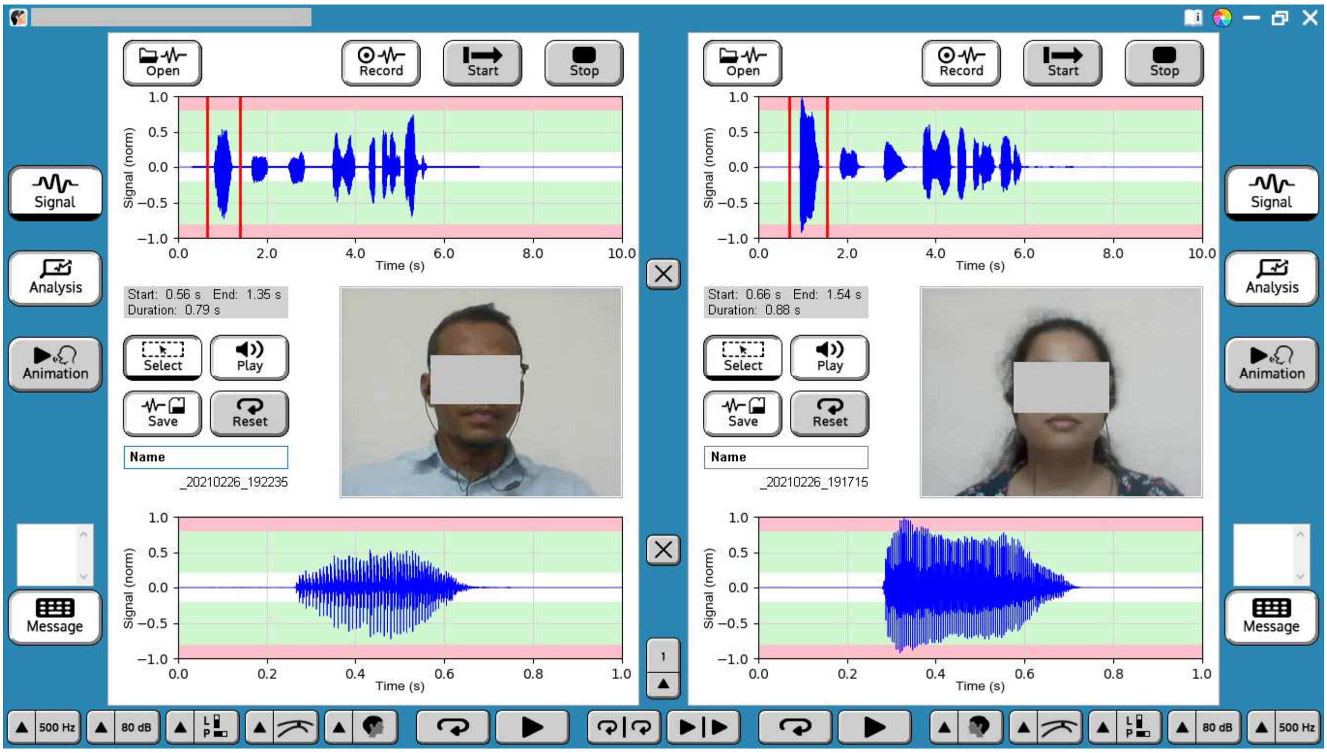
The pitch estimation uses a glottal epoch detection method using the Hilbert envelope for excitation enhancement and reported in [24] as suitable for speech signals with voice disorders. The vocal tract shape estimation uses Wakita's inverse filtering method based on linear prediction (LP) for obtaining the vocal tract area function [25]. The analysis is carried out using a window length of two average pitch periods and a window shift of 5 ms. For improving consistency of the vocal tract shape estimation, the area values for analysis windows positioned at the minimum windowed energy are linearly interpolated to obtain area values every 5 ms, as reported in [26]. The cubic B-spline interpolation is used to estimate 20 uniformly placed area values along the length of the oral cavity for smoothing the estimated vocal tract area function. For synchronized audiovisual playback, the speech signal is uniformly time-expanded by the selected scaling factor using synchronous overlap add with fixed synthesis (SOLAFS) method [27].

C. Time-Scaled Animation

The animation module formats the information obtained by the speech analysis module and displays the vocal tract animation and indicators for pitch, level, and place of articulation. A time-aligned video of the speaker's face is also displayed to provide simultaneous feedback of articulatory efforts visible on the speaker's face. The vocal tract animation and video of the speaker's face can be displayed at the selected speed for speech training, along with a playback of the time-scaled audio signal.

The estimated area values are used for vocal tract animation with a 2D mid-sagittal view of the head, as in [2]. The estimated oral cavity opening is represented as the area between the fixed upper curve (comprising upper lip, upper teeth, and palate) and moving lower curve (comprising lower lip, lower teeth, and tongue). The place of maximum constriction along the vocal tract length is indicated as the place of articulation.

The vocal tract area function is estimated with a 5-ms window shift resulting in 200 frames/s for real-time animation, while the video of the speaker's face is available at ten frames/s. The vocal tract animation, level, and pitch are time-scaled for the selected animation speed by down-sampling. The video of the speaker's face is time-scaled for the selected animation speed by frame repetition.



Start: 0.56 s End: 1.35 s
Duration: 0.79 s

Name
_20210226_192235

Start: 0.66 s End: 1.54 s
Duration: 0.88 s

Name
_20210226_191715

Fig. 1. Display of the app during signal acquisition, with a male speaker on the left panel and a female speaker on the right panel.

D. *GUI for Speech-Training Aid*

The aid's GUI is developed for integrating it with the modules for signal acquisition, speech analysis, and time-scaled animation. A two-panel design is used to enable comparison of the articulatory efforts of the learner and the teacher or a pre-recorded reference speaker. It can also be used to validate the animation of the vocal tract shape using a side-by-side display of analysis and animation of the same signal segment. Each panel has a vertical menu bar, with three graphical controls for module selection: Signal (signal acquisition), Analysis (speech analysis), and Animation (time-scaled animation). The information display is controlled by graphical controls located on the bottom horizontal and center vertical toolbars.

The display panels for the signal acquisition module are as shown in Fig. 1. Each panel has graphical controls for recording (record, start, stop), file opening, segment selection, playback (play/pause, reset), and saving, and three display sections. The upper section is for the audio signal, the middle section is for the video of the speaker's face, and the lower section is for the selected audio segment. A segment of the acquired speech signal is selected using movable cursors for input to the analysis module.

In the analysis module, the time-varying spectrum is displayed as a spectrogram, and a 2D plot of time-varying vocal tract area function is displayed as an 'areagram' [23]. These two plots are accompanied by plots of the speech signal, level, and pitch. An example of the display of this module is shown in Fig. 2. The plots in this display can be used by speech therapists and for seeing the analysis results. The plot values in either of the panels can be read by placing the cursor at the corresponding position.

The animation module displays the selected speech segment, frontal view of the speaker's face, and vocal tract shape animation, as shown in Fig. 3. Specific graphical controls are provided for customizing the display: display of level and pitch bars along with the animation, a marker for the place of articulation, the face for animation (man, woman, boy, or girl, and as left or right facing), and the animation slowdown factor (1, 2, 5, 10, and 20). There is a provision for simultaneous animation in the two panels.

III. Test Results

The app has been extensively tested for its speech analysis results, functionalities of its modules, the correctness of the animation for speech signals from adult male and female speakers, and ease of use of its graphical controls. The displayed vocal tract shapes conformed to the corresponding shapes obtained through MRI images in [8] for different vowel and semivowel utterances. An earlier version of the app, not having the facility for video recording and time-scaled audio playback, was evaluated by a group of 35 users (speech therapists and special education teachers). The processing delay of about 5 s to generate the animation for a 1-s speech segment was considered acceptable, with preference for a shorter delay. The processing delay of the app's current version for a 1-s speech segment was measured on several notebook PCs. The smallest delay was 2.6 s for execution on a Windows 10 machine with 8 GB RAM (Intel Core i5-8265U CPU, 1.6 GHz, Windows 10 (x64) Home). It was up to 4.5 s on machines with lesser RAM and exceeded 5 s on Windows 7 machines and having lesser RAM.

IV. Conclusion

A speech-training aid has been developed as a standalone PC-based app for providing visual feedback of key articulatory efforts for an utterance based on the information obtained from its audiovisual recording. It has two panels to enable a comparison between the articulatory efforts of the learner and the teacher or a reference speaker. The app has

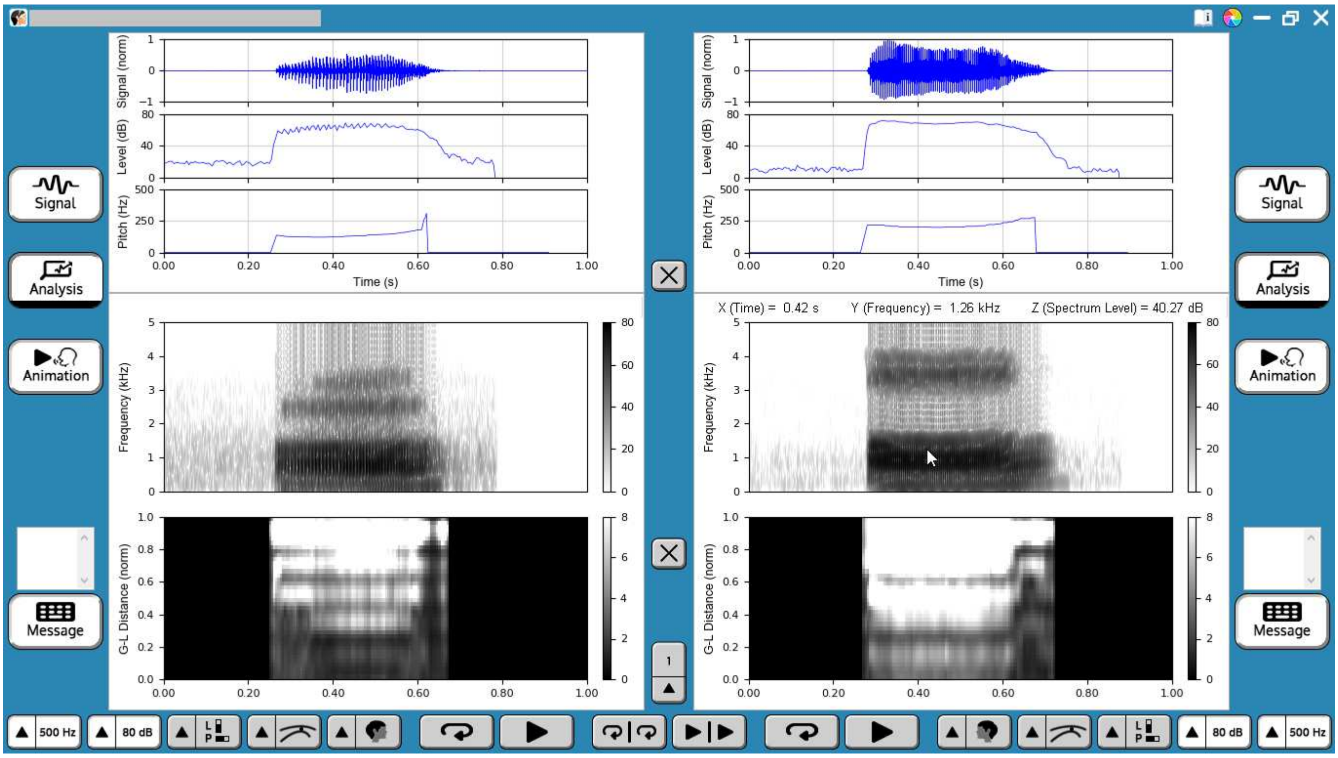

X (Time) = 0.42 s Y (Frequency) = 1.26 kHz Z (Spectrum Level) = 40.27 dB

Fig. 2. Display of the app during speech analysis.

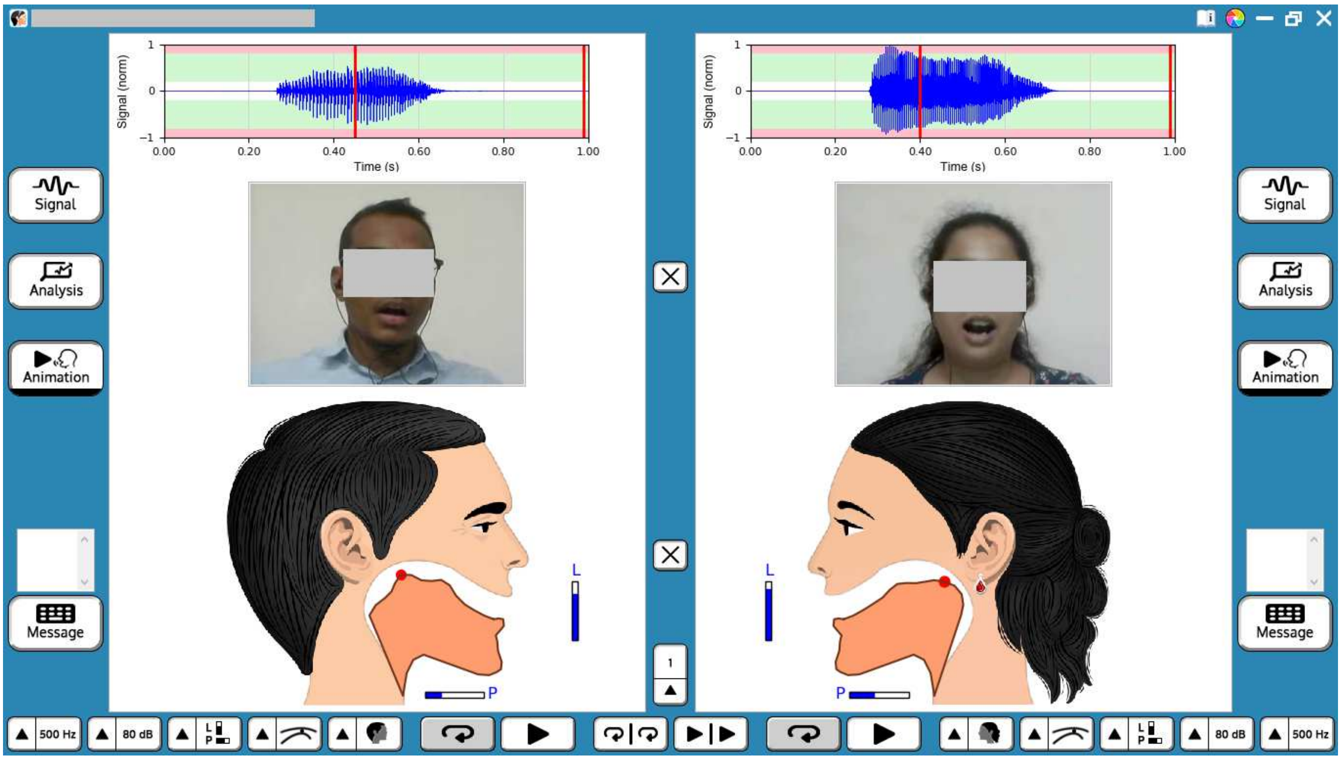


Fig. 3. Display of the app during time-scaled animation.

been tested for its functionalities and user interface. The speech processing part may be revised for reducing the processing delay for the animation generation. The app needs to be further tested for children's speech. Subsequently, it should be evaluated by the special education teachers for improving the speech production in hearing-impaired children and also by second-language learners for improving the pronunciation of unfamiliar sounds. The app may also find application in the diagnosis of voice disorders and speech rehabilitation therapy for pitch and intonation control.

References

[1] R. S. Nickerson and K. N. Stevens, "Teaching to a deaf: Can a computer help?," *IEEE Trans. Audio Electroacoust.*, vol. 21, pp. 445–455, 1973.

[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. W. Massaro and J. Light, "Using visible speech to train perception and production of speech for individuals with hearing loss," *J. Speech Lang. Hear. Res.*, vol. 47, pp. 304–320, 2004.

[4] O. Engwall, O. Balter, A. Öster, and H. Kjellstrom, "Designing the user interface of the computer-based speech training system ARTUR based on early user tests," *Behaviour Info. Technol.*, vol. 25, pp. 353–365, 2006.

[5] B. Kröger, V. Graf-Borttscheller, and A. Lowit, "Two and three dimensional visual articulatory models for pronunciation training and for treatment of speech disorders", in *Proc. Interspeech*, Brisbane, Australia, pp. 2639–2642, 2008.

[6] A. Rathinavelu and H. Thiagarajan, "Computer Aided Articulatory Tutor: A scientific study," *Int. Sci. J. Computing*, vol. 6, pp. 100–105, 2007.

[7] A. E. Mahdi, "Visualisation of the vocal-tract shape for a computer based speech training system for the hearing-impaired," *Open Elect. Electron. Eng. J.*, vol. 2, pp. 27-32, 2008.

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

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

[10] Micro Video Corp., Video Voice Speech Training System, Ann Arbor, Michigan, 2003. [Online]. Available: www.videovoice.com

[11] Tiger DRS, Dr. Speech. Seatle, WA, USA, 1998. [Online] Available: www.drspeech.com/platform (Last accessed: Feb., 2021)

[12] K. Vicsi, Box of Tricks, Budapest, Hungary. [Online]. Available: www.enl.auth.gr/phonlab/box_of_tricks.html (Last accessed: Feb., 2021)

[13] U. Desai, C. Yarra, and P. K. Ghosh, "Concatenative articulatory video synthesis using real-time MRI data for spoken language training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP),* Calgary, Alberta, Canada, pp. 4999–5003, 2018.

[14] X. Peng, H. Chen, L. Wang, and H. Wang, "Evaluating a 3-D virtual talking head on pronunciation learning," *Int. J. Human-Computer Studies*, vol. 109, pp. 26–40, 2018.

[15] 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.

[16] Y. Lu, S. Yang, Z. Xu, J. Wang, "Speech training system for hearing impaired individuals based on automatic lip-reading recognition," in *Int. Conf. Applied Human Factors and Ergonomics* (*AHFE*), vol. 1207, pp. 250–258, 2020.

[17] Speechtools Ltd., Voice analyst, Bristol, UK, 2021. [Online] Available: play.google.com/store/apps/details?id=co.speechtools.voiceanalyst

[18] DevExtras, Voice tools: Pitch, tone, and volume, Stafford, UK, 2021. [Online] Available: play.google.com/store/apps/details?id=com.DevExtras.VoiceTools

[19] Purr Programming, Voice pitch analyser, Hamburg, Germany, 2017. [Online] Available: play.google.com/store/apps/details?id=de.lilithwittmann.voicepitchanalyzer&hl=en

[20] 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.

[21] L. Czap, J. M. Pintér, and E. Baksa-Varga, "Features and results of a speech improvement experiment on hard of hearing children," *Speech Commun.*, vol. *106*, pp. 7–20, 2019.

[22] 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.

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

[24] 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. Selected Topics Signal Process.*, vol. 14, no. 2, pp. 461–471, 2019.

[25] 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.

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

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