Dynamic Display of Vocal Tract Shape for Speech Training

Rahul Jain, K. S. Nataraj, and Prem C. Pandey
Dept. of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India
<rahulmailsonix@gmail.com>, <natarajks@ee.iitb.ac.in>, <pcpandey@ee.iitb.ac.in>

***Abstract*—Children suffering from prelingual hearing impairments have difficulty in speech acquisition due to lack of auditory feedback. They can benefit by speech training aids providing corrective feedback, especially those providing visual feedback of key articulatory efforts. These aids should enable a comparison between the articulatory efforts of the student and that of a teacher or a reference speaker. A system is developed for dynamic display of vocal tract shape to provide visual feedback for production of short voiced utterances. It displays speech waveform, spectrogram, and areagram as validation tools. Intensity, pitch, and vocal tract shape are displayed for use in speech training. LPC analysis of the speech signal is used to estimate the vocal tract shape. Articulatory feedback is provided through 2D mid-sagittal view of the vocal tract, in the form of a variable rate animation emphasizing the place of maximum constriction and the opening at this place. The display has two panels, one for the articulatory efforts of the student and another for that of the teacher, for newly recorded utterances as well as pre-recorded ones.**

Keywords—articulatory feedback; hearing impairment; speech training aid; vocal tract animation

I. INTRODUCTION

The process of acquiring speech in children with normal hearing is primarily supported by auditory feedback. Children born with hearing impairment and those who suffered impairment at a prelingual stage have difficulty in speech acquisition. Although their articulatory organs may be functional, difficulty arises due to the lack of auditory feedback and they have to rely on visual cues. Lipreading and visual feedback using a mirror help to some extent, but they do not help in deciphering the movements of internal articulators. While the therapist plays a crucial role in speech training, computer based speech training systems are useful in motivating children to practice speech training and in providing feedback on their progress [1]. Such systems can also help in pronunciation training for second-language learning [2].

Speech training aids, employing visual display of articulatory efforts which are not externally visible, have been reported to be effective in improving articulation by the hearing impaired [3]–[9]. Most of these aids provide a display of the vocal tract shape. As imitation and self-correction both are integral parts of speech training, the aid should help in altering the articulatory effort, without causing a high perceptual load. Therefore, the training aids should facilitate a comparison of one's own articulatory efforts with those of a teacher or a pre-recorded reference speaker. For fulfilling this requirement, a speech training aid with vocal tract shape animation obtained by processing of the speech signal is developed, along with a graphical user interface (GUI) with separate displays for the teacher's and student's speech signals. The system has been developed for short voiced utterances and is to be extended further for other sounds.

The next section provides a review of aids for speech training. The system for dynamic display of vocal tract shape, involving speech analysis for estimation of vocal tract shape, graphics for animation, and the GUI, is described in the third section. Results are presented in the fourth section, followed by conclusions in the last section.

II. SPEECH TRAINING AIDS

Park *et al*. [3] developed a speech training system for analyzing the speech signal and displaying the vocal tract shape and speech parameters like intensity, fundamental frequency, nasality, and log spectra from the speech signal. Mahdi [4] reported a system for visualization of vocal tract shape using a mid-sagittal view of the head, and for displaying sound intensity, pitch, and first three formants, for vowel utterances.

Massaro and Light [5] used a 3D animated talking head model "Baldi" as a virtual articulation teacher. It could be transformed to highlight particular articulatory features in prerecorded animations, and also gave operating instructions to the children. It has been reported to be effective with hearing impaired children. However, it does not provide feedback of one's own articulatory efforts. Engwall *et al.* [7] have reported the speech training system "ARticulator TUtor Project (ARTUR)", which utilises 3D animations of the articulatory organs. They reported that use of the system with a human instructor performing articulatory inversion and providing pronunciation correction instructions yielded positive results with the test subjects. Kroger *et al*. [8] tested 2D and 3D animation models of the vocal tract shape for speech training. While both gave positive results, the 3D model did not lead to higher visual perception rates than the 2D model. Rathinavelu and Thiagarajan [9] reported the development and use of "Computer Aided Articulatory Tutor (CAAT)" for training hearing impaired children. The system uses a 3D model of the vocal tract shape with tongue movements derived from MRI data. The results showed an improvement in articulation of the alveolar sounds.

The research is supported by the National Program on Perception Engineering Phase II, sponsored by the Department of Electronics & Information Technology, MCIT, Government of India.

978-1-5090-2361-5/16/$31.00 ©2016 IEEE

Proc. 22th National Conference on Communications 2016 (NCC 2016), Guwahati, Mar. 4 - 6, 2016, paper no. 1570220186

Some of the commercially available computer based speech training systems include the "Video Voice Training System" [10], "SpeechViewer III" [11], and "Box-of-Tricks" [12]. They display the acoustic parameters of speech, such as speech intensity, voicing, timing, pitch, and spectral features and incorporate a number of interactive game formats and other graphics for motivating children in using the system.

A speech training system should be useful as an aid during speech therapy sessions, and should also be helpful as a self-learning aid for the children. Eriksson *et al*. [13] examined the design requirements for such a system, in relation to human computer interaction (HCI), pedagogy, and visual models. They have suggested that the display should be improvised for the children with minimal level of details. The tongue movements should be shown with reference to palate and jaw and important articulatory features of the pronunciation should be highlighted. The amount of feedback given should vary with the level of learning. The cues may be enhanced with help of sign language and/or complimentary text.

Based on the literature on use of vocal tract shape for corrective feedback to aid speech acquisition and training, a review of the available speech training aids, and design recommendations in [13], a system for a dynamic display of the vocal tract shape for feedback of articulatory efforts as a variable rate animation is developed and described in the next section.

III. System for Display of Vocal Tract Shape

The system developed for dynamic display of vocal tract shape for use as a speech training aid has three components: (i) speech analysis for estimation of vocal tract shape, (ii) vocal tract shape animation, and (iii) graphical user interface (GUI) for speech training. The GUI and the backend are developed using MATLAB.

A. Speech Analysis for Estimation of Vocal Tract Shape

For estimating the vocal tract shape, Wakita [14] proposed a speech processing technique based on linear predictive coding (LPC). In this technique, the vocal tract is modeled as an all-pole filter driven by an impulse train. The vocal tract is also modeled as a lossless acoustic tube having a set of equal-length sections of varying cross-section areas. Equivalence of the two models is used to obtain reflection coefficients at the section interfaces in the acoustic tube and these are converted to the ratios of the areas on the two sides. The ratios are scaled assuming a constant area at the glottis end of the tract to obtain the vocal tract shape. The speech production model used in this method is considered to be valid for sounds produced with glottal excitation and not involving the nasal cavity. Hence it is useful for obtaining shape of the oral cavity during the production of vowels, diphthongs, semivowels, and vowel-to-stop and stop-to-vowel transition segments, but not for that of nasalized vowels, nasals, and fricatives, and stop closure segments.

The vocal tract shapes estimated during voiced segments using Wakita's method show variations with position of the analysis frame. To reduce the variation and improve the consistency of the vocal tract shape estimation, Nataraj *et al*. [15] reported a method which selects the frame positions having the minimum windowed energy index, which is the ratio of the windowed signal energy to the frame energy. Vocal tract shapes estimated from Wakita's method are not satisfactory during stop closure segments due to lack of energy and spectral information. Pandey and Shah [16] reported a technique for estimating the vocal tract shape during the stop closures of vowel-consonant-vowel (VCV) utterance by using a bivariate surface model fitted on the vocal tract shapes during the transition segments preceding and following the stop closure. The estimated places of closure for VCV utterances of type /ΛCa/, involving stop consonants /b/, /d/, and /g/ were reported to have a good match with those obtained from X-Ray Microbeam database.

The present system has been developed for short voiced utterances which may consist of vowels, diphthongs, and semivowels. The speech is acquired at a sampling frequency of 10 kHz. End-point detection is carried out on the recorded or stored sounds and the signal between the two end points is used to display the waveform and spectrogram, and for further analysis for vocal tract shape estimation. The time axis for the display is normalized with reference to the duration of the sound.

The system uses vocal tract shape estimation by LPC analysis [14] as modified in [15], with analysis window length of twice the average pitch period and 12th order LPC. This order provides shape estimates which are comparable to those obtained from MRI databases. A higher order LPC analysis generally results in low area values at back of the vocal tract [17]. The LPC analysis provides 12 section areas for each analysis frame. For each section, linear interpolation is applied on the area values for the frames with minimum energy index to obtain area values at 5 ms intervals. The set of 12 section area values for frames at 5 ms intervals thus obtained are used for generating the animation as described in the next subsection. We assume the oral cavity to be a concatenation of cylindrical sections. Square root of the area value of a section, which is proportional to the section diameter, is used as the estimate of the amount of opening of the oral cavity at the corresponding position.

The oral cavity shape in the form of 12 values of its opening, as obtained by LPC analysis, is too coarse a representation for clearly indicating the place of articulation in the graphical display, particularly its variation across the frames. It was found that it needs to be converted to a representation with at least 20 values and we examined the use of several techniques for this purpose. Polynomial curve fitting, cubic spline interpolation, and piecewise cubic Hermite polynomial interpolation were found to be not suitable. Use of a least-squares approximation based B-spline fitting with 4 knots [18] resulted in smooth representations correctly indicating the place of articulation without undue ripples in the shape. Hence, it is applied on the 12 estimated opening values to get the values corresponding to 20 uniformly placed points along the length of the oral cavity.

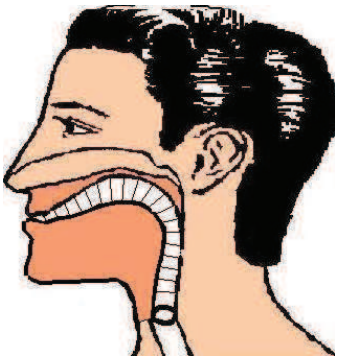

Fig. 1. Mid-sagittal view of vocal tract, with tongue in neutral position.

For validating the analysis results, the estimated oral cavity shape as a function of time is displayed as "areagram", which is a spectrogram-like 2D display of the estimated values of the oral cavity opening plotted as grey level as a function of time along x-axis and distance from the glottis along y-axis. The vocal tract shape area values estimated using LPC analysis match satisfactorily with area values estimated using MRI for vowel and vowel-semivowel-vowel utterances [16].

Indication of pitch and intensity is considered to be important for speech training, as hearing impaired children often unintentionally alter their pitch while trying to match intensity of a particular sound and vice versa [1]. Therefore, the system provides a plot of intensity and pitch as a function of time. Pitch is estimated using PRAAT software [19], invoked using a script integrated in the system.

B. Vocal Tract Shape Animation

Animation of the vocal tract shape is developed with a 2D mid-sagittal view of the head, as shown in Fig.1. A base image is created consisting of the upper jaw, palate, and neck portion. The oral cavity is shown as an area between two curves. The upper curve represents the upper surface of the oral cavity, corresponding to the positions of the upper lip, upper teeth, and palate. The lower curve represents the lower surface of the oral cavity, corresponding to the positions of the lower lip, lower teeth, and tongue. The upper curve remains fixed and the variation in the oral cavity shape is indicated by variation in the lower curve, without separately indicating the movement of different articulators (tongue, teeth, and lips).

For each frame, the image is generated from the estimated values of the oral cavity opening as described earlier. The first 16 values are used in updating the shape, as these approximately correspond to the segment of the oral cavity, the configuration of which varies during speech production. The graphical representation of the oral cavity configuration is shown in Fig. 2. Fig. 2(a) shows a neutral configuration with the upper and lower curves marked as U and L, respectively. An axial curve A is obtained as an axis of symmetry for the area between U and L, such that the curve approximately bisects the normals to it, using the method as reported in [20]. Normals to the axial curve at 16 equidistant points are used as fixed grid lines for generating the lower curve in accordance with the estimated values of the opening at these points. Fig. 2(b) shows the configuration corresponding to vowel /a/, as an example.

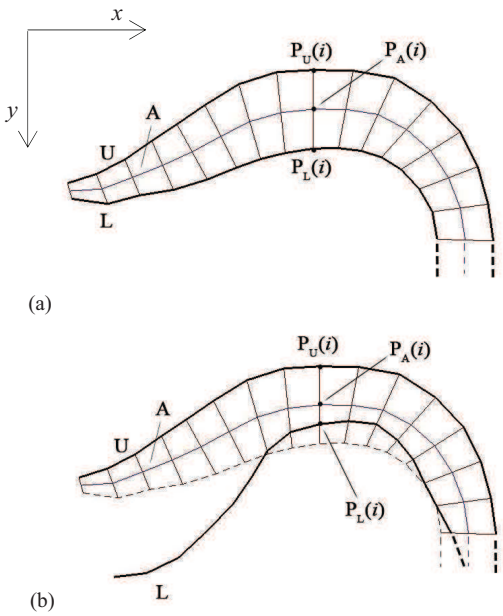

Fig. 2. Oral cavity displayed as area between 2 curves, segmented into 16 sections: (a) the neutral configuration, with the axial line marked in blue and the grid lines as normals to the axial curve, (b) configuration corresponding to the vowel /a/, obtained by redrawing the lower curve.

The estimated opening values are labeled as $h(i)$, i = 1, 16 with 1 representing the front point and 16 as the back point of the oral cavity. The $h(i)$ values are scaled to the number of pixels $d(i)$ for use in the animation graphics, as

$$d(i) = (h(i)/h_{max})\, d_{max} \tag{1}$$

where h_{max} is the maximum possible opening empirically estimated from the LPC analysis of multiple utterances and d_{max} is the maximum possible opening as set for the display. The point $P_L(i)$ on the lower curve is obtained along the corresponding grid line at the distance $d(i)$ from the upper curve. The coordinates $x_L(i)$ and $y_L(i)$ of the point are calculated using the slope $m(i)$ of the grid line passing through the point $P_A(i)$ on the axial curve. The coordinates of $P_A(i)$ are given as $x_A(i)$ and $y_A(i)$, and those of the corresponding point $P_U(i)$ on the upper curve are given as $x_U(i)$ and $y_U(i)$. The point coordinates are calculated using the direction notation as shown in Fig. 2.

The coordinates of $P_L(i)$ are calculated by finding the distances of $P_L(i)$ from $P_U(i)$ along the two axes as the following:

$$\Delta x = d(i)/(1+m(i)^2)^{1/2} \tag{2}$$

$$\Delta y = m(i)d(i)/(1+m(i)^2)^{1/2} \tag{3}$$

$$x_L(i) = \begin{cases} x_U(i) - \Delta x(i), & m(i) < 0 \\ x_U(i) + \Delta x(i), & m(i) > 0,\ x_A(i) > x_U(i) \\ x_U(i) - \Delta x(i), & m(i) > 0,\ x_A(i) < x_U(i) \end{cases} \tag{4}$$

$$y_L(i) = \begin{cases} y_U(i) - \Delta y(i), & m(i) < 0 \\ y_U(i) + \Delta y(i), & m(i) > 0,\ x_A(i) > x_U(i) \\ y_U(i) - \Delta y(i), & m(i) > 0,\ x_A(i) < x_U(i) \end{cases} \tag{5}$$

The calculated points are joined by line segments to form the lower curve. The lower jaw is displayed as moving as a single unit, with its left tip marked by $P_L(1)$.

Place of maximum constriction in the vocal tract shape indicates the major articulatory effort in production of a phoneme. Its position and the amount of opening at this position are the most important inputs for speech training and emphasizing them in the display is likely to improve the effectiveness of the aid. Hence, an option is provided to display the lower curve modelled as a triangle with the end points forming the left and right vertices, and the place of maximum constriction indicated by the vertex between the two end points.

For reducing perceptual load during speech training, only the organs important in the articulatory process are shown in the mid-sagittal sectional view. The system provides an option to choose between two types of vocal tract shapes: the estimated shape and the shape emphasizing the maximum constriction, as shown in Fig. 3. For animation, each frame of the estimated vocal tract shape is displayed after a specified delay, for creating animation at a controlled pace.

C. GUI for Speech Training Aid

The system integrates both acoustic and articulatory feedback of speech. The display screen has two panels; one each for the student and the teacher, as shown in Fig.4. The GUI for the system displays speech waveform, spectrogram, and areagram, on each panel derived from the corresponding speech signal. The duration of the non-silent portion of the

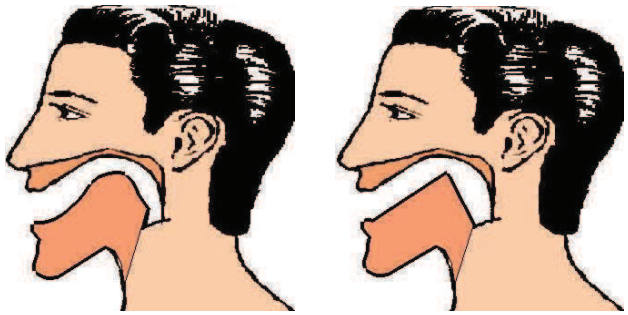

Fig. 3. Vocal tract shape obtained through speech processing (left) and after modelling it as a triangle (right)

recording (in s) is displayed below the button panel on each side. The waveform, spectrogram, areagram, pitch and intensity plots have a common time axis.

For each panel, estimated vocal tract shape for animation is generated as described in the previous section. The intensity and pitch are displayed in a single graph with different colors. A progress bar, normalized to the duration of the sound, indicates the position of the segment being displayed in the animation.

Both the animations can be played together or separately, and can be replayed as per convenience, using the six buttons at the center. "Left" and "Right" buttons replay the vocal tract animation in the left and the right panels, respectively. "Both" plays both vocal tract animations simultaneously. "Scroll Bar" controls the animation rate by varying the delay between

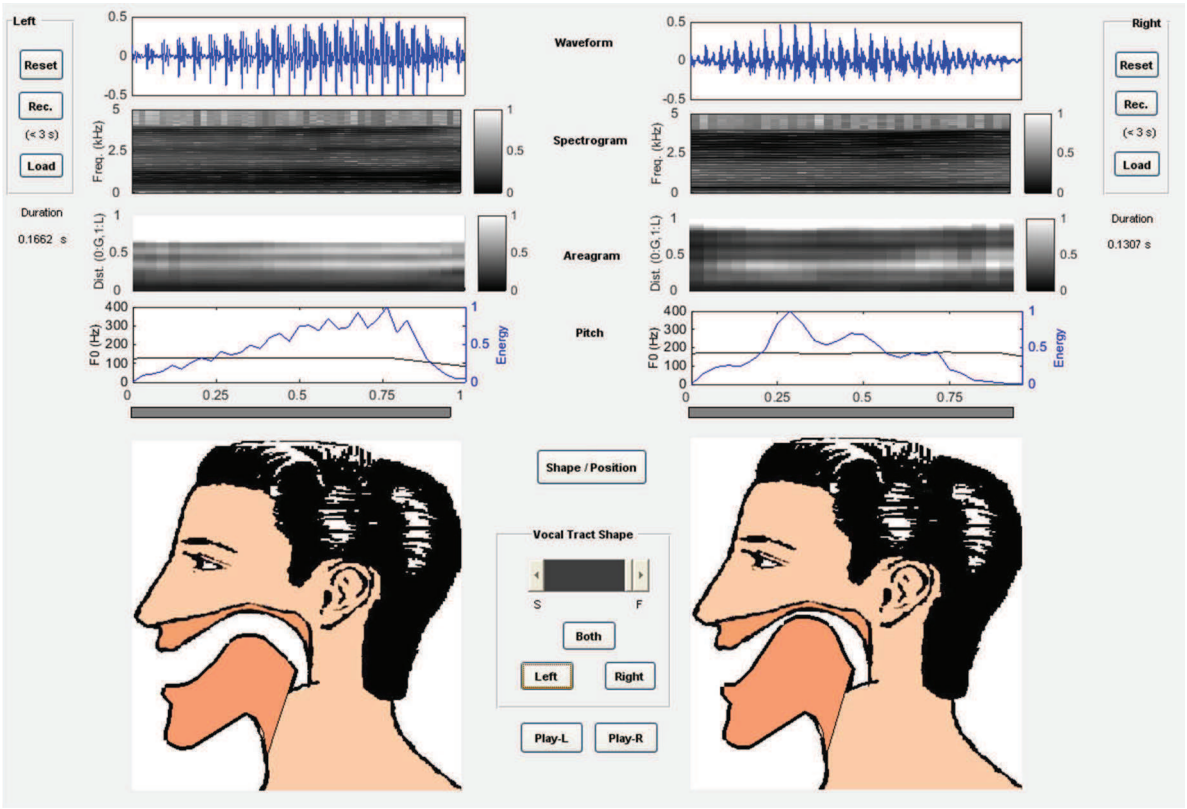

Fig. 4. GUI for displaying the speech parameters and vocal tract shape, one each for the student and the teacher.

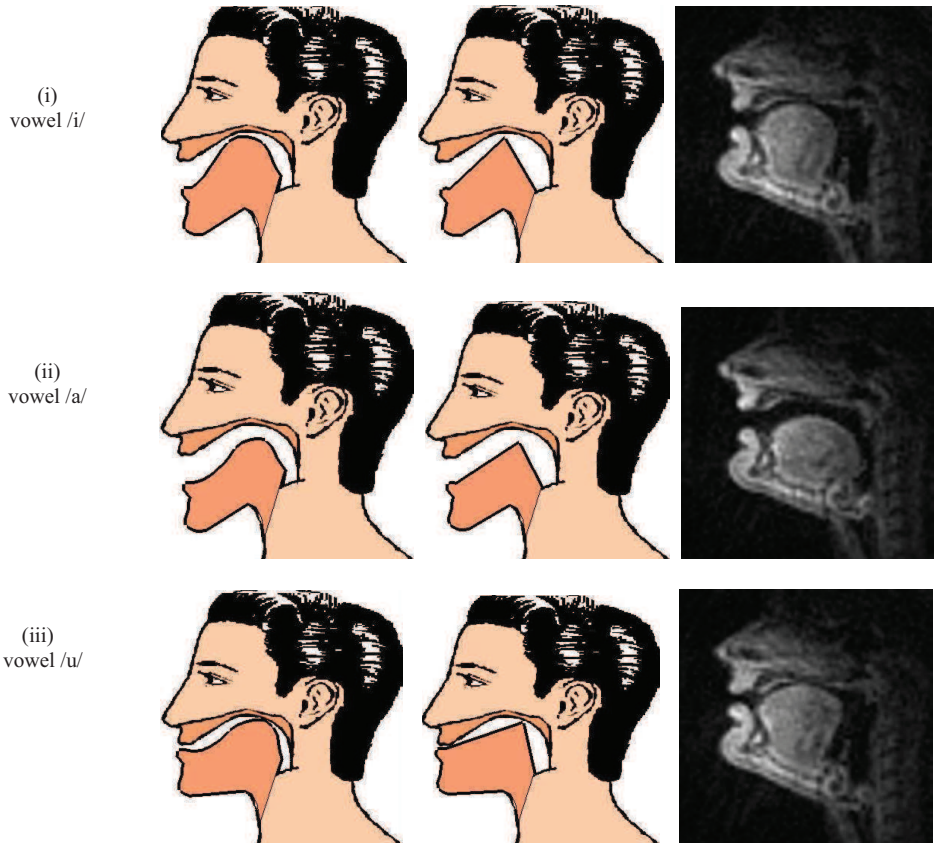

Fig. 5.Vocal tract shapes for three vowels: obtained using LPC analysis based speech processing (left), and after modelling as a triangle (middle), and MRI scan (right).

successive frames. "Shape/Position" toggles between animation of the estimated shape and the shape emphasizing the maximum constriction. "Play-L" and "Play-R" buttons play the speech utterance of the left and the right panel as audio output, respectively.

The GUI can be used for analysis and display of freshly recorded or pre-recorded sounds. On either side of the display, three buttons are provided. "Record" starts recording (after displaying a countdown, as an alert indicator) the input audio for 3s. The sound is played back after the recording of the specified length, and is analyzed to get the parameters for the display. The recorded audio is stored with a date and timestamp for further use in training. "Load" allows the user to select and generate displays for pre-recorded audio. "Reset" clears all the displays generated on the corresponding panel.

IV. Test Results

The system has been tested for recordings from three speakers (two male, one female) for a set of vowels and vowel-semivowel-vowel utterances. Screen recordings of the displays are available at [21].

Fig.5 shows an example of estimated vocal tract shapes and shapes emphasizing maximum constriction, for three vowel utterances from a male speaker. The figure also shows MRI scan images for the corresponding vowels from the MRI database [22]. The vowel /i/ is high-front vowel produced by raising the front part of the tongue to the palate. The vowel /a/ is low-mid vowel, produced with a relatively neutral position of the tongue. The vowel /u/ is high-back vowel produced by raising the posterior end of the tongue. Vocal tract shapes generated by the system match satisfactorily with those obtained from MRI scans, and highlight the important features of the articulation i.e. place of maximum constriction and the opening at that place.

V. Conclusion

A speech training system with speech analysis for obtaining vocal tract shape from acquired speech utterances and graphics for dynamic display of vocal tract shape with separate panels for two speakers has been developed. The system enables the teacher to provide feedback to the student and store individual practice sounds for maintaining a record of the progress. It also enables the student to practice with the stored recordings.

The present system is useable for utterances involving vowels, diphthongs, and semivowels. The next version will be developed, using the processing technique as reported in [16], for speech training of vowel-consonant-vowel utterances for oral stops. Pitch, intensity, and vocal tract shape animation serve as the speech training parameters and waveform, spectrogram, and areagram serve as tools to validate the output. Both the animation and the GUI have been designed so that the student is not overwhelmed with data. An option of triangular

profile for the lower surface of the oral cavity is provided to keep the focus on the place of articulation. The variable rate of animation enables the rate to be increased with the improving proficiency of the student.

Many aspects of the speech training system, including the relevance of different sections of the display in speech training, modes of providing feedback, method employed to emphasize the articulatory effort, orientation and positioning of the two vocal tract animations, orientation of vocal tract shape, and the most appropriate color combination are being discussed with user groups. An option of animation displaying only the oral cavity (and not the oral cavity as a part of the head) may be helpful in better display of the information for articulatory control. Using the oral cavity shape information for animation of the movement of the actual articulators is expected to significantly improve the effectiveness of the system. The graphics framework of the system needs to be enhanced by incorporating inputs from user groups and speech therapists before its evaluation for speech training of children with hearing impairment.

ACKNOWLEDGMENT

The authors are thankful to Mr. Vishal Mane (Media Lab Asia, New Delhi) for insightful discussions and suggestions for the display format.

REFERENCES

[1] R. S. Nickerson and K. N. Stevens, “Teaching speech to the deaf: can a computer help?,” IEEE Trans. Audio Electroacoust., vol. 21, no. 5, pp. 445–455, 1973.

[2] P. Badin, “Visual articulatory feedback for phonetic correction in second language learning,” in Proc. Speech Comm. Assoc. Workshop Second Language Learning Studies: Acquisition, Learning, Education and Technology, 2010,Tokyo, Japan, pp.1-10.

[3] S. H. Park, D. J. Kim, J. H. Lee, and T. S. Yoon, “Integrated speech training system for hearing impaired,” IEEE Trans. Rehabil. Eng., vol. 2, no. 4, pp. 189–196, 1994.

[4] A. E. Mahdi, “Visualisation of the vocal-tract shape for a computer-basedspeech training system for the hearing-impaired,” Open Elect.Electron. Eng. J., vol. 2, pp. 27-32, 2008.

[5] D. 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, no. 2, p. 304, 2004.

[6] P. Badin, F. Elisei, G. Bailly, and Y. Tarabalka, “An audiovisual talking head for augmented speech generation: Models and animations based on a real speaker's articulatory data,” in Proc. AMDO 2008, Mallorca Spain, pp. 132-143.

[7] O. Engwall, O. Balter, A. M. Oster, and H. Kjellstrom, “Designing the user interface of the computer-based speech training system ARTUR based on early user tests,” Behaviour and Information Technology, vol. 25, pp. 353-365, 2006.

[8] 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 2008, Brisbane, Australia, pp. 2639–2642.

[9] A. Rathinavelu and H. Thiagarajan, “Computer Aided Articulatory Tutor: A scientific study”, Int. Sci. J. Computing, vol. 6, no. 3, pp. 100-105, 2007

[10] Video Voice Speech Training System, (Micro Video Corp., Ann Arbor, Michigan, 2003). [Online]. Available: www.videovoice.com. (Last accessed: Aug, 2015)

[11] Speech Viewer III, (Synapse Adaptive, San Rafael, Cal., 2000). [Online]. Available: www.synapseadaptive.com/edmark/prod/sv3. (Last accessed:Sep, 2015).

[12] G. Vlahavas, “Anna Sfakianaki - Box of Tricks”,2015. [Online]. Available: www.enl.auth.gr/phonlab/box_of_tricks.html (Last accessed: Sep, 2015).

[13] 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 2005, Patras, Greece, pp. 483-486.

[14] 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.

[15] K. S. Nataraj, Jagbandhu, P. C. Pandey, and M. S. Shah, “Improving the consistency of vocal tract shape estimation,” in Proc. 17th National Conf. Commun. (NCC 2011), Bangalore, India, 2011, paper SpPrII.4.

[16] 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.

[17] M. S. Shah, “Estimation of place of articulation during stop closures of vowel-consonant-vowel syllables,” Ph.D. thesis, Dept. of Elect. Engg., IIT Bombay, India, 2008.

[18] C. E. DeBoor and J. R. Rice, "Least square cubic spline approximation: II Variable knots," Dep. Comput. Sci., Purdue Univ., Tech. Rep. 21, 1968.

[19] P. Boersma and D. Weenink, “Praat: doing phonetics by computer”, 2015. [Computer program]. Version 5.4.20, Available: www.praat.org(Last accessed: Sep, 2015)

[20] K. S. Nataraj and P. C. Pandey, “Place of articulation from direct imaging for validation of its estimation from speech analysis,” in Proc. 5th National Conf. Comput. Vision, Pattern Recognition, Image Process. and Graph. (NCVPRIPG 2015), Patna, India, 2015, paper no. 88.

[21] R. Jain, “Screen video clips of VTS”, 2015 [Online]. Available:https://www.ee.iitb.ac.in/~spilab/material/rahul_jain/ncc2016.

[22] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammart, J. Kim, S. Lee, K. S. Nayak, Y. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. K. Ghosh, A. Katsamanis and M. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research," J. Acoust. Soc. Am. 136, pp. 1307–1311, 2014.