White Paper

Visual Speech Training Software for the Hearing Impaired: Current Status and Way Forward

Prepared under the project "Visual Speech Training Software for the Hearing Impaired", Sponsoring Agency: DEITY, MCIT, Govt. of India, Chief Investigator: Bhavani Prasad Yerrapalli, Media Lab Asia, Co-Investigator: Prof. P.C. Pandey, Department of EE, IIT Bombay

Revised: 14 September 2015

Vishal Mane
Sr. Research Scientist, <vishal.mane@medialabasia.in>

Not-for-Profit, Section 25 Company set up by Department of Electronics and IT, Ministry of Communications and IT, Government of India

1. Introduction

Hearing loss (HL) and deafness are global issues which affect at least 278 million persons worldwide and two-thirds of them live in developing countries. In India, persons using hearing aids have been treated as persons with disability (PwD) in the Census 2011 and their number is 50, 71,007 (26, 77,544 males, 23, 93,463 females).

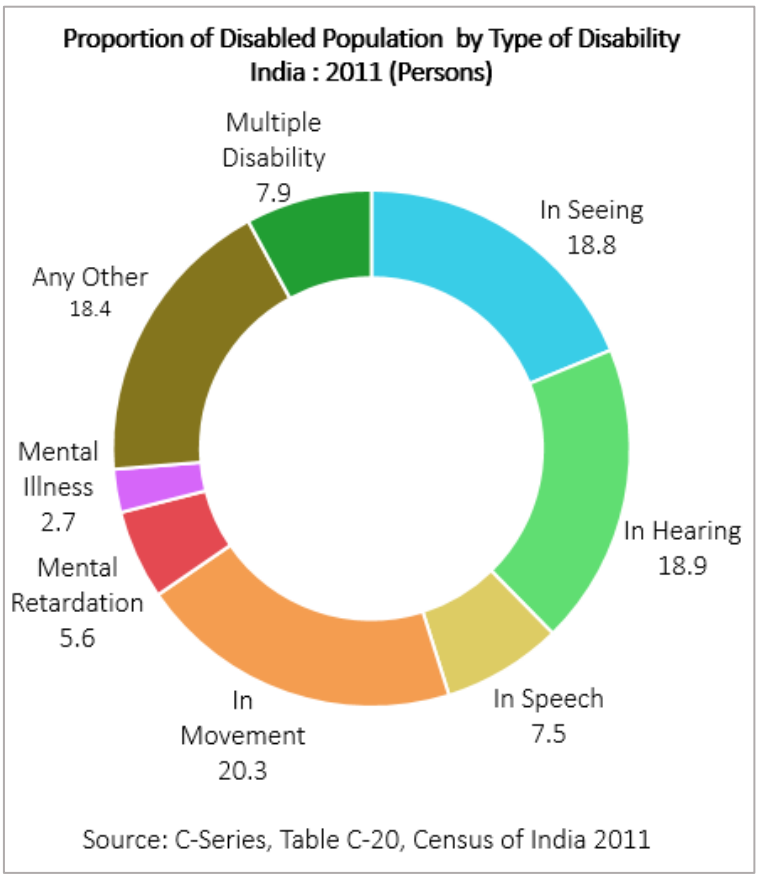

Figure 1: Disabled Population in India -2011

It's difficult for children born with hearing disabilities, or those who develop impairments at an early stage, to acquire speech. Thus they rely on visual feedback of the auditory cues. Lip reading and feedback from a mirror are effective in this regard, but they do not help in deciphering utterances from internal articulations. While the involvement of a speech therapist is crucial for speech acquisition, computer based speech training (CBST) systems have been found to motivate children in practicing by providing feedback of their progress. These systems employ various mechanisms for providing training and feedback, and can also help in pronunciation training for second language learning.

2. Feedback from Users and Experts

Hearing impaired children have good drawing skills to express their feelings. Mobile SMS and sign communication mediums are very useful to them. As students, they are more comfortable while using computers and visual displays like graphs, colors, vibrations and interactive animations. An inhalation and exhalation pattern through paper displacement/ flows has been used for teaching. Intonation patters i.e. high pitch, low pitch, exhalation, bar rising and falling etc. has been used for teaching the students.

It has been recommended that graphical display and video techniques would be useful for teaching as well as moving pictures, tongue & jaw movement,

graphics/images, vocal track are useful patterns for visual speech training. It has been noted that there are difficulties for teaching vowels, consonants, diphthongs, timing patters, intonation patterns, articulatory patterns and nasalization to hearing impaired. A visual feedback can be given though animated movement/placement of the lip, tongue, jaw and vocal tract. Effective tool need to be developed to remember specking words/sentence/vowels/consonants/intonations by the student/persons with hearing impaired. Self-learning tool through vibrations can be useful for hearing impaired.

3. A Survey of Visual Speech Training Tools

A) Box of Tricks [1], developed as a part of the SPECO Project, is a CBST system that provides real-time visual display of various acoustic components of speech, for children with hearing impairments. It employs child friendly visualizations for displaying the various parts of speech; e.g. each phoneme has a cochleogram and a figure representative of the particular sound. A number of mini games emphasize producing various sound metrics as close as possible to required results; such as energy-vs-time curves, pitch, spectrogram differences. It utilizes a product oriented approach instead of a process oriented approach. During speech acquisition, normal-hearing children do not usually get instructions on how to move or place their organs to produce the sound. Traditional speech therapy on the other hand generally adopts a process oriented approach; a speech therapist instructs how to utilize a particular organ for producing the required sound.

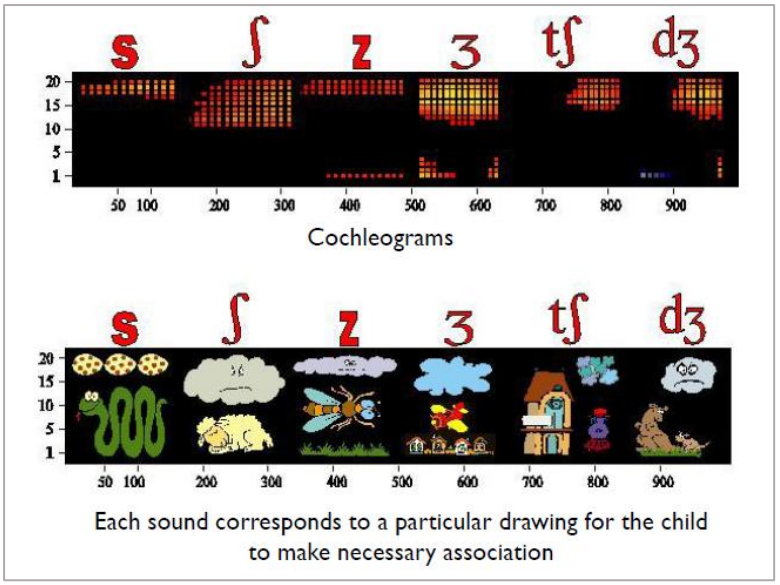

Figure 2: Box of Tricks - SPECO Project

The database for SPECO has two components – language independent editor and language dependent speech database, which can vary from language to language, which itself, is derived from two components – reference speaker and multi-speaker database. All of this is constructed specifically for the project, from children in the age group of 7-11 years. SPECO relies on the principle that the user can process low level speech metrics and utilize it using their high level information processing capacity to produce sound instead of giving articulation instructions.

B) *Baldi* is a CBST System developed by Light and Massaro [2]. It has an animated talking model of the human face. It works on the principle that speech and its acquisition are multimodal phenomenon and the human face provides information critical for communication. The facial animation works on a texture mapped wireframe model, and realistic speech is obtained by animating the proper facial targets. *Baldi* can present views not visible from the outside of the face, and also has the ability to vary transparency so that we can look inside of the vocal tract, and a cutaway, thee mid-sagittal view of the vocal tract is also possible. The movements of the tongue, hard palate, teeth, and other internal articulatory organs have been trained by Electro-palatography (EPG) data and Ultrasound measurements of the upper tongue. The head can be rotated & viewed, aiding in a back-of-the-head view. Moreover, the system can employ additional cues, than normally used, that can help in training, such as visual indication of vocal cord vibration and turbulent airflow, helping distinguish between voiced and voiceless distinctions. Variation in speeds of articulation is also possible.

Figure 3: CBST System - Light and Massaro

The system can be used to provide feedback by creating a happy and sad face. It also pronounces all the instructions and can ask pre-determined questions to the user, with proper pronunciation animations. The system can be altered to match various targeted facial models. It is accompanied by textual equivalent of the spoken content. Both visual and auditory speech components can be controlled individually and manipulated, thus creating possibility of customized changes to enhance informative characteristics of speech. *Baldi* can be configured incorporated as part of different systems for different means. For example, Language Tutor and Wizard employs *Baldi* also as an instructor alongside mini games aimed at teaching children pronunciation, spelling and recognition of vegetables from photos. *Baldi,* however, does not relate speaker's own pronunciation to the actual pronunciation, thus providing little opportunities for self-correction.

C) ARticulator TUtoR Project (ARTUR) [7], [8], uses 3D animation of the articulatory organs for providing feedback on the user's pronunciation and deviation from the desired pronunciation. It relies on the principle that self-correction is an important factor in speech acquisition, and the user should know how to alter his/her articulation.

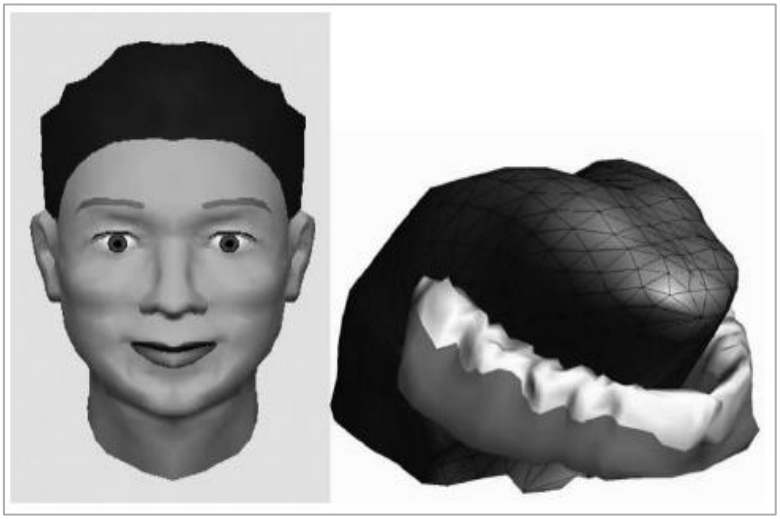

Figure 4: ARticulator TUtoR - KTH

Determining the level of acceptability of the pronunciation and its classification as correct or incorrect is a major concern here, and it requires a theoretical framework of preconceived types of errors possible and statistical training of the system with both correct and deviant pronunciations. The most complicated step in the process is articulatory inversion, i.e. creating the animation from the user's speech, as it is a process with many-to-one mapping of acoustics to articulations. Facial data are also used to improve the process due to significant correlation between face and tongue positions.

ARTUR employs both 3D animations of a face and a tongue & jaw model for displaying articulations, which were developed from Electromagnetic Articulography (EMA), EPG and real-time MRI measurements of a single subject. Adaptations can be made to the display according to preliminary data of the user, such as gender and age, using statistical analysis of MRI data.

It is important that the output is relevant, motivating and comprehensible to the children. A Wizard of Oz study was conducted to improve the Human Computer Interaction (HCI) aspects of the display, as well as collecting training data. The Wizard i.e. a human subject posing as the computer, worked to provide mispronunciation detection, articulatory inversion as well as determining the pre-generated feedback. While the results are positive and motivating, the critical issue is to process all the above in real time, without human intervention in the process.

D) Vocal Tract Display [3] developed at SPI Lab, IIT Bombay. Direct visualizations pose problems in feasibility and real time-processing, thus creating the need for using indirect methods of vocal tract shape estimation. The system incorporates a linear predictive coding (LPC) based method by Wakita to estimate the vocal tract from the speech signal. Wakita's method works for vowels, semivowels and diphthongs but not for vowel-consonant-vowel (VCV) utterances because of low energy of the signal. To resolve this, a method proposed by Pandey and Shah [14], is used which uses bivariate polynomial modeling on VC & CV transition area values. The closure segment VCV utterances were compared to values from the X-Ray Micro-Beam database. The smoothening of the vocal tract estimates from the above methods is done using cubic spline interpolation. Lip area scaling, derived from video of the utterances, is used where the LPC method is insufficient to determine the place of articulation. Graphics are developed from the values estimated from the LPC method.

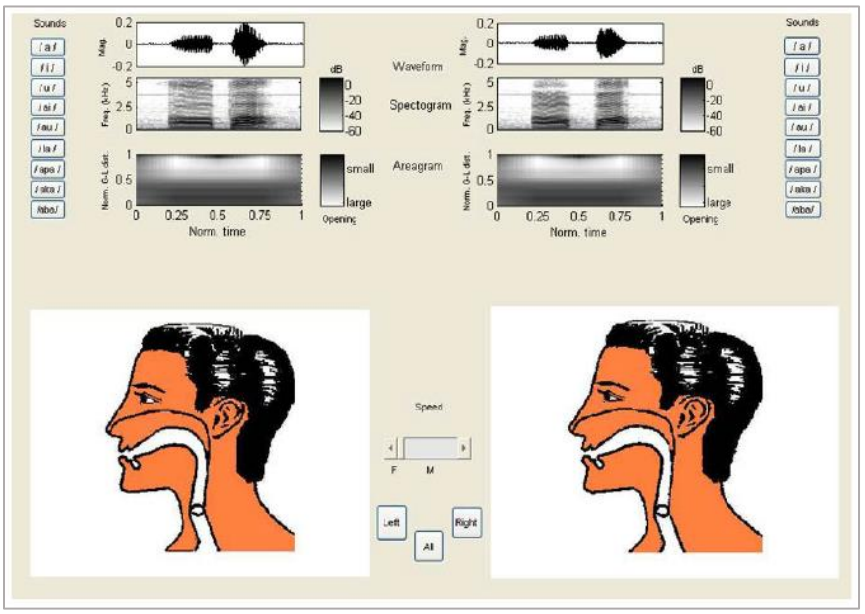

Figure 5: Vocal Tract Display - SPI Lab IIT Bombay

The vocal tract animation assumes the upper jaw fixed and the lower jaw movable. The vocal tract is in a 2-D mid-sagittal view, which varies according to the sound produced. No separate organs are shown. The interface has options to select the speech signal to be processed and displayed. Two vocal tract shapes can be animated simultaneously, one for the student and the other for the teacher, as shown in Fig.1. The delay between frames can be adjusted according to need using a scroll bar and option to play both animations together and separately also exists.

4. Way Forward

A visual speech training application will be developed for providing feedback of articulatory efforts of the trainee and a comparison with those of the trainer as the target, with more accurate estimation of the parameters related to vowel and consonants in speech utterances than the currently available speech training software and a display and user interface designed by involving the users as active collaborators.

The aim is to develop the software for a Visual Speech Training Aid, as a desktop or tablet based application. It will use the speech processing techniques as the backend engine to get information related to speech production and provide a visual feedback of the information to the hearing impaired persons to help them in speech and language development. The software will be based on integration of (i) state-of-the-art techniques for estimating the articulatory efforts of the hearing impaired person undergoing speech training and (ii) information display and user interface designed by involving speech teachers and therapists as active collaborators and stake holders.

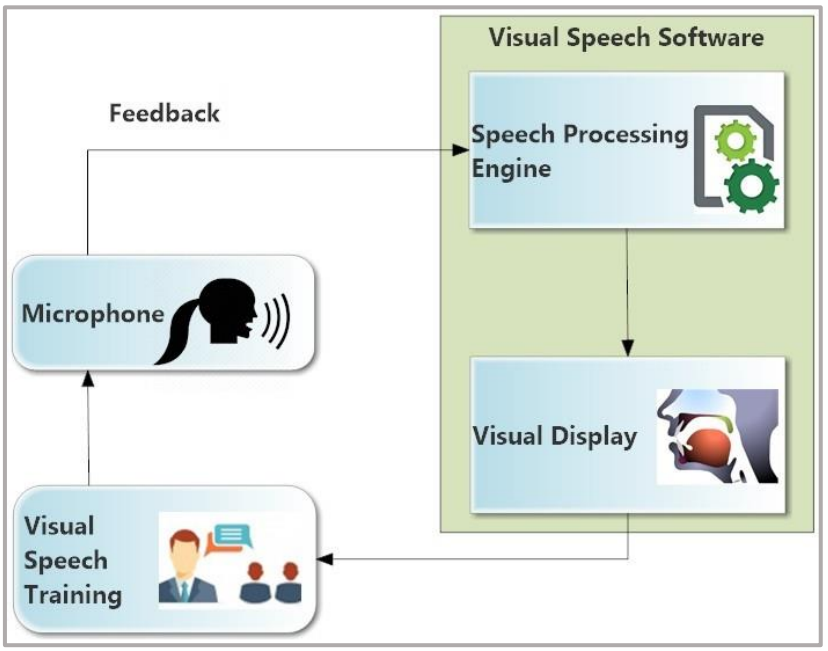

Figure 6 'Visual Speech Training Software'

References

1. http://www.censusindia.gov.in
2. http://punarbhava.in/index.php?option=com_content&view=article&id=1463&Itemid=758
3. Rahul Jain, Prof. P. C. Pandey, Dept. of Electrical Engineering, IIT Bombay, "Articulatory visual feedback for speech acquisition and training," EE 451 SRE Seminar Report, EE Dept, IIT Bombay, 15 April 2015
4. S. Nilashree Wankhede, Dept. Of Elect. and Telecom., Fr.C. Rodrigues Institute of Technology, Navi Mumbai, "Designing visual speech training aids for hearing impaired children," International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 3, Issue 4, April 2014
5. Sheila R. Pratt, Ph.D., Using Electropalatographic Feedback to Treat the Speech of A Child with Severe-to-Profound Hearing Loss, SLP- ABA Issue 2.2, 2007