Symposium on Frontiers of Research on Speech and Music, IIT Kanpur, Feb. 15-16, 2003, pp. 121-124.

AREAGRAM DISPLAY FOR INVESTIGATING THE ESTIMATION OF VOCAL TRACT SHAPE FOR A SPEECH TRAINING AID

Milind S. Shah and Prem C. Pandey

Department of Electrical Engineering
Indian Institute of Technology, Bombay
Powai, Mumbai 400 076, India
{milind, pcpandey}@ee.iitb.ac.in

ABSTRACT

The display of intensity, pitch, and vocal tract shape is considered to be helpful in speech training of the hearing impaired. A speech analysis package is developed in MATLAB for displaying speech waveforms, pitch and energy contours, spectrogram, and areagram (a two-dimensional plot of cross-sectional area of vocal tract as a function of time and position along the tract length). While vocal tract shape estimation works satisfactorily for vowels, during stop closures, the place of closure can not be estimated due to very low signal energy. There is a need to investigate methods for predicting vocal tract shape during stop closure from the shapes estimated on either side of the closure.

1. INTRODUCTION

Lack of auditory feedback in the hearing impaired results in failure to produce intelligible speech. Speech training, in such cases, can be assisted by other forms of feedback. Lip-reading is the simplest and most widely used method for providing visual cues, however it does not provide information on articulatory efforts inside the mouth. Efforts have been made to develop speech training aids involving spectrographic displays, display of formant tracks, voicing and pitch variations, etc [1]. In order to provide information on articulatory efforts, a number of speech training aids have been reported for displaying vocal tract shape [2-5]. The vocal tract shape display is often supplemented by voicing/pitch and energy tracks. Early training aids based on this principle displayed a simple line graph indicating the variation of vocal tract area. The more recent ones have more realistic displays of the vocal tract shape.

Formant frequencies were used to generate vocal tract shape by Ladefoged et al [2]. Authors have stated that, since the acoustic structure of a vowel is fairly well determined by the first three formants, it should be possible to recover a plausible vocal tract shape for a vowel, knowing the formants. The authors have commented that the method is useful for a limited set of vowel utterance. The system reported by Park et al [4] displays intensity, fundamental frequency, and nasality along with vocal tract shape. Fundamental frequency and nasality are detected using separate vibration sensors. Vocal tract area function from lips to glottis is found using Wakita's method [6], supported by lip-to-lip distance found from first three formant frequencies as given by [2]. Tests were carried out to train deaf children successfully, for five Korean vowels. A speech training aid, which integrates acoustic and several type of instrumentally measured articulatory data like palatography, nasal vibration, airflow and presence/absence of voicing has been reported in [3]. In another development, speech visualization system which extracts the consonantal features using neural network and creating visual images by adding all the consonantal patterns whose brightness is controlled by the strength of the extracted phonemic features has been developed [7].

In the systems reported in the literature, the verification of the vocal tract shapes have been reported for certain vowels, and despite incorporation of various interactive games, etc. they have been found to be of only limited help as speech training aid. Our objective is to investigate the dynamics of vocal tract shape estimation, particularly in vowel-consonant-vowel syllables, by displaying the vocal tract area as a function of time and distance in a spectrogram like display, identify the problem, and devise a solution in order to develop vocal tract shape display system that can help in speech training for hearing impaired persons.

2. IMPLEMENTATION

Our implementation of vocal tract shape estimation is based on reflection coefficients obtained from LPC analysis of speech signal [8]. Wakita's speech analysis model [6] is selected for the estimation of vocal tract area and Robinson's algorithm [6] for optimum inverse filtering is implemented. Order of linear predictor is chosen to be 12. The 12-section area function is interpolated to 176 points by Beizer form algorithm [9] to obtain a more realistic display. The pitch is estimated by short time autocorrelation method [8], and energy as the zeroth autocorrelation coefficient.

Difficulties have been earlier experienced in vocal tract shape estimation algorithm implementation using fixed-point real-time processing due to recursive errors in computation and dynamic range limitation [10]. For the investigations reported here, the algorithm is implemented using MATLAB with floating-point arithmetic.

In order to study the reliability of the vocal tract shape estimation during vowels, and to study the transitions at vowel-consonant boundaries, we have used "areagram", a two dimensional display of vocal tract area with time and lip to glottis distance. Time is plotted along x axis; y axis represents distance from lip to glottis. Minimal opening is represented by black and maximal opening is represented by white. This form of display is similar to that of spectrograms [8]. The speech signal is recorded with PC sound card and Goldwave software at a sampling rate of 11.025 k Sa/s. Each analysis frame consists of 256 samples, and successive frames are positioned with 50% overlap.

For processing and display, package "VTAG-1" is developed in MATLAB for displaying speech waveforms (recorded signal and selected segment), pitch and energy contours, spectrogram, and areagram. Recorded speech signal, selected speech segment, and pitch/energy contours are displayed in the left half of the screen while spectrogram and areagram for selected segment is displayed in the right half. The lower and upper range of various scales can be modified with graphical user interface menu in order to select the most appropriate dynamic range. An algorithm for lip shape estimation proposed by McAllister et al [11] is being implemented in the same package. There is provision for displaying any one pair of figures out of spectrogram, areagram, and estimated lip shape

3. RESULTS

The package was tested with natural and synthesized speech signals for consistency and validity of the estimates. Figure 1 shows the spectrogram and areagram from the analysis of synthesized vowel /*a*/, with a constant pitch. As /*a*/ is a central vowel, vocal tract is more open towards the front region than the middle region. Figure 2 shows result for natural V-C-V sequence /*apa*/. In this figure, for the vowel segment areagram results are proper. But during the stop closure, the area estimation becomes almost random.

It is to be noted that the indication of the place of closure is critical for the success of the system as a speech training aid. Thus techniques for estimating the place of closure, on the basis of the transition in the area function preceding and following the closure need to be investigated. Also, the possible

use of harmonic parameter of harmonic plus noise (HNM) model [12] needs to be investigated for the vocal tract shape estimation.

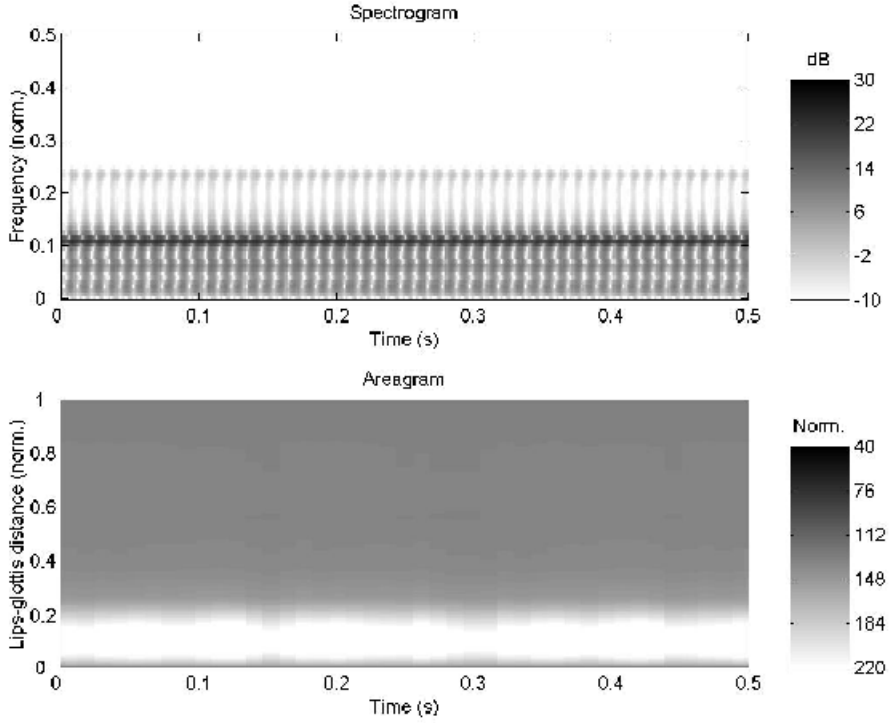

Figure 1. Analysis result for synthesized vowel/*a*/

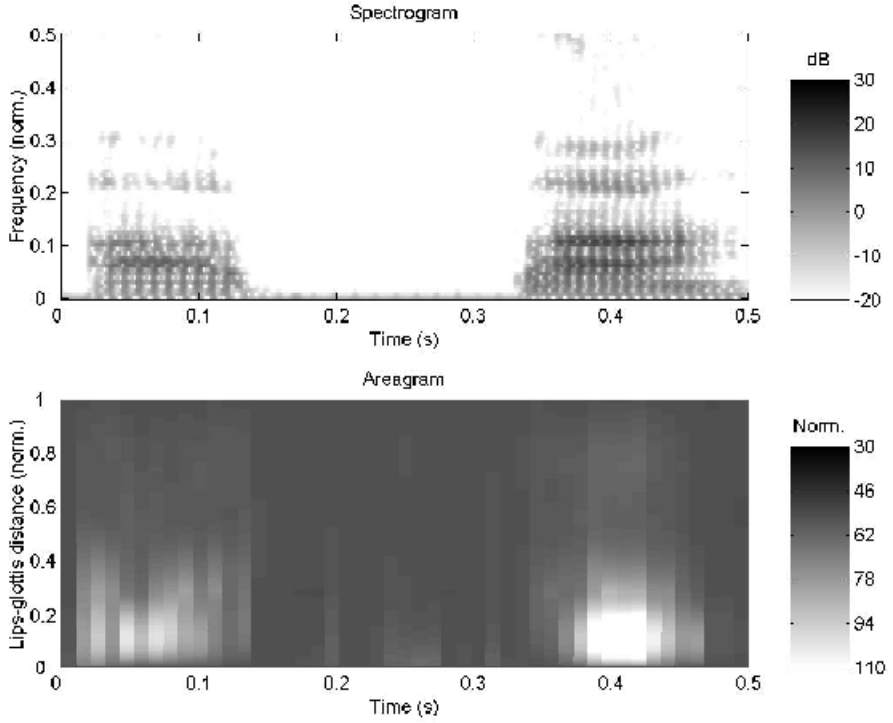

Figure 2. Analysis result for V-C-V sequence/*apa*/

4. CONCLUSION

The VTAG-1 speech analysis package works satisfactorily for the vowels and vowel sequences. With the help of areagram display, we see that for stop consonant, the estimation of area is not reliable for the closure duration, for which the energy of the speech signal is very low. There is a need to investigate methods for predicting vocal tract shape during stop closure from the shapes estimated on either side of the closure. Also, possible use of harmonic plus noise model for vocal tract shape estimation for consonant sounds needs to be investigated.

REFERENCES

[1] H. Levitt, J. M. Pickett, and R. A. Hounde, *Sensory Aids for the Hearing Impaired*. New York: IEEE Press, 1980.

[2] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice, ”Generating vocal tract shapes from formant frequencies,” *J. Acoustic Soc. Am*., vol. 64, part. 4, pp. 1027-1035, 1978.

[3] H. Javkin, N. A. Barroso, A. Das, D. Zerkle, Y. Yamda, N. Murata, H. Levitt, and K. Youndelman, ”A motivation-sustaining articulatory/acoustic speech training system for profoundly deaf children,” *Proc. ICASSP 93*, 1993, pp. 145-148.

[4] S. H. Park, D. J. Kim, J. H. Lee, and T. S. Yoon, ”Integrated speech training systems for hearing impaired,” *IEEE Trans. Rehab. Engg*., vol. 2, no. 4, pp. 189-196, 1994.

[5] D. Rossiter, D. M. Howard, and M. Downes, ”A real time LPC-based vocal tract area display for voice development,” *Journal of Voice*, vol. 8 (4), pp. 314-319, 1994.

[6] H Wakita, ”Direct estimation of the vocal-tract shape by inverse filtering of acoustic waveforms,” *IEEE Trans. Audio Electroacoust*., vol. AU 21, no. 5, pp. 417 - 427, 1973.

[7] A. Watanbe, S. Tomishige, and M. Nakatake, ”Speech visualization by integrating features for the hearing impaired,” *IEEE Trans. Speech Audio Processing*, vol. 8, no. 4, July 2000.

[8] L. R. Rabiner, and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice Hall, 1978.

[9] J. D. Foley and A. Vandam, *Fundamentals of Interactive Computer Graphics*, pp. 514-536, New York: Addison-Wesley, 1983.

[10] S. A. Kshirsagar, ”A Speech Training Aid for Hearing Impaired,” M.Tech. Dissertation, Dept. of Electrical Engg., IIT Bombay, January 1998, Guide: Dr. P. C. Pandey.

[11] F. D. McAllister, R. D. Rodman, D. L. Bitzer, and A. S. Freeman, ”Lip synchronization as an aid to the hearing impaired,” *Proc. AVIOS 97*, September 1997, pp. 233-248.

[12] Y. Stylianou, “Applying the harmonic plus noise model in concatenative speech synthesis,” *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 21-29, Jan. 2001.