Visualisation of the Vocal Tract Based on Estimation of Vocal Area Functions and Formant Frequencies

Abdulhussain E. Mahdi

Department of Electronic and Computer Engineering
University of Limerick, Limerick, Ireland
hussain.mahdi@ul.ie

Abstract

A system for visualisation of the vocal-tract shapes during vowel articulation has been designed and developed. The system generates the vocal tract configuration using a new approach based on extracting both the area functions and the formant frequencies form the acoustic speech signal. Using a linear prediction analysis, the vocal tract area functions and the first three formants are first estimated. The estimated area functions are then mapped to corresponding mid-sagittal distances and displayed as 2D vocal tract lateral graphics. The mapping process is based on a simple numerical algorithm and an accurate reference grid derived from x-rays for the pronunciation of a number English vowels uttered by different speakers. To compensate for possible errors in the estimated area functions due to variations in vocal tract length, the first two section distances are determined by the three formants. The formants are also used to adjust the rounding of the lips and the height of the jawbone. Results show high correlation with x-ray data and the PARAFAC analysis. The system could be useful as a visual sensory aid for speech training of the hearing–impaired.

1. Introduction

For hearing-impaired people, learning to speak naturally is a very difficult process. With limited auditory capability, a hearing-impaired person often lacks models of speech targets necessary to produce normal speech. In an effort to overcome this difficulty, many attempts have been made to provide a substitute for the feedback mechanism with visual speech display devices [1]. However, without any articulatory correlate, the benefits of such devices were limited. In order to produce a natural and intelligible speech, a speaker needs to know how to use the vocal organs in regards to correct position of the articulators, breathing, loudness, rhythm and nasalization. Hence the availability of visual information regarding these aspects would greatly help the hearing-impaired improving their speaking abilities.

A system which visualises a speaker's vocal tract by means of mid-sagittal graphical plots of the human head is described in this paper. The vocal tract shapes, and other related speech parameters, are displayed on a PC-monitor using information extracted directly from the acoustic speech signal picked up by a microphone or loaded from an audio file. To estimate the necessary parameters, the speech production process is assumed to be an autoregressive (AR) model. The vocal tract area functions, log spectra and the first three formants are then estimated, using a linear prediction (LP) analysis, and used to display the corresponding vocal tract and other speech parameters.

2. Speech analysis model

Speech is the acoustic wave that is radiated from the vocal system when air is expelled from the lungs and the resulting flow of air is perturbed by a constriction somewhere in the vocal tract. This process can be effectively modelled using the well-known all-pole source-filter approach, which represents the speech signal in terms of an AR model [5]. According to this model, speech is split into a rapidly varying excitation signal, generated by an impulse train input or a random noise generator, and a slowly varying filter representing the vocal tract. Voiced speech is produced by using the impulse train as excitation. In unvoiced segments, the random white noise is used as the excitation. The output speech is produced by passing the excitation through the vocal tract filter. Hence, changes in the vocal tract configuration, reflected by the filter, produces corresponding changes in the spectral envelope of the speech signal. Therefore to estimate the vocal tract shape an inverse filter model has to be used [6].

The speech analysis model used in this work is shown schematically in Figure 1. Here, it is assumed that the speech is limited to periodic non-nasalised voiced sounds so that the filter in Figure 1 is driven by an impulse train. This means that the filter includes all the contributions from the glottal wave, the vocal tract and the radiation impedance at the lips. The inverse filter is assumed linear with only zeros in its transfer function, and the power spectral envelope of the speech is assumed to be approximated by poles only. Accordingly, the transfer function of the inverse filter can be expressed in terms of a z-transform notation as:

$$A(z) = \sum_{i=0}^{p} a_i \, z^{-i} \,, \quad a_0 = 1 \tag{1}$$

where a_i are the coefficients of the inverse filter, $z = \exp(jwT)$ and T is the sampling period. Here, a_0 affects only the gain of the system, hence no generality is lost by setting $a_0 = 1$. To obtain a close representation of the vocal tract, one needs to estimate the coefficients of the optimal inverse filter described by equation (1). Wakita [7] has shown that *A(z)* is also an inverse transfer function of a non-uniform acoustic tube model of the all-pole vocal tract model, and thus the optimal inverse filter process in the above speech analysis model can be equivalently replaced by the filtering process of the acoustic tube, provided that:

a) Continuity conditions for the volume velocity and sound pressure are satisfied at the junctions between adjacent sections,

b) The length of the individual tube sections are kept short compared to the wavelength at the highest frequency of interest,
c) The sampling frequency of the speech signal is fixed to $f_s = c\ p\ /\ 2l$, where l is the assumed vocal tract length and c is the sound velocity, and
d) No losses are accounted for.

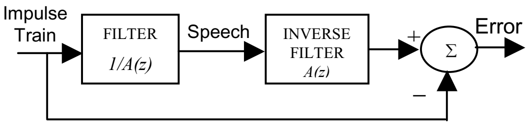


Figure 1: The speech analysis model

3. The vocal tract model

If we abstract from the vocal tract curvature, the acoustic tube can be divided it into cylindrical sections of equal lengths. Depending on the shape of the tube, a sound wave travelling through it will be reflected in a certain way that generates resonance at certain frequencies. The resonances are called formants whose locations largely determine the speech sound that is heard.

It is well known that the linear prediction (LP) analysis of speech signals is based on an AR speech production model [8]. Also, it has been shown by several researchers that the LP process is equivalent to the filtering process of a non-uniform acoustic tube model where the tube is divided into an arbitrary number of sections of equal length [4, 7]. Thus, if the conditions stated at the end of Section 2 are satisfied, and if the speech signal is pre-emphasis to compensate for the spectral characteristics of the glottal excitation source and for the lips radiation impedance, then estimates of the vocal tract area functions can be obtained by computing the reflection coefficients at the junctions between adjacent sections of the equivalent acoustic tube. This can be done by using an LP model of the appropriate order and the following relation:

$$\mu_i = \frac{A_{i+1} - A_i}{A_{i+1} + A_i} \quad \Leftrightarrow \quad A_i = A_{i+1}\frac{1-\mu_i}{1+\mu_i} \qquad (2)$$

where A_i and A_{i+1} are the cross-sectional areas of two adjacent sections of the non-uniform acoustic tube indexed in ascending order from the lips to the glottis, and μ_i is the reflection at the junction between these two sections. For simplicity, the vocal tract is assumed to be 17 cm long. If the speech is sampled at 8 kHz, then condition (c) above requires the use of an acoustic tube of 8 sections.

4. System description and design

A PC-based system for visualisation of the human vocal tract shapes and other associated speech parameters has been designed and developed. The system uses the PC's sound card operating with 8 kHz sampling frequency and 16-bit resolution, to extract the necessary speech parameters directly from the acoustic speech waveform as outlined in Figure 2.

4.1. Estimation of the area functions

Referring to Figure 2, the speech signal is segmented into 30 ms frames using a hamming window of an appropriate length. A pre-emphasis of an approximately 6dB/octave is then applied to the current frame. The reflection coefficients are computed by applying a 7th order LP analysis model. Equation (2) is then used to estimate the corresponding vocal tract area functions. The LP model is also used to obtain the log spectra, whose peaks are then marked to identify and estimate the first three formant frequencies.

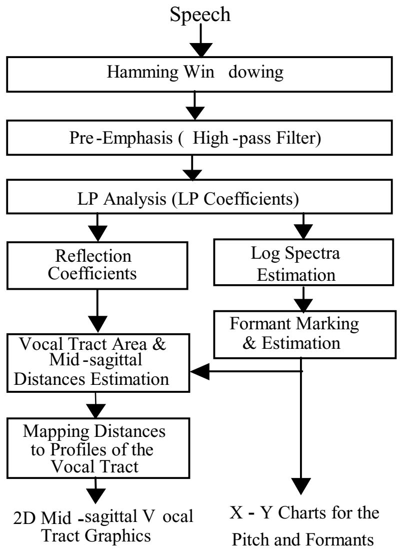

Speech
Hamming Win dowing
Pre-Emphasis (High -pass Filter)
LP Analysis (LP Coefficients)
Reflection Coefficients
Log Spectra Estimation
Formant Marking & Estimation
Vocal Tract Area & Mid -sagittal Distances Estimation
Mapping Distances to Profiles of the Vocal Tract
2D Mid -sagittal V ocal Tract Graphics
X - Y Charts for the Pitch and Formants

Figure 2: Functional block diagram of the system

4.2. Mapping to mid-sagittal distances

As the human vocal tract does not resemble an exact circular tube, there is a need to modify the above computed area functions such that they map correctly into mid-sagittal distances of the vocal tract profiles. Several areas to profile transformation techniques have been developed [3]. Most such techniques rely on derivation of suitable application-specific transformation parameters using complex analysis of x-ray and cine-fluorograms of various speakers. A common technique is the *αβ* model [3], which is described by:

$$A_i = \alpha_i\, d_i^{\beta_i} \quad \Leftrightarrow \quad d_i = (\frac{A_i}{\alpha_i})^{1/\beta_i} \qquad (3)$$

where A_i is the cross-sectional area of a given section, d_i is the mid-sagittal distance and α_i and β_i are section dependent parameters. In our system, we employed a new method based on the above model to compute the mid-sagittal distances along the lines of a semi-polar grid (*See Figure 3)*, as follows:

a) The vocal tract was divided into 18 equal sections.
b) In the vocal organs, the shortest path from the upper to the lower part of each section was selected.
c) Upper jaw was assumed fixed and lower jaw movable.
d) A reference grid for the upper jaw based on x-ray data of the lateral shape of the vocal tract and on results of the PARAFAC analysis [2] was designed, as shown in Figure 3. In this grid, straight perpendicular lines were drawn through the centre of each section, in accordance with (b) above.
e) The 8 area functions estimated by the 7th order LP model were re-sampled and redistributed to fit the 18-section vocal tract configuration used in the system.
f) Based on equation (3), a simple numerical procedure was used to estimate values of the coefficients α and β that minimize the root mean squared error between the

estimated area functions and those derived from measurement data for pseudo-sagittal dimensions of the tongue position for five speakers each saying ten English vowels obtained from [2]. The estimated area functions are then interrupted as functions of α and β, as given in equation (3), to compute the mid-sagittal distances.

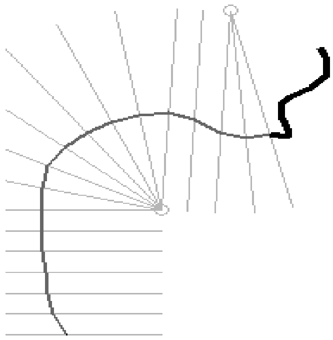


Figure 3: The reference upper jaw grid used in the system

In general, the male vocal tract is slightly longer than 17 cm, while children and females have shorter vocal tracts [9]. Hence, fixing the vocal tract's length to 17 cm may cause an error in the distribution of the area functions. To compensate for any possible error, the first two mid-sagittal distances have been determined from the three estimated formants F_1, F_2 and F_3 as follows [10]:

$$X_1 = C_1 F_2 + C_2 F_2 F_3 + C3 F_1 / F_2 + C_4 \qquad (4)$$

where X_1 is the mid-sagittal distance between the lips in cm, $C_1 = 0.3\times10^{-3}$, $C_2 = -0.343\times10^{-6}$, $C_3 = 4.143$, $C_4 = -2.865$. The distance between upper and lower teeth, X_2, is estimated by:

$$X_2 = \frac{X_1 + X_3}{2} \qquad (5)$$

where X_3 is the mid-sagittal distance extracted from the vocal tract area function that corresponds to section 3. In addition, the estimated formants have been used to adjust the rounding degree of the lips and the height of the jawbone on the designed vocal tract lateral graphics.

5. Results and discussion

The multi-display window and other user's features of the complete system are shown in Figure 4. As can be seen, the system's screen is divided into four windows for displaying the vocal tract graphics, the sound intensity, the pitch and the first three formants of the speech signal. The system can operate in: (a) near real-time mode, whereby the speech signal is picked up by a microphone connected to the PC sound card (as with the case shown in Figure 5), and (b) non real-time mode, whereby the speech signal is either recorded by the system or read from a stored audio file. It also allows the saving of speech/sound signals. For the vowel articulation, the user can compare the shape of his/hers vocal tract to a reference trace (shown with a dashed line in Figure 4) for the correct tongue position derived from the measurements data reported in [6]. The deviation from the reference trace is given for this case in the form of a computed mean squared error (MSE) of all the estimated mid-sagittal distances. Figure 5 shows the vocal tract profiles for 10 American English vowels, as estimated by the system. For evaluation purpose, MSE deviations from the reference tongue position data adopted from [2] are also indicated.

In general, the obtained results seem to correlate well with the reference and x-ray data, and with the PARAFAC analysis. Referring to the MSE values shown in Figure 6, the system seems to perform particularly well in the cases of all the 'front vowels', such as /IY/, /EY/, /IH/, /EH/ and /AE/, with the MSE increasing as the vowel height decreases. With the exception of /AA/ and /UH/, the results show relatively less accurate correlation with the reference data for the cases of the 'back vowels'. As vowel classification into front and back vowels is related to the position of the tongue elevation towards the front or the back of the mouth, we believe that the higher accuracy in the cases of the front vowels is attributed to the formant-based added adjustments of the lips, jawbone and front sections of the vocal tract we used in our approach. On the other hand, the relative length of the vowel's vocalisation seems to affect the accuracy of the estimated area functions and hence the displayed vocal tract shape. In specific, the system seems to give relatively lower accuracy for relatively longer vowels, such as /AO/, and complex vowels which involve changes in the configuration of the mouth during production of the sound, such as /OW/. We believe this is due to the fact that the system, in its current design, bases its estimation on information extracted from the 2-3 middle frames of the analysed speech waveform.

6. Conclusions

A computer-based system for the near real-time and non real-time visualisation of the vocal tract shape during vowel articulation has been presented. Compared to other similar systems, this system uses a new approach for estimating the vocal tract mid-sagittal distances based on both the area functions and the first three formants as extracted from the acoustic speech signal. It also utilises a novel and simple technique for mapping extracted information to corresponding mid-sagttial distances on the displayed graphics. The system also displays the sound intensity, the pitch and the first three formants of the uttered speech. It extracts the required parameters directly from the acoustic speech signal using an AR speech production model and LP analysis. Reported preliminary experimental results have shown that in general the system can reproduce well the shapes of the vocal tract, with real-time sensation, for vowel articulation. Work is well underway to optimise the algorithm used for extraction of the required information and the mapping technique, such that dynamic descriptions of the vocal tract configuration for long and complex vowels, as well as vowel-consonant and consonant-vowel are obtained. Enhancement of the system's real-time capability and features are also being investigated.

7. References

[1] Choi, C. D., "A Review on Development of Visual speech Display Devices for Hearing Impaired Children". *Commun. Disorders*, 5: 38-44, 1982.

[2] Harshman, R., Ladefoged, P. and Goldstein, L. "Factor Analysis of Tongue Shapes", *J. Acoustics Soc. Am.*, 62: 693-706, 1977.

[3] Heinz, J. M., and Stevens, K. N., "On the relations between lateral cineradiographs area functions and acoustic spectra

of the speech", *Proc. 5th Int. Congress of Acoustics*. Paper A44, Liege 1965.

[4] Markovic, M., "On determining heuristically decision threshold in robust AR speech model identification procedure based on quadratic classifier", *Proc. 5th Intl. Symp. Sognal Process. And its Applications (ISSPA'99)*. 131-134, Brisbane, Australia, 1999.

[5] Quatieri, T. E., "*Discrete-time Speech signal Processing, Principles and Practice"*. Prentice Hall, NJ, 2002

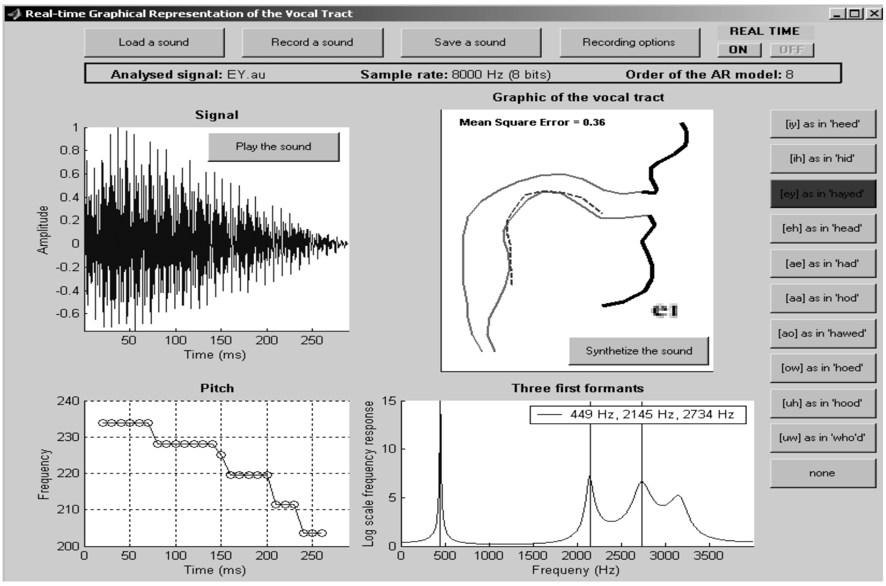
[6] Miller, J. E., and Mathews, M. V., "Investigation of the glottal waveshape by automatic inverse filtering", *J. Acoust. Soc. Am.*, 35: 1876-1884, 1963.

[7]Wakita, H., "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms". *IEEE Trans. on Audio and Electroacoustics*, AU-21: 417-427, 1973.

[8] Markel, J., and Gray, A.,*"Linear Prediction of Speech"*. Springer-Verlag, New York, 1976.

[9] Kirlin, R. L., "A posteriori estimation of vocal tract length". *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP—26: 571-574, 1978.

[10] Ladefoged, P., R. Harshman and Goldstein, L., "Generating vocal tract shapes from formants frequencies". *J. Acoustics Soc. Am.*, 64:1027-1035, 1978.



Real-time Graphical Representation of the Vocal Tract

Analysed signal: EY.au **Sample rate:** 8000 Hz (8 bits) **Order of the AR model:** 8

Mean Square Error = 0.36

449 Hz, 2145 Hz, 2734 Hz

Figure 4: System's multi-display screen and user's features.

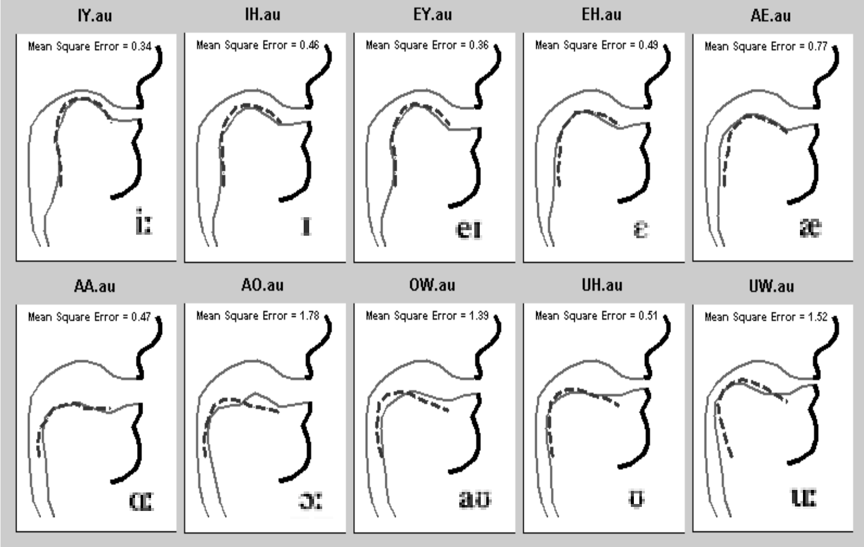


Figure 5: The vocal tract profiles for 10 American English vowels, as estimated by the system