Synthesis of sentence-level speech based on measured vocal tract area functions

Brad H. Story, Ingo R. Titze, and Russel Long

Wilbur James Gould Voice Research Center, Denver Center for the Performing Arts, Denver, Colorado 80204

Abstract: A parametric model of the vocal tract is developed based on an inventory of area functions acquired for one male subject with Magnetic Resonance Imaging (MRI). The model is then used to synthesize a sentence recorded from the same subject.

AREA FUNCTION MODEL OF THE VOCAL TRACT

The question that is addressed in this paper is whether an inventory of vocal tract area functions, which represent only static speech sounds, can be successfully used to synthesize dynamic (i.e. connected) speech recorded from a speaker. This is a form of articulator synthesis in which the effective shape of vocal tract airway is directly manipulated rather than explicitly defining position and movement of individual articulators (e.g. tongue tip and body, velum, lips, jaw, etc.). For this discussion, an area function inventory for one male speaker acquired with Magnetic Resonance Imaging (MRI) (1) will be used.

The area function at any instant of time is represented as the combination of a vowel substrate and an imposed consonantal element. Such an approach was proposed by (2) and further developed by (3), and (4). The vowel substrate $V$ for this study is represented by a principal components analysis (PCA) of the ten vowel area functions in (1,5) as,

$$V(z) = \sum_{j=1}^{N} c_j \phi_j(z) + \Omega(z)$$

where $z$ the distance from the glottis, the $\phi_j(z)$'s are the principal components (or eigenvectors), the $c_j$'s are the coefficients produced by the PCA that will reconstruct each of the original ten vowels, and $\Omega(z)$ is the mean area function. Thus, any vowel is represented as a perturbation around the mean area function. It has been found that four of the principal components (i.e. 4 coefficients) can reconstruct each area function to within 3 percent of its original fidelity. For a time-varying area function, the $c_j$'s become time-dependent parameters,

$$V(x, t) = \sum_{j=1}^{N} c_j(t) \phi_j(x) + \Omega(x)$$

The consonantal element $C$ is represented by a Gaussian function that has a value of 1.0 throughout the extent of the area function except in the region of desired constriction. Justification for a Gaussian function comes from observation of the consonantal area functions also acquired with MRI. The time-varying $C$ takes the following form,

$$C(x, t) = 1 - s_c(t)e^{-\frac{1}{2}}\left(\frac{x - l_c(t)}{p_c(t)}\right)^2$$

where $l_c(t)$ is the location of the constriction in terms of distance from the glottis and $p_c(t)$ is the constriction width. The $s_c(t)$ is called the “strength” of the consonant and can vary from 0.0 for an “all vowel” condition to approximately 0.95 for a partial occlusion as in a fricative, and finally to 1.0 for a full occlusion as in a stop. The final area function $S(x, t)$ is the product of $V(x, t)$ and $C(x, t),$

$$S(x, t) = \left\{ \sum_{j=1}^{N} c_j(t) \phi_j(x) + \Omega(x) \right\} \cdot \left\{ 1 - s_c(t)e^{-\frac{1}{2}}\left(\frac{x - l_c(t)}{p_c(t)}\right)^2 \right\}$$

Nasalization is made possible by an additional parameter $N(t)$ which governs the cross-sectional area of the velar port.
SPEECH-TO-SPEECH SYNTHESIS

In order to synthesize a recorded sentence, a method is needed in which the speech waveform can be mapped to the parameters specified above. For the vowel substrate, a database of 2500 formant triplets (F1, F2, and F3) was computed for area functions generated by 2500 pairs of principal component coefficients 1 and 2 (c1's in Eqn. (1), where j = 1, 2). Only two principal components have been used in order to simplify the database, however, this does compromise the fidelity of the reconstructed area functions; two components will reconstruct to within 12 percent of the original area function fidelity. With this database, the time-varying formant locations can be extracted from the speech waveform by an LPC (linear predictive coding) analysis and immediately mapped to their corresponding coefficient pair (c1(t) and c2(t)), which in turn can be used to generate an area function with Eqn. (2).

At this point, the consonantal parameters are determined from manual segmentation of the speech waveform, phonetic knowledge of the sentence, and an estimate of constriction location lC from MRI-based consonant area functions. An automated identification of consonants and subsequent extraction of lC(t), pC(t), and sC(t) in Eqn. (3) is a future effort.

In addition, the fundamental frequency (F0) of the voice is determined on a cycle-to-cycle basis from a simultaneously recorded electroglottograph and an estimate of amplitude is determined by computing a frame-based RMS pressure from the acoustic speech waveform. These two parameters are used to control a parametric glottal flow model (6).

Figure 1 shows two spectrograms of the sentence “I enjoy the simple life”. The top figure is an analysis of the recorded from the same subject who was vocal tract was imaged in (1). The spectrogram on the bottom shows the synthesized version based on the methods discussed in this paper. The spectrograms show many similarities, with the main difference being that the excursions of the formants are not as extreme in the synthesized version. Informal listening tests indicate that the synthesized sentence is quite intelligible and reproduces many of the qualities of the recording. It’s main downfall is in the quality of the voice source and the fricative consonants. More accurate voice and noise source modeling will be required for close replication of the original speaker.

![Spectrograms of “I enjoy the simple life”; a) recorded, b) synthesized.](image)

ACKNOWLEDGEMENTS

This study was supported by grant No. R01 DC02532 from the NIDCD.

REFERENCES