ARTICULATORY SYNTHESIS OF SUSTAINED SPEECH

Submitted by

Avinash Yentrapati

Mentor: Dr. Espy-Wilson

Research in Telecommunication Engineering (RITE)

MERIT 2005

TABLE OF CONTENTS

ABSTRACT

Articulatory synthesis has been touted to be the synthesis technique which can handle all the problems associated with high quality speech synthesis. The focus of this work was on synthesizing sustained sounds (vowels, semivowels and nasalized vowels) using the technique of articulatory synthesis. The area function corresponding to a sustained sound was converted to its spectrum using VTAR (a computer simulation program developed in SCL). Rosenberg Type B waveform was used as the glottal source. The period and amplitude of this wave were varied randomly to make it sound more natural. This synthesis module was then integrated with VTAR to be able to synthesize sounds directly from the GUI.

1. INTRODUCTION

Articulatory Synthesis is a method of synthesizing speech by controlling the speech articulators (e.g. jaw, tongue, lips, etc). This form of synthesis has been a focus of research since only very recently, and a greater understanding in this area of synthesis promises much insight into the production of natural sounding continuous speech. It can also be a very useful aid for understanding speech production and developing parameters for recognizing speech.

In the past, articulatory synthesis was not an obvious choice due to lack of information of the shape of the vocal tract and moreover, it would have been impractical to pursue such synthesis due to the fact that it would require heavy mathematical calculations. However, technological advances in various fields have collectively given hope for pursuing articulatory synthesis. For example, the information of the vocal tract could be readily, though not cheaply, obtained from MRI data. And since the advent of modern computers, mathematical calculations are not going to pose a problem.

As mentioned earlier, articulatory synthesis has promising applications if research in this area moves at a steady pace. Better understanding of how our major articulators affect our speech could lead to better speech synthesizers or voice recognition systems. It is difficult to produce synthetic speech

of high quality and may not posses all the attributes that we may hope for. An ideal synthesizer would (Hill, Manzara and Schock, 1995):

1. be quite as intelligible as a human
2. sound almost perfectly natural
3. be able to sound like different speakers such as male or female, young or old, in between.
4. be capable of speaking various other languages
5. be able to uniquely reproduce a particular persons' voice
6. be able to switch to a different speaker, alter quality of sound without much effort
7. teach us new things and provide opportunities to learn more as we strive to produce a commercially useable system.

2. A HISTORICAL PERSPECTIVE

Articulatory speech synthesis has been a goal for speech researchers from the earliest days, ever since they started considering machine synthesis of speech. The physical model by Kratzenstein in 1779 was the earliest documented example of articulatory synthesis of speech (Hill, Manzara and Schock, 1995). The quest for the understanding of the nature of five basic vowels in English started when Kratzenstein's work was entered for a prize. By 1791, a more complete and elaborate model was made by Wolfgang von Kempelen. His model used hand control of a leather “vocal tract” (the tube originating from the vocal cords to the mouth, see figure 1) to be able to vary the sounds produced, which included a bellows for lungs, auxiliary holes for nostrils and reeds, and many other mechanisms (Dudley & Tarnoczy, 1950). Alexander Graham Bell later was able to produce his own physical articulatory model, which was able to produce vowels, nasals and few other simple utterances (Flangan, 1972). Graham Bell’s model was based on skull castings and internal parts.

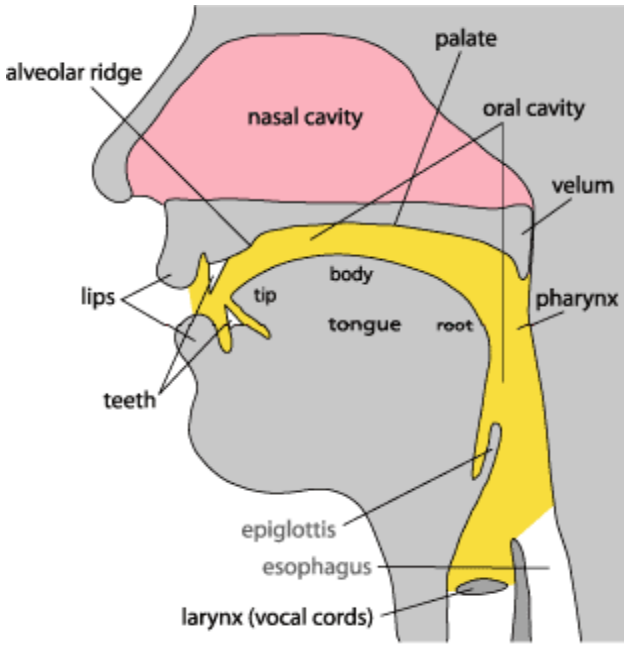

Figure 1. Illustration of vocal and nasal passages.

Over the past 40 years or so, speech synthesis by spectral reconstruction, rather than vocal tract modeling has dominated. Techniques for spectral reconstruction mimic that which is seen in spectral speech analysis, and hence can be controlled based on data, and moreover can be understood easily by anyone that works with them (Hill, Manzara and Schock, 1995). When 'The sound Spectrograph' was invented at the Bell Laboratories, people speculated that all the problems of speech recognition and synthesis had been solved after the machine was able to provide such a dramatic picture of speech energy distribution in time and frequency. Not only was the machine able to reveal the variation with time of the energy in various frequency band, but also clear depictions of the changing resonances and energy sources that were involved. However, in reality even the most

experienced and talented person would need at least two years of training to “read” spectrographs for purposes of recognition.

3. MODERN ARTICULATORY SYNTHESIS

The fields of Electrical Engineering and Computer Science have played a major role in shaping the modern approaches to articulatory speech synthesis. Instead of physically modeling the acoustic tube (the vocal tract), the sound propagation is modeled algorithmically, applying the same techniques that are used to model high-speed pulse transmission lines (the transmission line is analogous to the acoustic tube in that it models the sound propagation in the vocal tract which may be thought of as a spatial filter or as a waveguide). The whole length of the vocal tract (acoustic tube) can be divided into smaller sections, usually equal sections, where each chunk is thought of possessing appropriate “impedance” that corresponds to physical reality (kelly & Lochbaum, 1962). In earlier times, these elements were represented as inductors and capacitors (Dunn, 1950), but then the resulting systems tended to be very unstable. Only since recently, more computationally challenging methods have been tried, thanks to the emulation of the digital computer - each section of the vocal tract can be thought of as a transmission line which comprises of forward moving wave, a reflected wave and reflection paths which connect the forward and reverse paths. Implementing this form of method would require a lot of calculation, partly because of the nature of the method and also due to the number of sections that the vocal

tract is divided into.

3.1 The Acoustic Tube (vocal tract)

The resonant behavior seen in transmission line is much similar to the resonance in the vocal tract. To avoid reflections in transmission lines or waveguides, they are designed to have uniform impedances and variations in physical characteristics are not to exist. But if these conditions are not met, then reflection of waves occurs between boundaries, which give rise to the resonant behavior.

The transmission line theory applies very well to the study of the vocal tract. A human vocal tract has varying cross-sectional areas and other varying properties throughout the tract. Hence, it makes sense to divide the vocal tract into smaller sections, in which each section does not vary in cross-sectional area or does not vary in properties.

The input energy to the vocal tract comes from the vibrating glottis, driven by the air pressure released from the lungs, causes the sound waves to propagate through the tube. The vibrating glottis (vocal cords) gives rise to a periodic characteristic. This periodic waveform is known as the **glottal waveform**.

If we imagine that each section of the vocal tract is different from the other, then part of the propagating wave would be reflected at the boundaries due to area changes and other

properties. This leads to the resonances characteristic of the vocal tract. It is also important to note that fraction of the energy that reaches the mouth is again reflected back into the vocal tract. To be able to model this behavior is not a simple task. All the things discussed so far about the oral branch also equally applies to the nasal branch, which is connected part way along the oral branch, which allows for energy to be distributed between the oral and nasal branches.

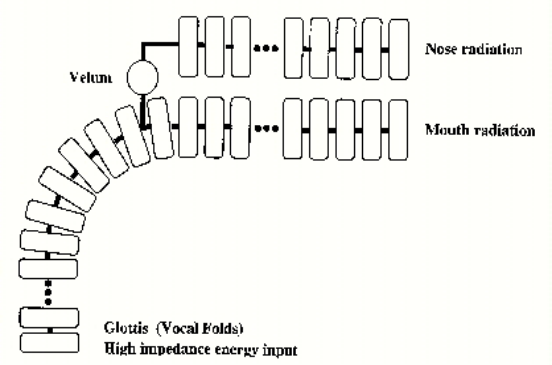

Figure 2: Multi-section physiological model of the human vocal tract

4. METHODOLOGY

ARTICULATORY SYNTHESIS PROCEDURE (see Figure 3)

- Use Magnetic Resonance Imaging (MRI) to image the vocal tract during the articulation of sustained speech sounds.
- Obtain the area function of the vocal tract from the MRI images.
- Use this area function as input to a computer program called VTAR (Vocal Tract Acoustic Response), which was developed at the University of Maryland Speech Communication Lab. VTAR calculates the transfer function of the vocal tract.
- Take the Inverse Fourier Transform (IFFT) of the transfer function to obtain the impulse response of the vocal tract
- Convolve the frequency response with the Glottal waveform to get synthesized speech.

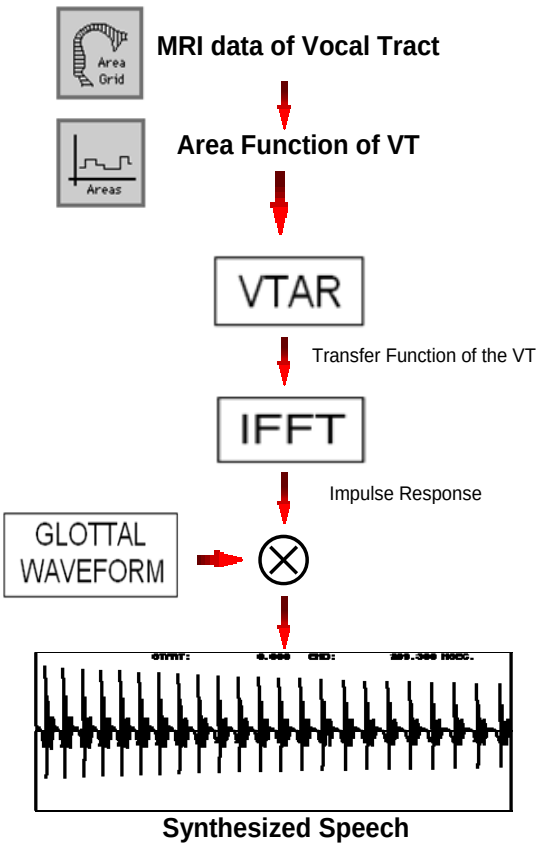

Figure 2. Steps involved in Articulatory synthesis

5. GLOTTAL WAVEFORM MODEL

The glottal waveform model (see Figure 4) that we used for the study was Rosenberg type B model (Rosenberg, 1971).

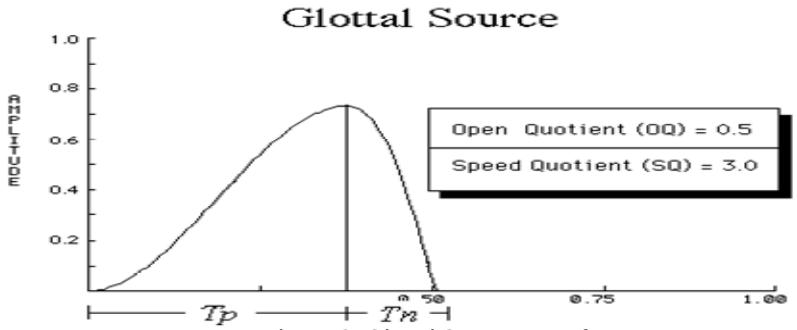

Figure 3. Glottal Source waveform

POSITIVE SLOPE	NEGATIVE SLOPE
$0 \le t \le Tp$	$Tp \le t \le Tp + Tn$
$3(t/Tp)\,\wedge\,2 - 2(t/Tp)\,\wedge\,3$	$1 - (\frac{t - Tp}{Tn})\,\wedge\,2$

OPEN QUOTIENT = $(\frac{Tp + Tn}{T})$

SPEED QUOTIENT = Tp / Tn

Figure 4. Glottal Waveform model

Glottal Parameters

- Time Period – the fundamental period of the waveform.
- Open Quotient - fraction of the time period during which time the vocal cords are open.
- Speed Quotient - asymmetry of the glottal pulse.
- Jitter – random variation in the time period of glottal pulse.
- Shimmer – random variation in the amplitude of the glottal pulse.

6. VTAR (Vocal Tract Acoustic Response)

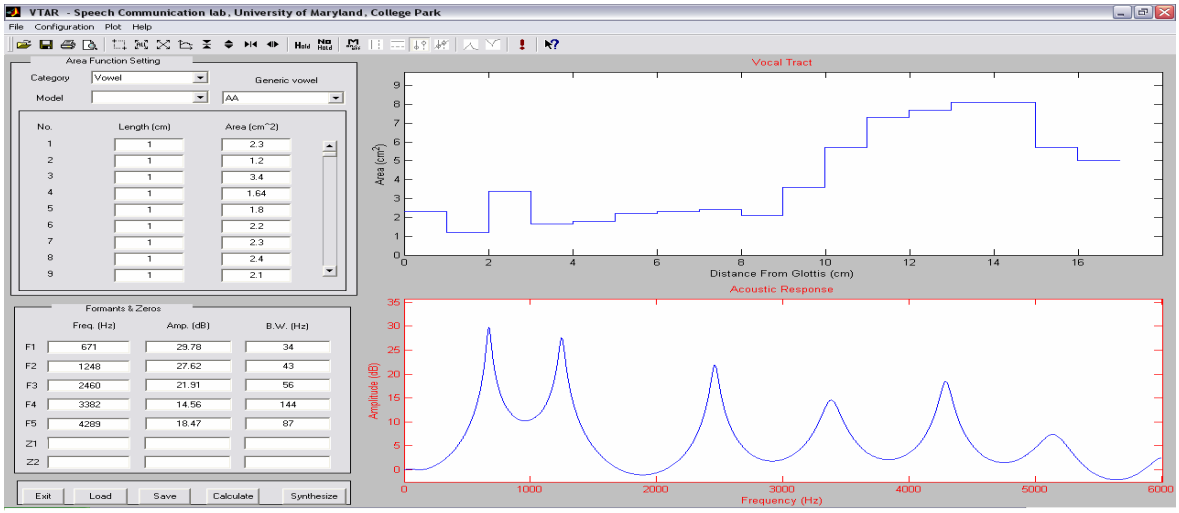

Figure 5. Snapshot of VTAR

Figure 5 shows a snapshot of the computer program called VTAR (Vocal Tract Acoustic Response). The purpose of VTAR in this study was to obtain transfer functions of the vocal tract for different vowels. The inputs to VTAR are simply the area functions of the vocal tract which are derived from the MRI images. Using these area functions, VTAR is capable of calculating the transfer functions. The loss characteristics of the vocal tract are also taken into account in these calculations.

Add-on Module to VTAR

An add-on module to VTAR was developed that allows for the synthesis of speech. After the transfer function is calculated in VTAR, clicking on the "synthesize" (see Figure 6) would play the sound. The glottal waveform used for this synthesis is the

same as the one mentioned earlier. This module also allows for one to change the parameters of the glottal waveform to study the effect of the parameters on the sound synthesized.

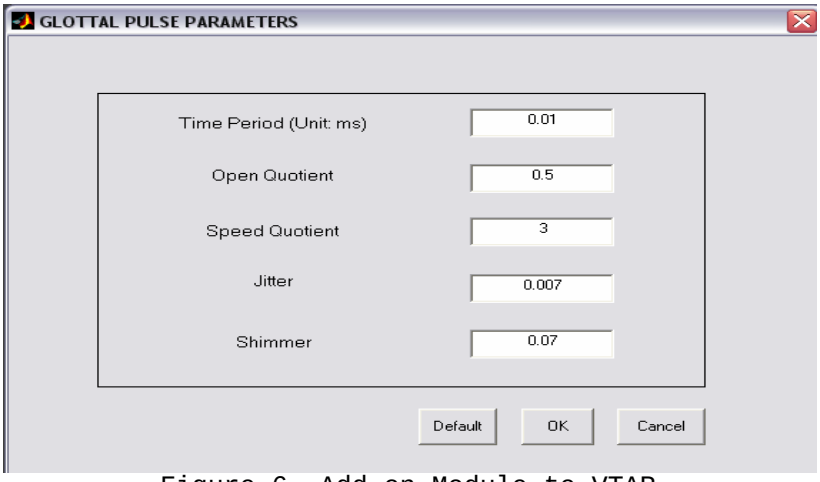

Figure 6. Add-on Module to VTAR

7. RESULTS AND CONCLUSIONS

One problem consistently arose when synthesizing the sounds- there was a lot of buzz involved in the synthesis. Though we were partly successful in eliminating the buzz, it still does slightly exist. This “buzziness” is experienced due to the monotonous glottal waveform (Sambur, 1978).To eliminate such buzz, we had to randomly vary the jitter, shimmer, open quotient and speed quotient in order to make it sound more natural (Sambur, 1978). This technique did make it sound a lot better, but there is still much more work needed to be done in this area.

This synthesis can be used to conduct perceptual experiments to study the effects of changes in area functions and the various glottal parameters on synthesized speech, and gain better understanding of the vocal tract and the major articulators. Understanding the precise relationship between the parameters and the sound produced can be very important for improving speech/speaker recognition applications.

8. REFERENCES

Dudley, H & Tarnoczy, T.H. (1950). “The speaking machine of Wolfgang von Kempelen. J. Acoust. Soc. Am. 22 (1), 151, 166

Dunn, H.K. (1950). The calculation of vowel resonances and an electric vocal tract. J. Acoust. Soc. Am. 22 740-753

Hill, D., Manzara, L., and Schock, C. (1995). “Index to “Real-time” Articulatory speech-synthesis by rules”, AVIOS, 27-44

Kelly, J.L. & Lochbaum, C.C. (1962). Speech synthesis. Proc. 4th International Congress on Acoustics, Paper G42: 1-4

Rosenberg, A.E. (1971). “Effect on Glottal Pulse Shape on the Quality of Natural Vowels,” J. Acoust. Soc. Am. 49, 583-590

Sambur, M.R., Rosenberg, A.E., Rabiner, L.R. & McGonegal, C.A. (1978). “On Reducing the buzz in LPC synthesis.” J. Acoust. Soc. Am 63, 918-925