The Tube Resonance Model Speech Synthesizer

Leonard Manzara

Department of Computer Science
University of Calgary
2500 University Drive NW
Calgary, Alberta, Canada
T2N 1N4

manzara@cpsc.ucalgary.ca

Abstract

The Tube Resonance Model (TRM) synthesizer is an articulatory speech synthesizer implemented in software. It directly emulates the resonant behavior of the oropharyngeal and nasal tracts using digital waveguides. The oropharyngeal cavity is subdivided into 8 regions of unequal length, where particular regions correspond to the human articulators of tongue, teeth, and mouth. The radius (cross-sectional area) of each region can be varied independently over time. The differences in radii between regions gives rise to differences in acoustic impedance, which are modeled using two-way scattering junctions. The nasal cavity is composed of 5 equal-length sections, and is connected to the vocal tract via another section (the velum) using a three-way scattering junction. The total length of the tube can be varied over a continuous range, allowing one to synthesize male, female, and juvenile voices.

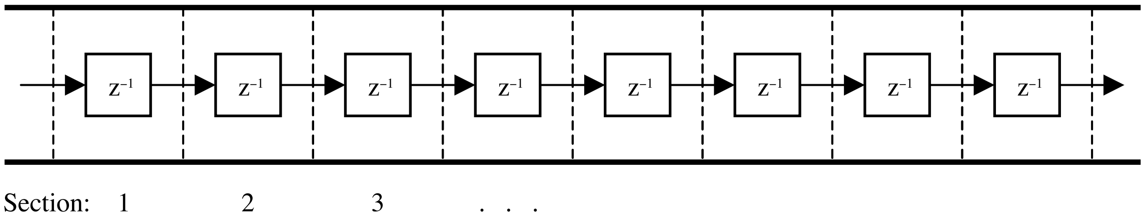
This paper was originally presented as a poster paper at the 149th Meeting of the Acoustical Society of America (ASA) held at the Vancouver Hyatt Regency Hotel, Vancouver, British Columbia, Canada, May 16–20, 2005.

Simulating the Acoustics of a Uniform Tube

Given a rigid tube with a uniform cross-sectional area, and given wavelengths that are large compared to the diameter of the tube, fluid motion in the tube is primarily parallel to its axis. In other words, sound pressure waves travel down the air column of a pipe as one-dimensional longitudinal plane waves, at least to a first degree of approximation (Morse and Ingard, 1968, p. 467).

These *traveling waves* can be simulated in the digital domain using a delay line (see Figure 1 below). This system samples both in time and space. Each delay unit stores the instantaneous pressure for the corresponding section of the tube. At each sample increment, the pressure values are shifted one unit to the right.

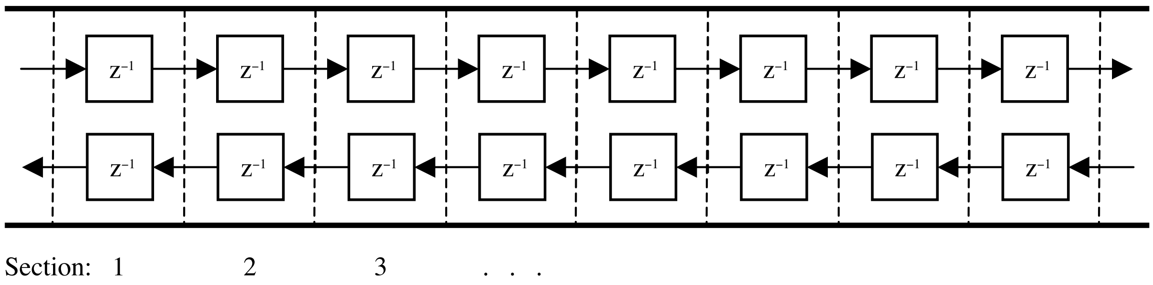
Figure 1: *Using a delay line to simulate traveling waves in a tube.*



Sound waves can travel through the tube in both directions simultaneously. The instantaneous pressure at any point in the tube is the sum of the right- and left-going traveling waves. This phenomenon of *superposition* results in constructive and destructive interference throughout the length of the tube.

Superposition can be simulated digitally using a bi-directional delay line or *waveguide* (see Figure 2 below). There are two “rails” in this structure, where the top rail represents the right-going traveling wave, and the bottom rail represents the left-going traveling wave. Note that the instantaneous sound pressure at each section of the tube is the sum of the pressures stored in the top and bottom delay units for the section.

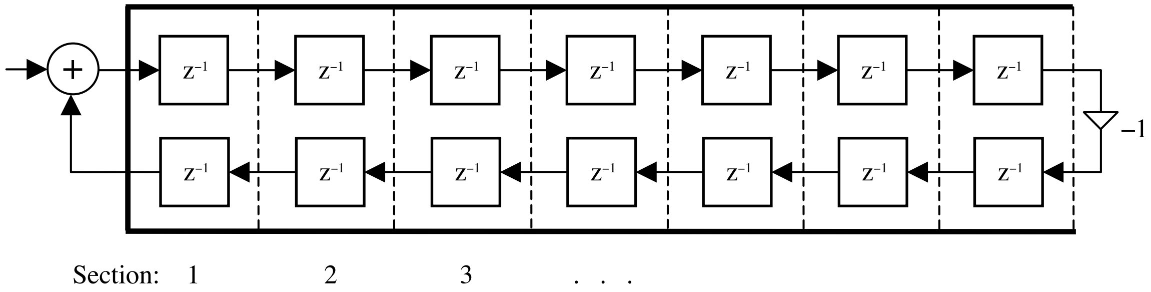
Figure 2: *Using a bi-directional delay line (waveguide) to simulate right- and left-going traveling waves.*



Real-world tubes must necessarily be finite in length, and are terminated with either open or closed ends. When a traveling wave reaches the open end of a tube, it is reflected back into the tube 180 degrees out of phase (i.e. it is inverted). This is modeled digitally by multiplying the value in the rightmost delay unit in the top rail by –1, and putting this new value into the rightmost delay unit in the bottom rail (see Figure 3 below). Of course, some sound energy will be radiated into the free air in a frequency dependent way; exactly how this is modeled is discussed later.

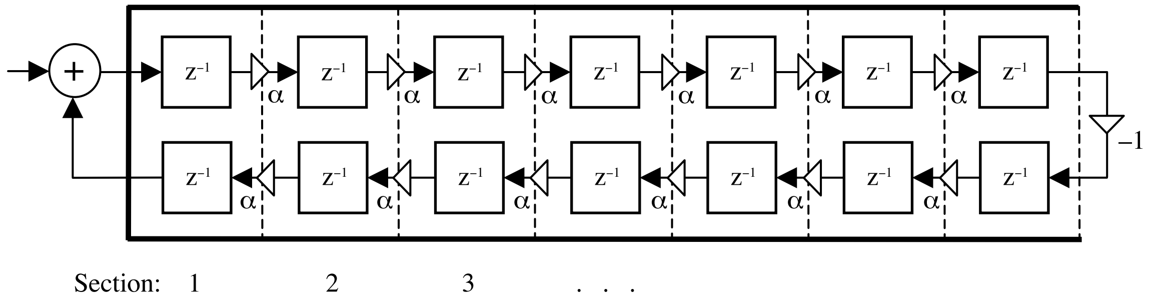
A traveling wave reaching a closed end of a tube is reflected back into the tube in phase (i.e. it is not inverted). In the digital domain, this is modeled by taking the value in the leftmost delay unit of the bottom rail, adding it to the incoming wave, and putting the sum into the leftmost unit of the top rail (see Figure 3).

Figure 3: *Modeling reflections at the closed and open ends of a tube.*



The above model assumes that no energy is lost as sound propagates through the tube. In actuality, a small amount of sound energy is converted to heat due to the viscosity and thermal conduction of the air in the tube (Morse and Ingard, 1968, p. 274). By ignoring any frequency-dependent effects of the phenomenon, this is modeled in a simple way by multiplying the pressure values in each delay unit by a “loss factor” as they are shifted to the next delay unit (see Figure 4 below). Normally, this factor is a value just under 1.0. For example, if we assume a 3% loss in each section of the tube, the factor will be 0.97.

Figure 4: *Incorporating energy losses into the model. α represents the loss factor for each section of the tube.*



Reflections and superposition cause standing waves (or *resonances*) in the tube. For a tube 17 cm in length, closed at one end and open at the other, there are resonant peaks at 500 Hz, 1500 Hz, 2500 Hz, etc. These peaks closely approximate the formant frequencies of the neutral vowel ("schwa") spoken by a male of average size.

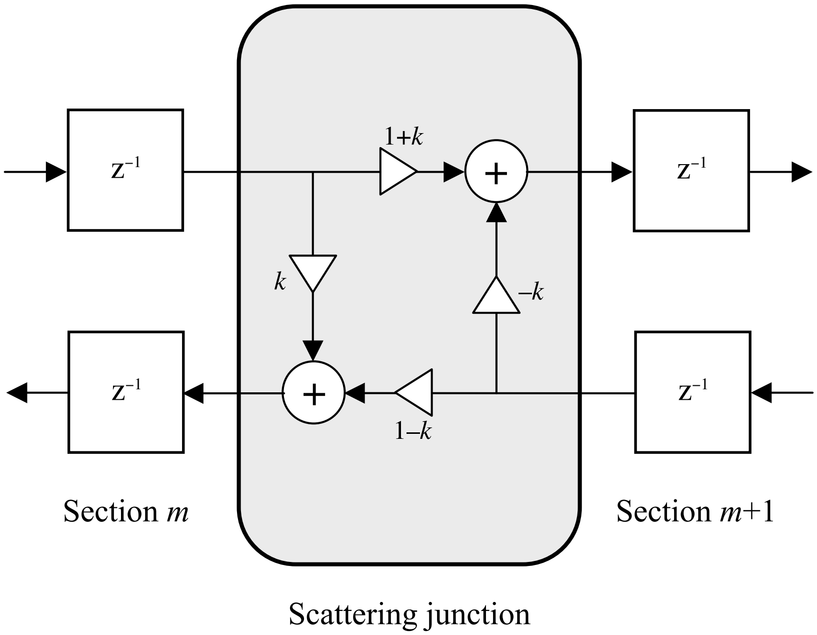
One can emulate different tube lengths in the digital domain by changing the sample rate at which the synthesizer operates. The length of a tube section is given by the relation $L = c / F_s$, where L is the length of a single section of the tube, c is the speed of sound, and F_s is the sampling frequency. As the sample rate increases, the tube length decreases. Sound samples calculated using this "internal sample rate" must be converted to one of the fixed sample rates supported by a given digital-to-analog converter (typically 44.1 kHz).

Modeling Non-Uniform Tubes

A non-uniform tube such as the vocal tract varies in its cross-sectional area from one end of the tube to the other. To simplify the simulation of plane waves traveling through the tube, the tube is characterized as a series of equal-length cylindrical sections, the cross-sectional area of each being the average cross-sectional area of the corresponding part of the smoothly varying tube. The sectioned tube is said to "sample" the smooth tube (Cook, 1991).

Each section of the non-uniform tube has its own characteristic acoustic impedance. If adjacent sections have different impedances (which occurs whenever they have different cross-sectional areas), then part of the pressure wave is reflected and part of it is transmitted at the junction of the two sections. This is modeled using a two-way *scattering junction* (see Figure 5).

Figure 5: *A two-way scattering junction.*



The scattering coefficient k_m is calculated with the formula:

$$k_m = \frac{Z_{m+1} - Z_m}{Z_{m+1} + Z_m}$$

where Z_m is the impedance of section m. Note that if $Z_{m+1} = Z_m$, then $k_m = 0$, and no reflection occurs at the scattering junction (the two sections form a uniform tube).

Since $Z_m = \rho v / S_m$, where S_m is the cross-sectional area of tube section m, and since the density of air ρ and speed of sound v will be the same for both sections, the formula can be recast as:

$$k_m = \frac{S_m - S_{m+1}}{S_m + Z_{m+1}}$$

And since $S = \pi r^2$, the formula can be expressed in terms of a radius r_m for each section m:

$$k_m = \frac{r_m^2 - r_{m+1}^2}{r_m^2 + r_{m+1}^2}$$

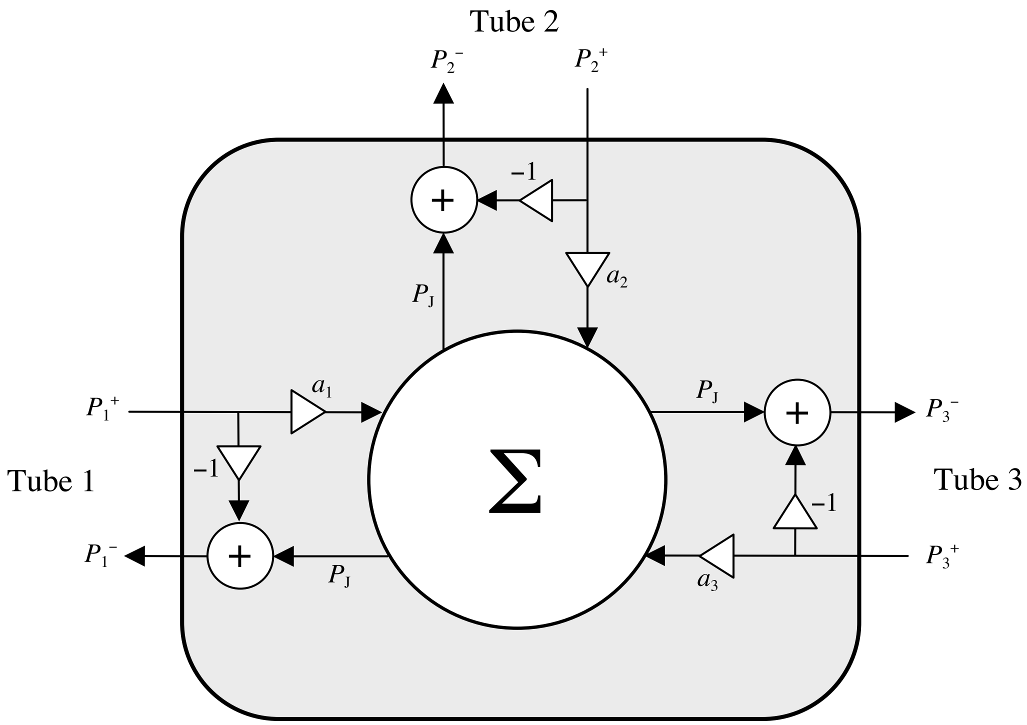
Note that if the radii (or, equivalently, the cross-sectional areas) of the two sections are equal, then $k_m = 0$, and no reflection occurs at the junction. If section m+1 is completely closed (i.e. $r_{m+1} = 0$), then the wave is entirely reflected back into section m since k_m is 1.

N-way Junctions

The nasal tract is connected to the oropharyngeal tract through the velum. The non-uniform tube forming the nasal tract is modeled with a separate waveguide, and its velar connection to the oropharyngeal waveguide is made with a 3-way scattering junction.

An *N-way junction* is a generalization of the 2-way scattering junction discussed above (see Figure 6). The relative impedances of the connecting tubes determine the reflection characteristics of the junction, and these impedances are a function of the relative cross-sectional areas (or radii) of the sections of the tubes adjacent to the junction.

Figure 6: *An example of an N-way scattering junction connecting 3 tubes together*.



The scattering coefficient a_i for tube i is calculated with the formula:

$$a_i = 2\frac{\Gamma_i}{\sum_{j=1}^{N}\Gamma_j}$$

where Γ_i = 1 / Z_i and is defined as the *admittance* of tube i. This formula can be re-expressed in terms of cross-sectional areas S_i as:

$$a_i = 2\frac{S_i}{\sum_{j=1}^{N}S_j}$$

or in terms of radii r_i as:

$$a_i = 2\frac{r_i^2}{\sum_{j=1}^{N} r_j^2}$$

The total pressure P_J of the junction is calculated with:

$$P_J = \sum_{i=1}^{N} a_i P_i^+$$

where P_i^+ is the incoming pressure for tube i. Each outgoing pressure wave P_i^- is calculated by subtracting the incoming pressure from the junction pressure:

$$P_i^- = P_J - P_i^+$$

Radiation and Reflection at the Mouth and Nose

The interface between the last section of a tube and free air can be modeled in a manner similar to that for an ordinary two-way scattering junction, except that sound energy from the outside is not introduced into the tube. That is to say, sound energy exits the tube at the mouth and nose, but none enters. The two-way scattering junction is modified by removing the $-k$ and $1-k$ multiplicative terms (see Figure 7 below).

The scattering coefficient k is calculated as before, using the radius of the last tube section for r_m, and the "radius" of free air for r_{m+1}. Since free air is unbounded, one would be tempted to use infinity for r_{m+1}. However, doing so results in $k = -1$, which implies that sound energy never exits the tube! Obviously this is not so. The solution is to set the radius to an arbitrary value that represents the "effective radius" of the air, and making sure that the radius of the final tube section never exceeds this number. The $1+k$ multiplier can be removed from the scattering junction since the output of the synthesizer is taken from this point, and the volume of the samples will be scaled arbitrarily to give a comfortable playback level.

The frequency response of the reflected and radiated sound at the tube termination must also be modeled. This is done by using a fixed one-pole low-pass filter for the reflected pressure waves, and a fixed one-pole one-zero high-pass filter for the radiated sound (see Figure 8 below). Since the junction between the last tube section and free air is lossless, the frequency response of the radiation filter should be the exact inverse of the frequency response of the reflection filter. That is to say, any energy at a particular frequency that is not radiated should be reflected.

Figure 7: *The scattering junction at the tube/air boundary.*

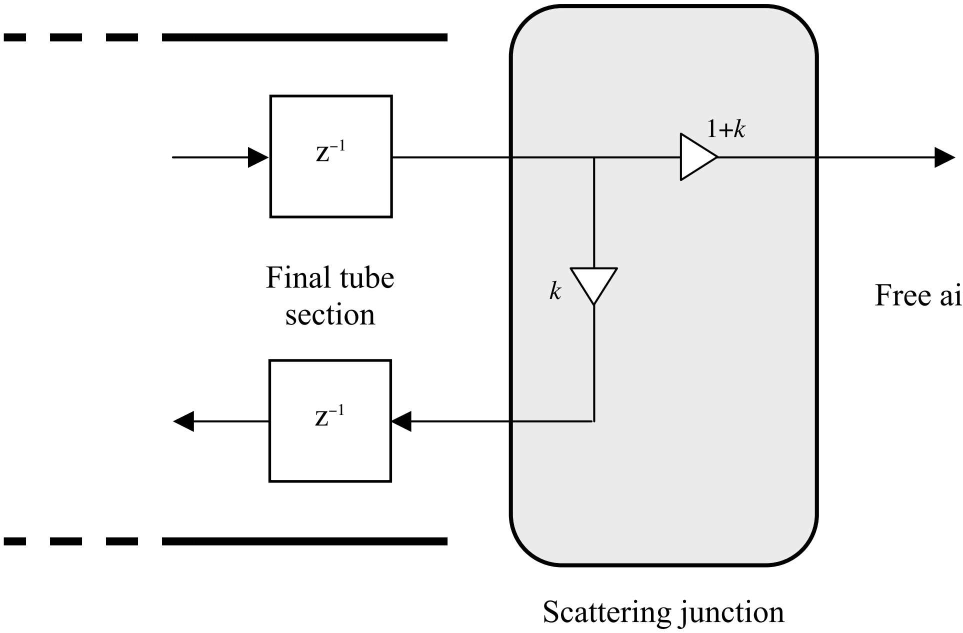
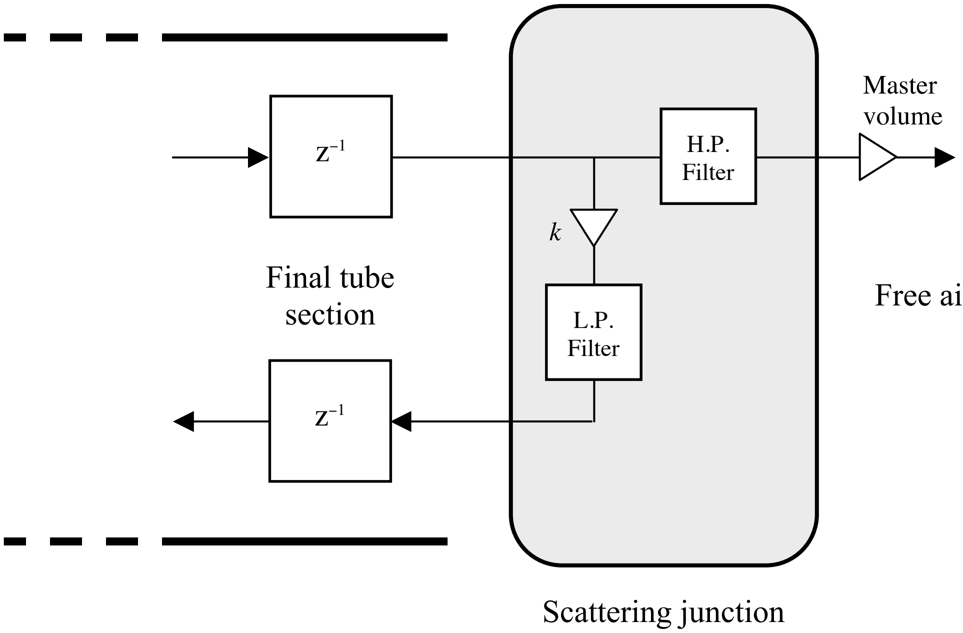


Figure 8: *The scattering junction at the tube/air boundary incorporating filters to model frequency response.*



The Distinctive Region Model

The Distinctive Region Model (Carre and Mrayati, 1992) subdivides the vocal tract into 8 regions of unequal length (see Figure 9). Such a partitioning provides optimal control over the first 3 formants (see Figure 10) and seems to be closely correlated to the position of the human articulators of tongue, teeth, and mouth. The Tube Resonance Model synthesizer closely approximates this sectioning of the tube by using one section each for regions 1-3 and 6-8 and two sections each for regions 4 and 5.

Figure 9: *The acoustic tube is divided into 8 unequal-length regions by using the zero-crossings of the sensitivity functions of the first three resonances.*

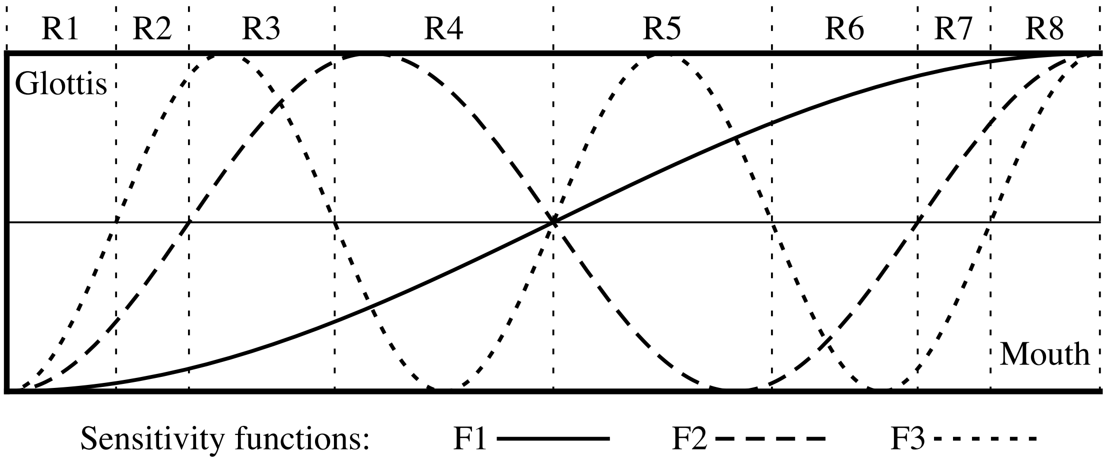


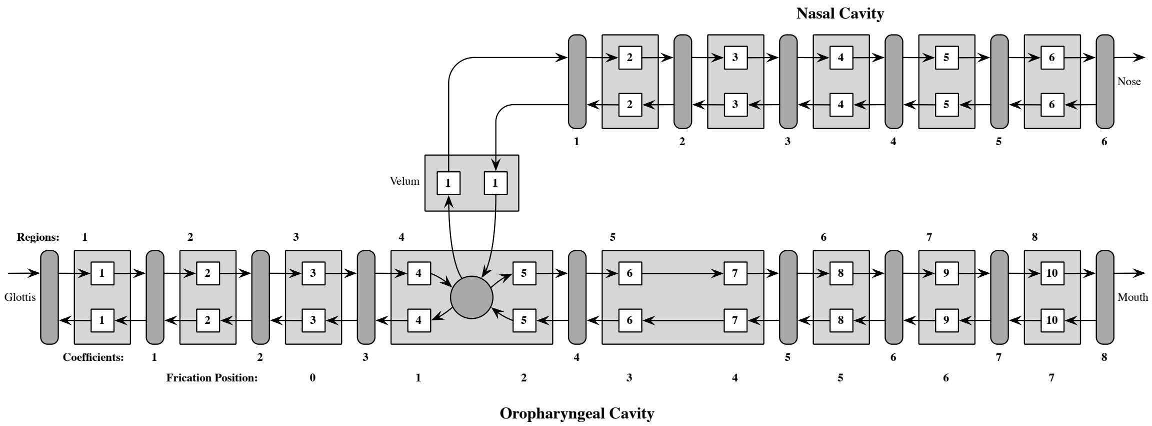
Figure 10: *Constricting the tube in each region raises or lowers the formant frequencies according to the pattern shown.*

	R1	R2	R3	R4	R5	R6	R7	R8
F1	↗	↗	↗	↗	↘	↘	↘	↘
F2	↗	↗	↘	↘	↗	↗	↘	↘
F3	↗	↘	↘	↗	↘	↗	↗	↘

The Topology of the Tube Resonance Model Synthesizer

The topology of the Tube Resonance Model synthesizer is shown in Figure 11. There are 8 regions (10 sections total) for the oropharyngeal tract, 1 section for the velum, and 5 sections for the nasal tract. The mouth corresponds to region 8, the teeth to region 7, and the tongue to regions 2-6. The radii of these regions and the velum are varied over time to produce speech-like sounds.

Figure 11: *The topology of the TRM synthesizer.*



The glottal source is synthesized using a wavetable oscillator, where the shape of the glottal pulse varies according to vocal effort (Rosenberg, 1971). “Breathiness”, in the form of pulsed noise, can be added to the glottal source. Aspiration is controlled separately, and is synthesized using low-pass filtered noise. Frication is similar, except that it is band-pass filtered, and can be injected into the tube anywhere from region 3 to region 8.

Conclusions and Future Work

The TRM has been implemented both in software and on a real-time DSP system (Hill et al, 1995). In its original incarnation, it was part of a larger commercial text-to-speech system, but can be used separately. The quality of the synthesis is subjectively very good, and is a marked improvement over the spectral reconstruction techniques used in typical formant synthesizers.

The model will be improved by replacing the glottal source wavetable oscillator with a physical model of the vibration of the vocal folds. Also envisioned is a new model for frication, where the noise source is synthesized directly by modeling the turbulence created wherever there are occlusions in the tube.

References

Carrè, R., and Mrayati, M. (**1992**). "Distinctive regions in Acoustic Tubes. Speech Production Modelling," J. Acoustique **5**, 141-159.

Cook, P. R. (**1991**). Identification of Control Parameters in an Articulatory Vocal Tract Model, with Applications to the Synthesis of Singing (unpublished dissertion, Stanford University).

Hill, D. R., Manzara, L., and Taube-Schock, C. (**1995**). "Real-time articulatory speech-synthesis-by-rules," AVIOS '95 Conference Proceedings.

Morse, P. M., and Ingard, K. U. (**1968**). Theoretical Acoustics (Princeton University Press, Princeton, New Jersey).

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

Source Code

The TRM synthesizer is part of the larger *gnuspeech* project. The link to the main page is:

http://savannah.gnu.org/projects/gnuspeech

The C Language source code for the synthesizer can be found at:

http://svn.savannah.gnu.org/viewvc/nextstep/trunk/src/softwareTRM/?root=gnuspeech