Synthesis of Nasal Consonants: A Theoretically Based Approach

by

Andrew Ian Russell

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1999

Author ....
Department of Electrical Engineering and Computer Science
January 13, 1999

Certified by. ...
Kenneth Stevens
Clarence LeBel Professor
Thesis Supervisor

Accepted by ....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Synthesis of Nasal Consonants: A Theoretically Based Approach

by

Andrew Ian Russell

Submitted to the Department of Electrical Engineering and Computer Science
on January 13, 1999, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Computer Science and Engineering

Abstract

The theory describing the production of nasal consonants is reviewed and summarized. A method of simulating the acoustics of the vocal tract for nasal sounds is described and implemented. An attempt is made to formulate an approach to synthesis of nasal consonants from observations made from the simulations, in combination with the established theory and empirical observations. Some synthesis was done using the new approach, as well as with conventional techniques, and the results were compared. By means of a simple listening experiment, it was shown that synthesis done using the theoretically based approach sounded more natural.

Thesis Supervisor: Kenneth Stevens
Title: Clarence LeBel Professor

Acknowledgments

I would like to express my deepest appreciation to my thesis supervisor Ken Stevens for being so patient and understanding. Also, it was a pleasure interacting with him as his deep understanding of the speech process is truly awe inspiring. I would also like to thank my wife Wendy-Kaye. Without her support and encouragement, this thesis would not have been written. I also thank the many subjects who participated in my listening experiment. Their contribution has been invaluable

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In English, the nasal consonants are /m/, /n/, and /ŋ/. The consonant /ŋ/ is at the end of the word **sing**. Much is understood about these sounds, but their synthesis is still done with most of this understanding ignored. These consonants are produced in much the same way as the stop consonants /b/, /d/ and /g/, except that the velum is lowered to provide an opening into the nasal cavity. A consequence of this velopharyngeal opening is that, even though there is a complete closure in the oral cavity, no pressure is built up because there exists an alternative path through the nostrils.

This additional acoustic path is what makes nasal consonants and vowels different from other speech sounds. This side branch of the airway is also what makes synthesizing nasal sounds a difficult problem. As yet, no unified theoretically based approach for solving this problem has been described. The purpose of this thesis is to try to identify what acoustic theory says about the processes involved in the production of a nasal consonant, and then to use that information to determine good rules for synthesis. The synthesis will be done using a formant synthesizer developed by Dennis Klatt, known as the Klatt synthesizer, which is described in [5].

1.2 Theory

A nasal consonant in a vowel environment (like the /n/ in the word **any**) is produced in the following way. Some time before the consonant closure is made, during the production of the vowel, the velum is lowered so that the vocal tract now consists of a tube which branches into two tubes. The closure in the mouth is then made with the lips for /m/, the tongue blade for /n/ and the tongue body for /ŋ/. The sound that is produced during the closure is called the *murmur*. The closure is then released and some time after that, the velum is raised to close off the nasal cavity.

1.2.1 Engineering Model of Speech Production

The simple engineering model used for speech treats the glottis as a volume velocity source which produces a certain glottal waveform, and treats the effects of the vocal and nasal tracts as a simple, slowly varying linear filter. The glottal waveform is then passed through the filter, giving the final speech waveform. The frequency response of this filter is $H(j\omega) = U_M(j\omega)/U_G(j\omega)$, where $U_M(j\omega)$ and $U_G(j\omega)$ are the Fourier transforms of the volume velocities at the lips and at the glottis respectively. This model serves as the basis of the Klatt synthesizer which simply filters the glottal waveform using a transfer function with certain number of poles and zeros, whose frequencies and bandwidths vary with time and are controlled by the user. The poles produce peaks in the transfer function which are called *formants*, and the zeros cause dips, or *antiformants*.

Let us first consider the lossless model, where the tube walls (and termination at the glottis) are assumed to be perfectly hard; the sound pressure at the lips and nostrils is assumed to be zero; and the effects of friction and viscosity in the air are ignored. For a single tube with no side branch (see figure 1-1), the transfer function from the glottis to the lips only has poles and no zeros. This model works well for vowels. The frequencies of these poles are spaced at about one every $\frac{c}{2l}$ on average, where c is the speed of sound and l is the length of the tube.

For the case where the tube splits into two tubes, as in figure 1-2, the transfer

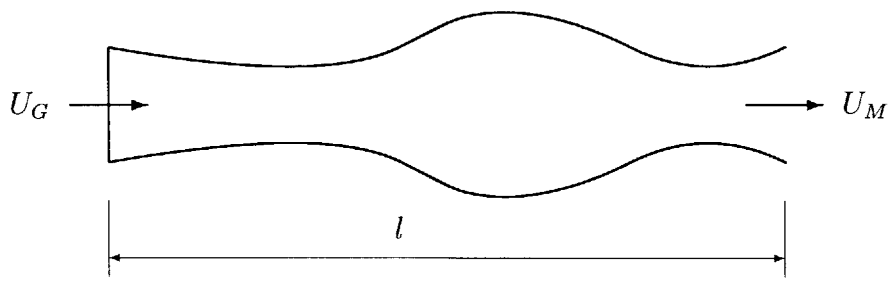


Figure 1-1: The vocal tract modeled as a single tube, with varying cross-sectional area. U_G is the volume velocity at the glottis; U_M is the volume velocity at the lips; and l is the total length of the tube.

function becomes more complicated, $H(j\omega) = (U_M(j\omega) + U_N(j\omega))/U_G(j\omega)$. There are now more poles whose frequencies are spaced on average a distance of $\frac{c}{2(l_G+l_M+l_N)}$ apart, where l_X represents the length the respective tube as shown in figure 1-2. There are also now zeros in the transfer function spaced on average at about one every $\frac{c}{2l_N}$.

1.2.2 Looking at the Susceptance Curves

Attempts have been made to estimate the frequencies of the poles and zeros of the transfer function by looking at the susceptance curves associated with the vocal tract and nasal tract (see [9] chapter 6 and [4]). If we take the susceptances looking in to the oral cavity , the nasal cavity and the pharynx to be B_M, B_N and B_G respectively (see figure 1-2), then it can be shown that the poles of the over-all transfer function occur when $B_M + B_N + B_G = 0$. However, it is somewhat more complicated to find the location of the zeros. We must first find the zeros of the transfer function from the glottis to the nose, which occur when $B_M = \infty$. We can similarly find the zeros of the transfer function from the glottis to the mouth, which occur when $B_N = \infty$. We then take the sum of the two transfer functions being careful to apply the correct scaling to each. The scaling is roughly proportional to the acoustic mass of the corresponding cavity.

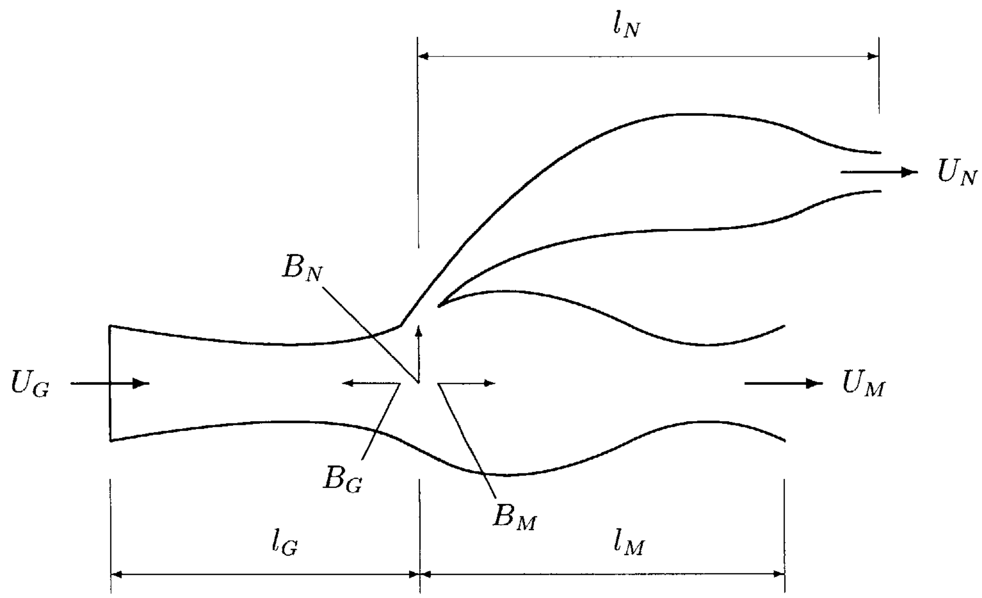


Figure 1-2: The vocal tract modeled as a tube with a side branch. The side branch is analogous to the nasal cavity. U_N is the volume velocity at the nostrils. B_X is the acoustic susceptance looking towards the glottis, mouth or nostrils from the branch point.

1.3 Background

Considerable research has gone into describing the acoustic signal associated with a nasal consonant. Understanding what makes an utterance sound like a nasal consonant is essential to knowing how to synthesize a natural sounding nasal.

1.3.1 Acoustic Features of Nasal Consonants

Dang et al. [3] examined the shape of the nasal tract using MRI technology. Then they used the data collected to do some computer simulations. From these they predicted what the spectra of the murmur during a consonant should look like and compared the predictions with measured spectra from speech recorded by the subjects.

What they found was that in order to get a good match between the simulated spectra and those from the recordings, they had to use a more sophisticated model than the one represented by figure 1-2 for doing the simulations. The fact that the nostrils are not perfectly symmetric causes extra pole-zero pairs to be introduced, and the sinuses also introduce pole-zero pairs. These extra pole-zero pairs cause the spectra of the recorded murmur to be very *bumpy* with many small peaks and valleys. The model used by Dang et al., called the dual-tube model, uses two tubes to model the nostrils instead of treating nostrils as one tube. An example of a measured spectrum of a nasal murmur together with the calculated spectra using the dual-tube model, with and without sinuses, is displayed in figure 1-3. It was not shown that the bumpiness, present in the measured spectra and the calculated spectra was important perceptually for identifying or recognizing the consonants.

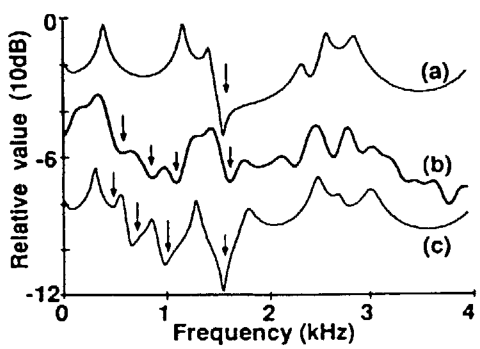


Figure 1-3: Taken from Dang et al. [3]. The transfer function from the glottis to the nostrils for /n/ (Arrows indicate zeros or dips.) (a) calculation from a dual-tube model without sinuses; (b) Spectrum obtained from real speech signals by the cepstrum method; (c) calculation from a dual-tube model with sinuses.

Murmur Spectra

Fujimura [4] did some experiments in which an attempt was made to fit the poles associated with the formants, together with a low-frequency pole-zero pair associated with nasalization to some spectra of nasal murmurs. The locations of these poles and the zero were chosen to fit within some theoretically predicted range, for both their frequencies and their bandwidths.

It was shown that the locations of these poles and the zero were different for the three consonants and also for different vowel contexts. It was also shown that for the duration of the murmur, the frequencies can move around, especially when the preceding vowel is different from the final vowel. Quite a large variability was also shown for different trials using the same utterance.

To grossly generalize Fujimura's findings, there was always a very low first formant, somewhere around 300–400 Hz, and another around 1,000 Hz. For /m/, there was also another pole and a zero close to 1,000 Hz, while for /n/, this pole and zero were between 1,500 Hz and 2,000 Hz. Figure 1-4 is taken from [4] and shows the frequencies of the first four resonances and one anti-resonance for several different vowel contexts and for both /m/ and /n/.

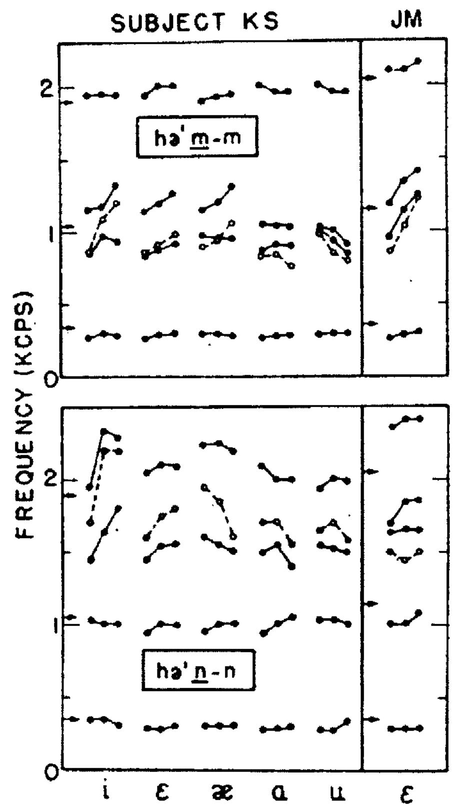


Figure 1-4: Taken from [4]. Locations of poles (closed circles) and zeros (open circles) during the murmur of the first /N/ in utterances /həNVN/. Each triplet represents values near the beginning, near the middle, and near the end of the murmur. Arrows indicate the locations of poles for each subject for utterances of /ŋ/.

It was also noted by Fujimura that the bandwidths of the formants were large, and that the formants were closely spaced in frequency. These characteristics cause the acoustic energy distribution in the middle-frequency range to be fairly even, or *flat*. What is meant by *flat* is that the spectrum does not have any large significant prominences. Many small peaks are allowed. In other words, the murmur spectra can be both bumpy and flat at the same time.

It is interesting to note that even with such a simplistic model which included only one zero (compare with Dang et al. [3] discussed above), fairly good spectral matches were found. The details of the spectra were slightly different, but in terms of overall shape, the matches were good.

Abruptness

We are dealing here with consonants, and so we would expect there to be some abruptness associated with the acoustic signal. Stevens ([9] chapter 9) looked at the amplitudes of the upper formants and found that these exhibit the kind of abruptness associated with consonants.

For /m/, it was found that the amplitude of the second formant jumped by about 20 dB within 10–20 ms. The higher formants experienced similar jumps, but it is expected that the second formant is the most important perceptually. For /n/, this was even more significant, with a jump of about 25 dB in second formant amplitude. For /ŋ/, the jump was also quite large, and comparable to that for /n/.

This abruptness in the amplitude of the second formant must be largely due to changes in the bandwidth of the second formant or changes in the locations of the poles and zeros which have frequencies that are lower than that of the second formant, or both. The roll-off rate in the frequency response of a single pole is 12 dB per octave, and so, if a formant drops to half its original frequency, this causes a drop in amplitude of everything above it by about 12 dB (ignoring the local effect of the peak).

For a typical /m/, the frequency of the first formant drops to about half its value when the consonant closure is made. This depends on what the adjacent vowel is, and for a high vowel, we would not expect such a significant change in the first formant. This would account for a discontinuity of about 12 dB in the amplitude of the second formant. There is still another 12 dB which is not accounted for. It is thought that this is due to the sudden jump of a low frequency zero (which basically has the same effect as the fall of a pole). This could also be due to some partial pole-zero cancellation, if the second formant is near to a zero during the murmur.

Chapter 2

Computer Simulations

In order to estimate the frequencies and bandwidths of the poles and zeros of the transfer function, some simulations were done on computer. These simulations were done by using a computer to calculate the locations of the poles and zeros from area functions of the vocal tract and nasal cavity.

2.1 Lumped Element Model

The first attempt at the computer simulations was made by using lumped elements to model each short section of tube. One capacitor and one inductor was used for each section of length 1 cm. Because of the approximations involved in using lumped elements, this model was only theoretically valid up to about 2,000 Hz. There were also practical problems because of numerical round-off errors.

The frequencies of the poles and zeros were found by actually calculating the susceptances B_G, B_M and B_N for the three different tubes (see figure 1-2) assuming a short circuit at the lips and nostrils, and a open circuit at the glottis. From these susceptances, frequencies were found as described in section 1.2.2. This model was quickly abandoned because of its complexity and relatively poor performance.

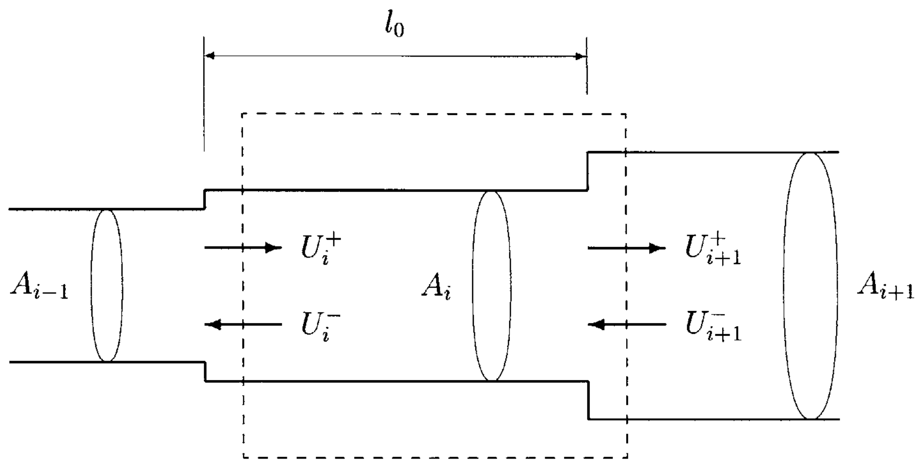
2.2 Transmission Line Model

In order to find the locations of poles and zeros which occur at higher frequencies, a different model was used for the final computation. Rabiner and Schafer [8] viewed the vocal tract as a concatenation of lossless tubes of equal length. Each tube was modeled as a short section of a transmission line. The characteristic impedance of the ith segment, Z_i is related to the cross-sectional area of that segment, A_i, by $Z_i = \rho c / A_i$, where ρ is the density of air and c is the speed of sound.

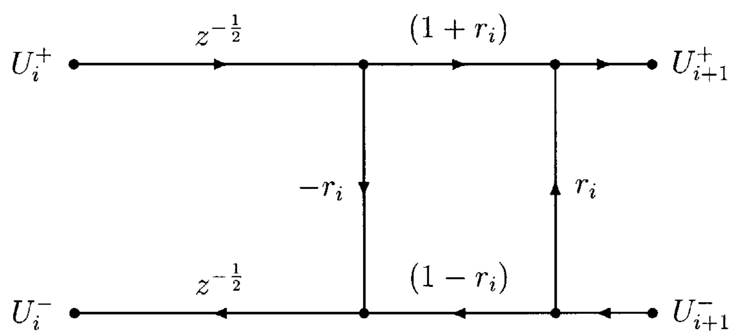
At the junctions between the sections, a part of the signal traveling down the line is reflected. This part is determined by the reflection coefficient associated with the junction. For the ith junction, which is between the ith tube and the $(i+1)$th tube, the reflection coefficient, r_i, is calculated by $r_i = (A_{i+1} - A_i)/(A_{i+1} + A_i)$. For each section of tube, there is a forward traveling wave and a backward traveling wave, each of which gets partially reflected at the junctions.

Each section of tube is treated as a simple delay. In the z-transform domain, this is just a multiplication by z^{-d}, where d is the delay measured in units of one sampling period. If the lengths of the tubes are all equal, and the sampling period is chosen so that the time it takes for sound to travel a distance equal to the length of one tube is half of a sampling period, this delay becomes a multiplication by $z^{-\frac{1}{2}}$. The length of a tube, l_0, and the sampling period, T, are related by $cT = 2l_0$. The ith tube and the ith junction can be represented by the flow graph shown in figure 2-1. U_i^+ is the *forward* traveling component of the volume velocity which is just about to enter the ith tube. U_i^- is the *backward* traveling component, which is just leaving the ith tube. *Forward* is taken to mean the direction moving toward the lips (or nostrils), and *backward* is the opposite direction, moving toward the glottis.

The entire tube can then be represented by the signal flow graph shown in figure 2-2. At each junction, a part of the wave flowing into the junction is transmitted and a part is reflected. There are two waves entering a junction, and two leaving. Each wave which is leaving the junction, is a weighted sum of the two waves entering it.



(a)



(b)

Figure 2-1: (a) The ith tube connecting the $(i-1)$th tube to the $(i+1)$th tube. The ith junction is the junction between the ith tube and the $(i+1)$th tube. The dashed box encloses the ith tube and the ith junction. (b) The flow graph representation of the ith tube and the ith junction.

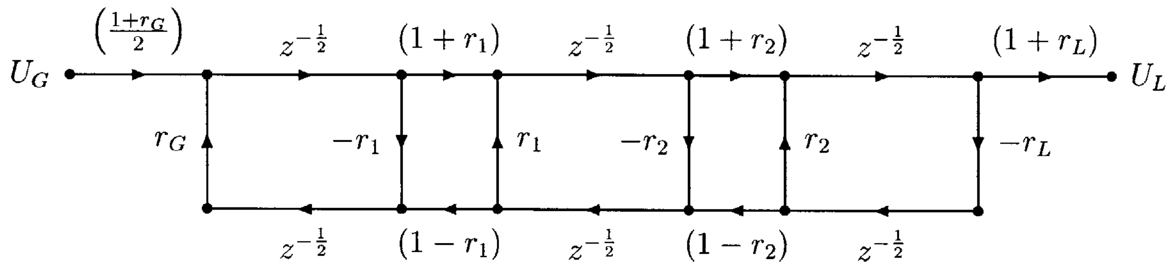


Figure 2-2: Flow graph used for reflection line model, from Rabiner and Schafer [8] page 90. The tube is made of three sections, each of which can be represented as shown in figure 2-1(b).

2.2.1 Nasalization: Side Branch

The transfer function of the system from the glottis to the lips can easily be found from the flow graph in figure 2-2, but if we also include a side branch to account for nasalization, the problem becomes more complicated. It can be shown that the point where the tube splits can be represented by the flow diagram in figure 2-3(b).

From the flow diagram, we see that this three-way junction behaves similarly to the junction shown in figure 2-1. There are three waves flowing into the junction, each of which gets split into three parts and assigned to the three waves flowing out of the junction. So, each of the three waves flowing out of the junction is a weighted sum of the waves flowing in.

Here we use a slightly different reflection coefficient. Actually, three different coefficients are need in this case. The reflection coefficient, r_X is defined by

$$r_X = \frac{2A_X}{A_G + A_M + A_N} - 1$$

One of the limitations of the transmission line model is that the tube lengths, l_G, l_M and l_N (as shown in figure 1-2, have to be integer multiples of the length of one tube section, l_0. Furthermore, the difference between the lengths of the tubes for the

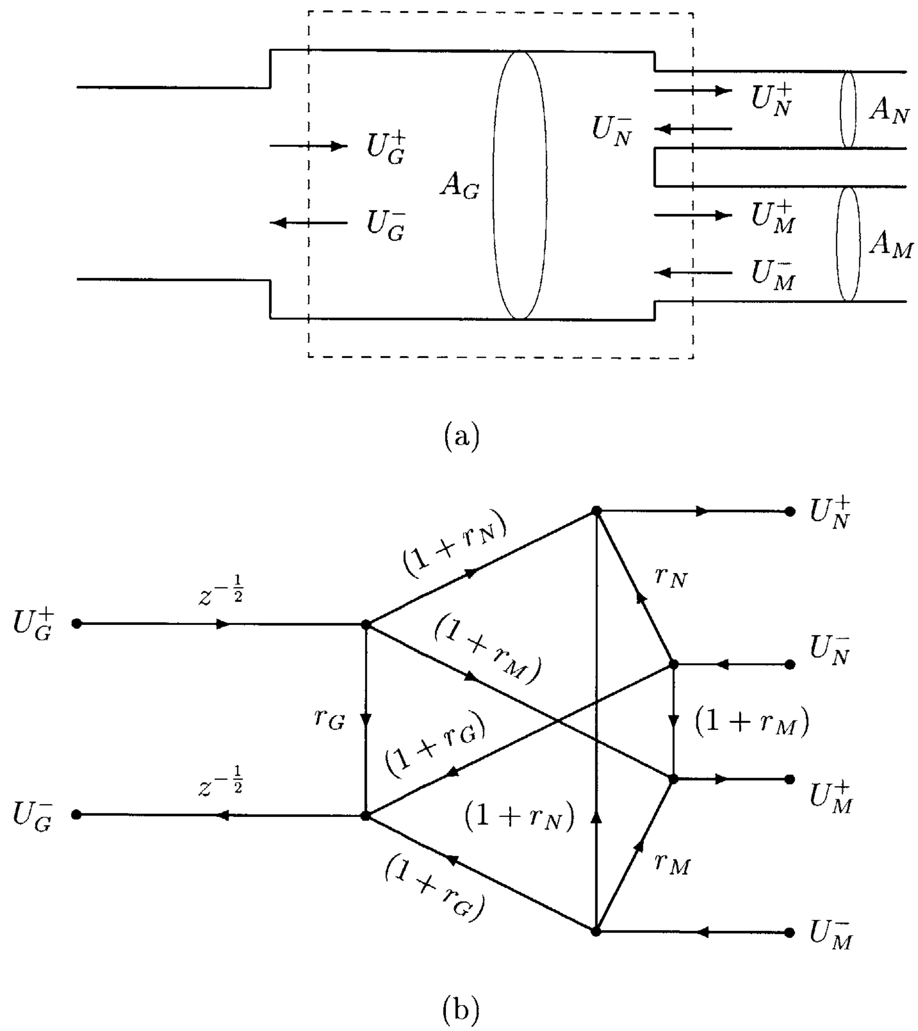


Figure 2-3: (a) The last section of the main tube, including the splitting point. The dashed box shows what portion is drawn as a flow graph. (b) Part of the flow graph used for reflection line model modified to include a side branch. This graph shows the branch point. The junction included here has three waves flowing into it and three waves flowing out.

nasal and oral cavities must be an even multiple of l_0, i.e., the quantity

$$\frac{(l_N - l_M)}{2l_0}$$

must be an integer. This is because the final system function, $H(z)$, must not have any fractional powers of z.

2.2.2 Bandwidths: Losses

Thus far, all of the models used have ignored losses. Let us now consider how some losses can be included in the model in order to be able to predict both frequencies and bandwidths.

Johan Liljencrants [6] deals with the issue of losses in some detail. Here we will make some approximations to simplify the problem.

Radiation at the Lips and Nostrils

Up to this point, we were assuming that the lips and nostrils were terminated by a short circuit. Now we will be a bit more careful.

The acoustic radiation impedance looking out into the rest of the world can be approximated by a frequency dependent resistance and a reactance. For a z-domain model, Liljencrants used one pole and one zero. The impedance looking out was taken as

$$Z_{rad} = \frac{\rho c a(z-1)}{A(z-b)}$$

where

$$a = 0.0779 + 0.2373\sqrt{A}$$

$$b = -0.8430 + 0.3062\sqrt{A}$$

Using this impedance, which is dependent on area of the opening, A, we can obtain a reflection coefficient at the lips and one at the nostrils. This model introduces an extra pole and zero in the final calculated transfer function, but these only affect the overall shape of the spectrum and mimic the effect of the radiation characteristic. The resistive part of the impedance is what affects the bandwidths.

Loss at the Glottis

Similarly, we were also assuming that the glottis was a perfectly reflective hard wall, but now we will treat it as resistive, with resistance, $R_G = p_G/U_G$, where p_G is the pressure drop across the glottis and U_G is the volume velocity. We can use D.C. values here since we are assuming that the resistance is independent of frequency. For $p_G = 10$ cm of H_2O and $U_G = .25$ l/s, and cross-sectional area of the first tube, $A_0 = 2$ cm^2, the reflection coefficient at the glottis, $r_G = 0.7$. Remember that this is a gross approximation, and the results obtained may not be completely accurate. We will need to confirm our findings with empirical observations.

Series Losses in the Line

Figure 2-4 shows some of the important loss mechanisms for a typical vocal-tract (uniform tube of length 17.5 cm and shape factor 2). Cross-sectional area and frequency are varied, and the regions where different loss mechanisms dominate are labeled. The figure is only valid for small DC flow velocity, where the flow is laminar instead of turbulent. *Radiation* is the radiation at the lips. *Viscous* is the loss in the boundary layer at the tube wall due to the viscosity of the air. *Laminar* is a purely resistive loss which represents a certain resistance per unit length on the line. *Wall vibration* is due to the loss of energy that occurs when the sound waves induce vibrations in the vocal-tract wall.

The viscous loss factor is proportional to the square root of frequency, $f^{\frac{1}{2}}$, while the wall vibration loss factor is proportional to the inverse of the square of the frequency, f^{-2}. These frequency dependencies are difficult to implement, and make the problem much more difficult. As a result, we will ignore the frequency dependence, and replace the viscous loss and the wall vibration loss by a catch-all loss which is independent of frequency. This loss factor will take on the minimum value of the loss factors in figure 2-4 for a fixed frequency. Thus we will ensure that the real bandwidths are not

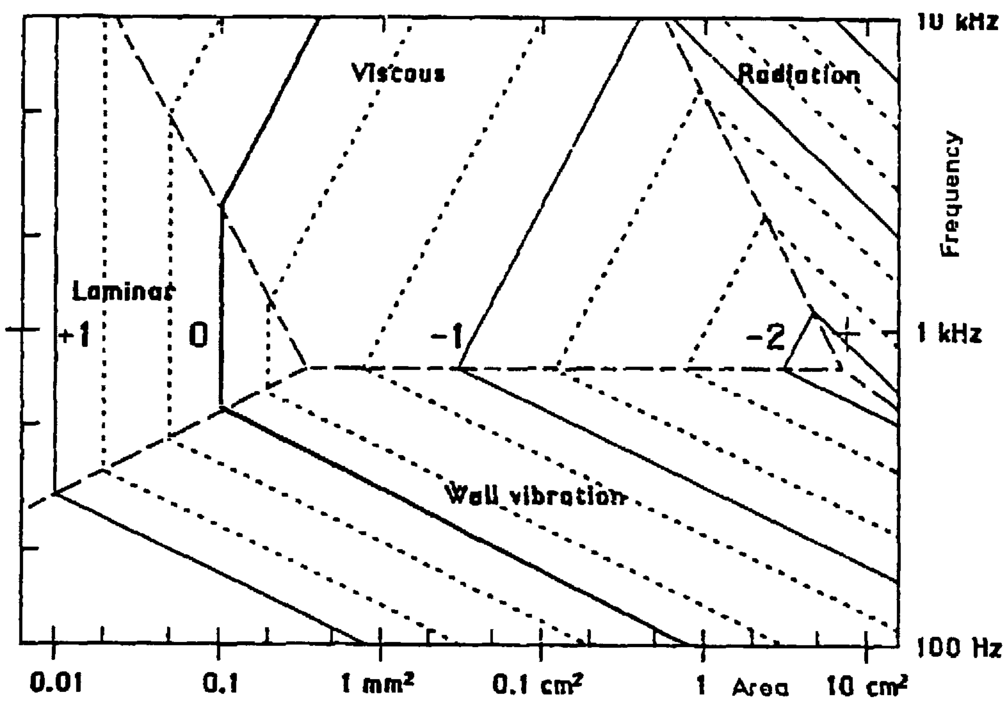


Figure 2-4: From Liljencrants [6]. Comparison of different loss mechanisms as a function of cross-sectional area and frequency for a uniform tube of length 17.5 cm. Contours are lines of equal loss factor. The different regions show where the different loss mechanisms dominate.

less than those produced by the computer simulations.

In order to justify the approximation made above, let us bear in mind that the purpose of doing these simulations is to get a general idea of how the frequencies and bandwidths change as the consonant closure is made or released, and so the absolute value is less important. It may be that because of these and other approximations associated with making an engineering model, the actual frequencies and bandwidths of the poles and zeros found will not be completely accurate, and so the simulations should only be used as a guide in doing the synthesis.

2.3 Examples and General Observations

The simulation method described above was used track the movement of the poles and zeros during simulated utterances of some /VN/ segments. This was done, for a particular time instant, by calculating the frequencies and bandwidths of poles and zeros associated with some area function of the vocal and nasal tracts. This area function was then modified and the simulation was repeated for the next time instant. For example, to do the simulation for the segment /ɪm/, the area function of the vocal tract was chosen to be that for the vowel /ɪ/, and the first section of the nasal area function was given zero area, representing a closed velopharyngeal port. The simulation was done to obtain information on the poles and zeros. The first value in the nasal area function was then increased slightly, corresponding to opening the velopharyngeal port, and the simulation repeated. This was done for several area values. Also, the last value in the vocal tract area function was decreased slowly to zero, representing the closure of the lips for the consonant /m/. Information from these simulations was then plotted as a function of time, as shown in figure 2-5. Simulations for the other /VN/ utterances are given in figures 2-6 through 2-8.

Figure 2-9 shows how the areas of the lips and the velopharyngeal port were varied for the simulation of /ɑm/ to produce figure 2-7. For all utterances, the areas were varied in much the same way. The velopharyngeal opening was varied linearly from zero to its maximum value of about 0.35 cm^2, from time instant -180 ms to -30 ms.

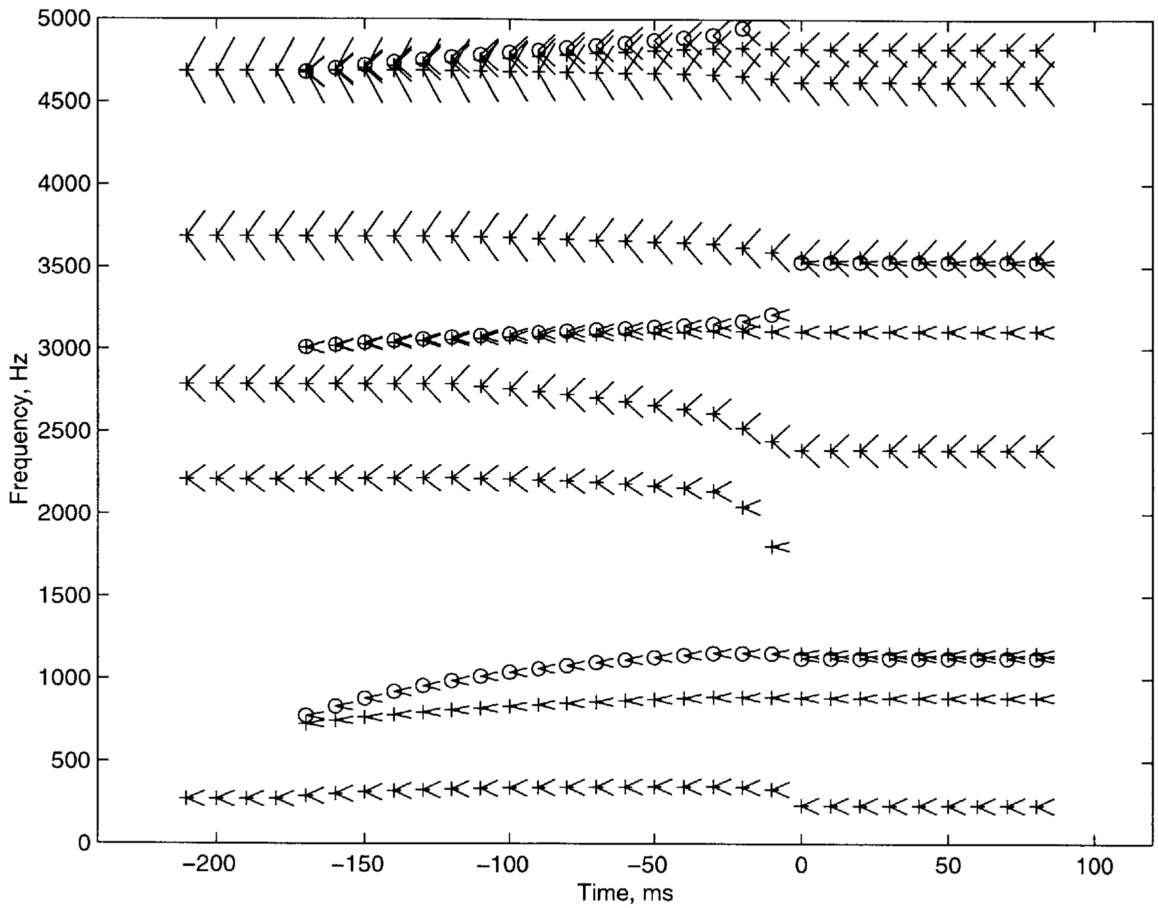


Figure 2-5: Simulation for the utterance /ɪm/. The circles represent zeros and the plus signs represent poles. The slanted lines indicate the bandwidths of the poles and zeros.

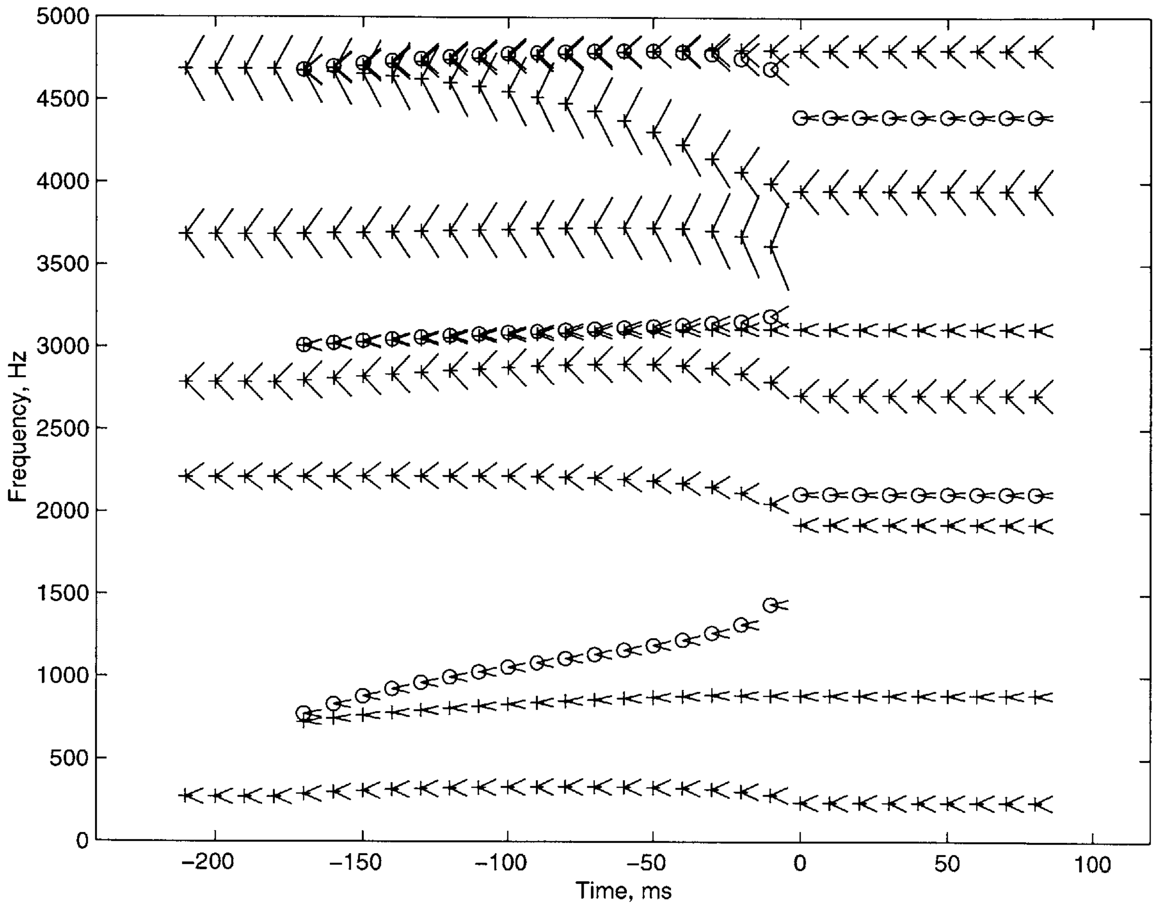


Figure 2-6: Simulation for the utterance /ɪn/. The circles represent zeros and the plus signs represent poles. The slanted lines indicate the bandwidths of the poles and zeros.

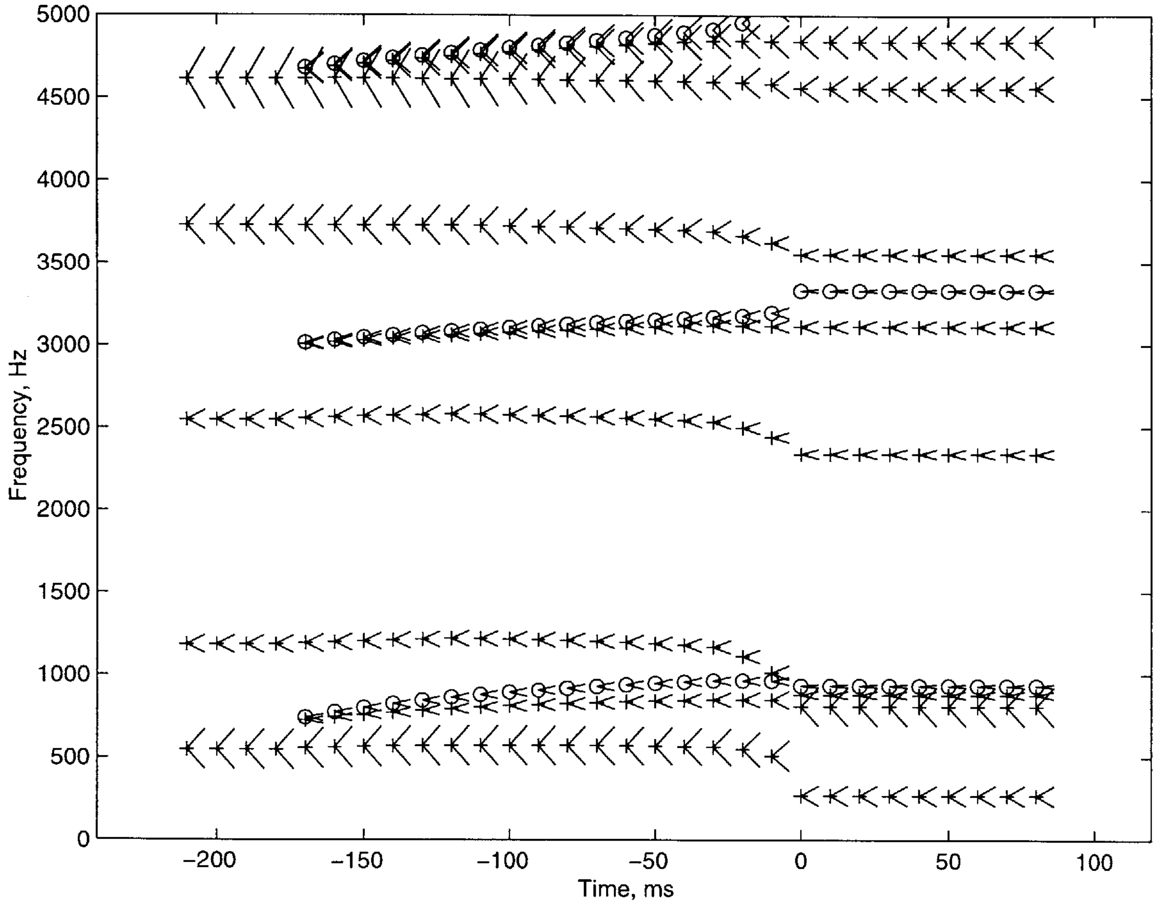


Figure 2-7: Simulation for the utterance /ɑm/. The circles represent zeros and the plus signs represent poles. The slanted lines indicate the bandwidths of the poles and zeros.

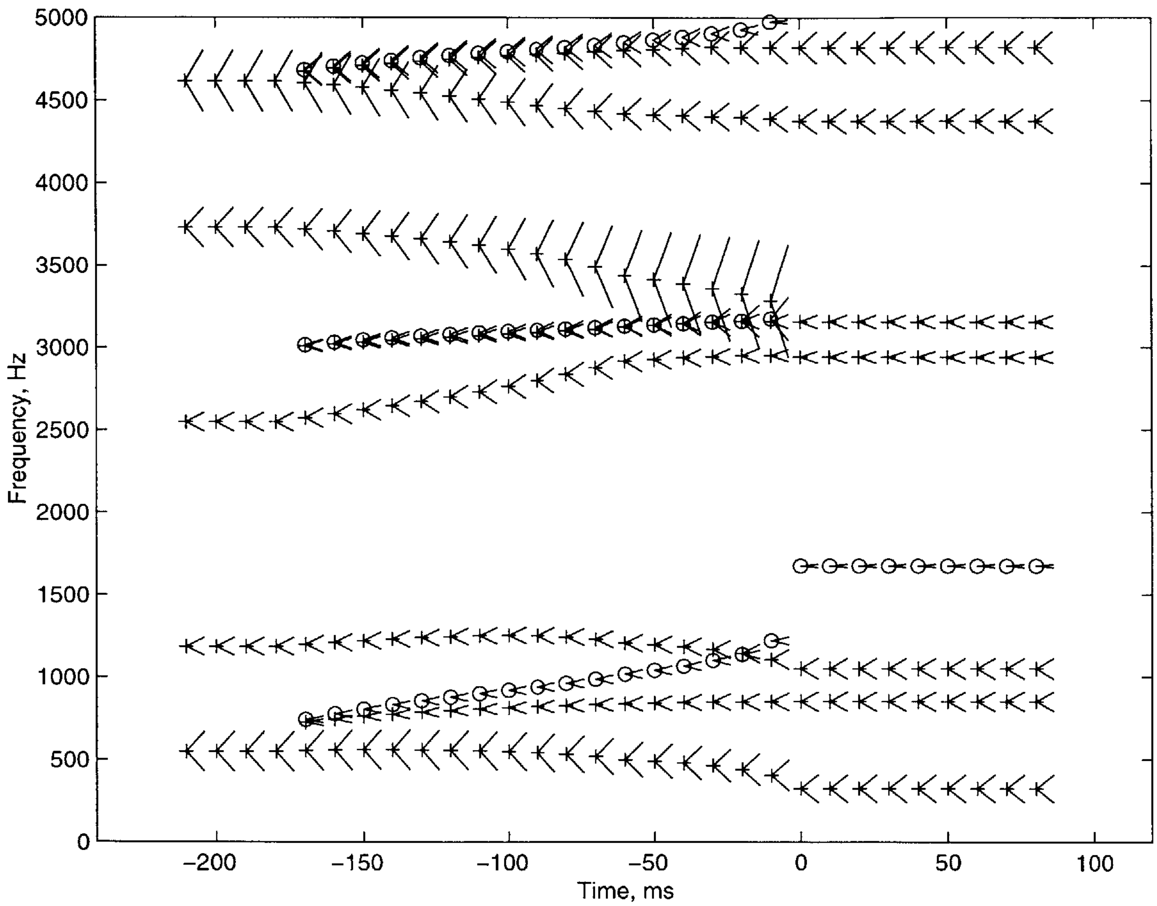


Figure 2-8: Simulation for the utterance /ɑn/. The circles represent zeros and the plus signs represent poles. The slanted lines indicate the bandwidths of the poles and zeros.

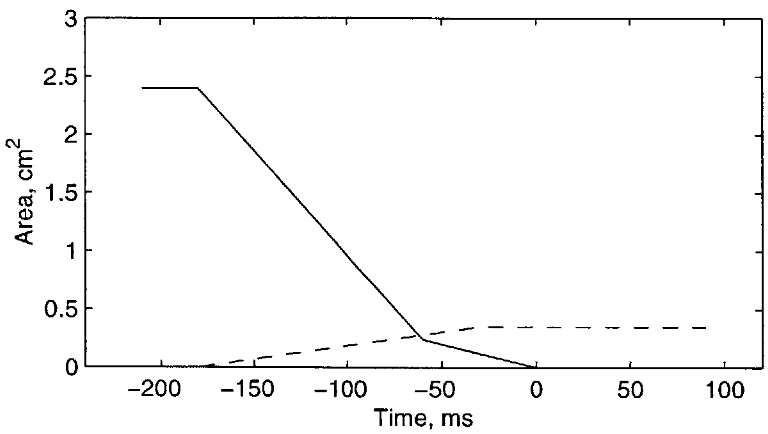


Figure 2-9: For the simulation of the utterance /ɑm/, two area values were varied. The first tube in the nasal area function (shown by the dashed line), and the last tube in the vocal tract area function (shown by the solid line.

The area of the constriction in the oral cavity made with either the lips or the tongue blade was varied as shown in figure 2-9, except that the maximum value was different in each case. For /m/, the last (17th) tube in the vocal tract area function was varied, while for /n/, the third-to-last (15th) tube was varied. The maximum value was what the area of the respective tube would be based on the area function associated with the vowel (see below).

2.3.1 Area Function Data

The area function data for the vocal tracts were obtained from Baer et al. [2]. The data for the nasal tract area function were obtained from a similar study by Dang et al. [3]. In both cases, a gross average of the area functions of the different subjects was used. The area function of the vocal tract which was used for the vowel /ɪ/ is shown in figure 2-10, and for /ɑ/ in figure 2-11. The area function which was used for the nasal cavity is shown in figure 2-12.

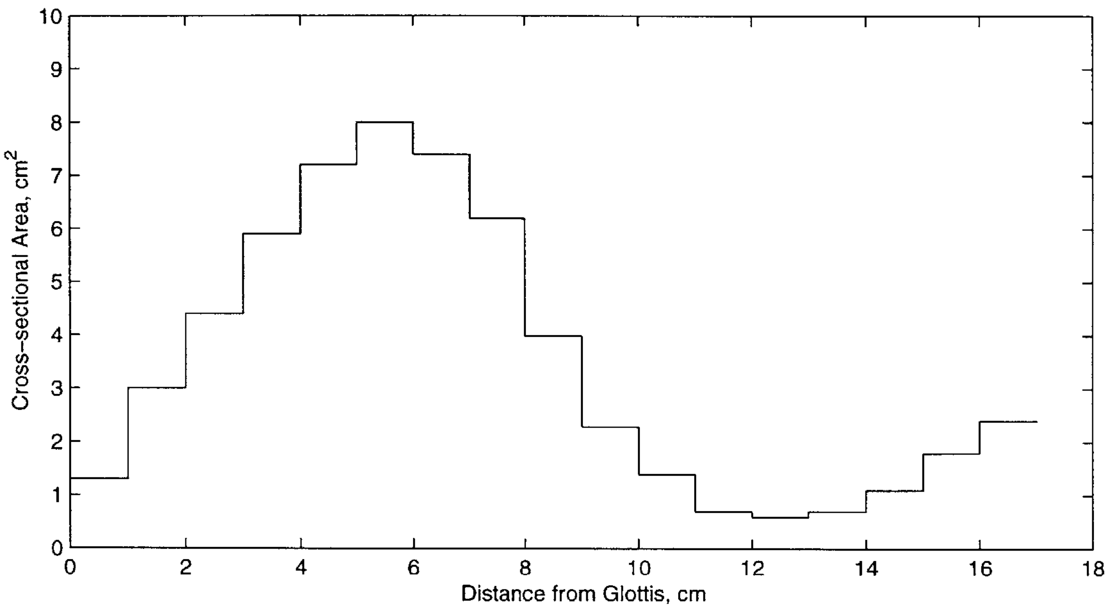


Figure 2-10: Vocal tract area function for /ɪ/.

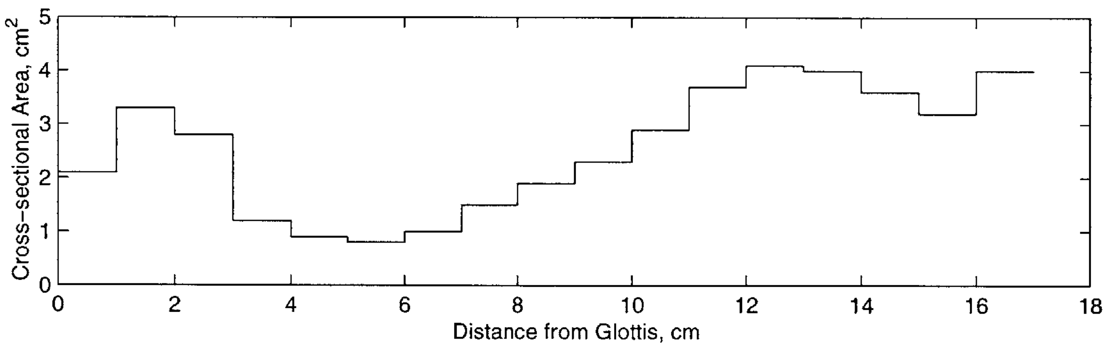


Figure 2-11: Vocal tract area function for /ɑ/.

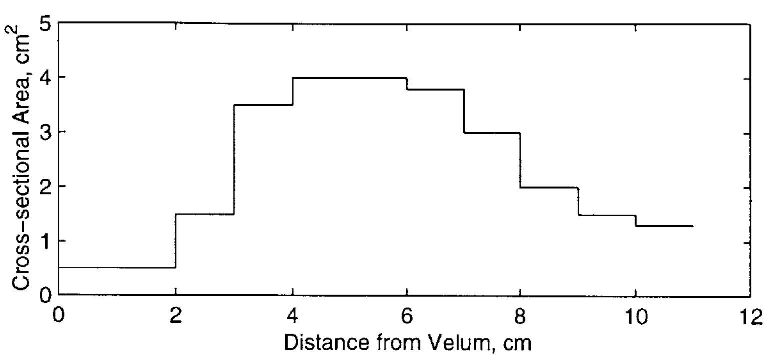


Figure 2-12: Nasal cavity area function.

2.3.2 Comparison with Empirical Data

We can see that we do get good correspondence between the results from the simulations and with observations based on real speech. We see from figures 2-5 through 2-8 that the locations of the poles and zeros during the murmur are close to what was obtained from recorded speech by Fujimura [4] (see figure 1-4). We also see that the the pole-zero locations obtained from an utterance of the word **bender** by Stevens [9], shown in figure 2-13, are very close to what our simulations produced (see figure 2-6).

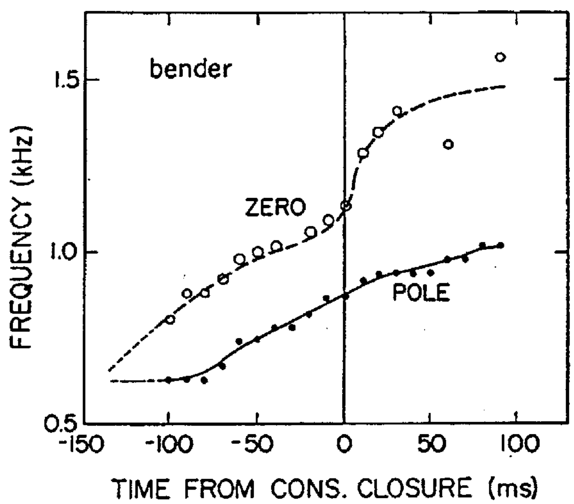


Figure 2-13: Pole and zero locations near the closure of the /n/ in the word **bender**.

Chapter 3

Synthesis and Perceptual Tests

In order to see how important the theoretically predicted pole-zero pairs are to the synthesis of nasal consonants, a simple experiment was performed. This involved some synthesis and some listening tests. Here we give details on the method and the results.

3.1 Synthesis

In order to do the synthesis, speech from an American English speaker was recorded and digitized. Utterances of several words were analyzed, and a few chosen for synthesis. Table 3.1 shows the words which were recorded, with the ones chosen to be synthesized in **bold**.

	/n/		/m/	
/ɪ/	tin	**knit**	Tim	**mitt**
/ɑ/	**tawn**	not	**Tom**	Motts
/æ/	tan	gnat	tam	mat

Table 3.1: List of words recorded and analyzed. The words which were also synthesized are shown in **bold**.

Each of the four words was synthesized by two different methods, using the Klatt formant synthesizer described in [5]. In each case, the voiced part of the utterance (i.e. vowel and nasal consonant) was synthesized as described in sections 3.1.2 and

3.1.3. Then the /t/ was taken from the original recording and concatenated with the synthesized portion to produce the word which the subject listened to.

3.1.1 The Klatt Formant Synthesizer

The Klatt synthesizer, described in [5], works in the following way. There are a certain number of parameters, about sixty, for which the user specifies the values. Some of these, about fifty, can be time varying, so the user specifies a time function, not just a single value. These parameters control different parts of the synthesizer.

The synthesizer is based on a simple engineering model of speech production described in section 1.2.1. Some of the parameters control the glottal wave form. For example, the parameter `F0`, controls the fundamental frequency or the *pitch* of the glottal wave form, while `AV` controls its amplitude. Other parameters control the time varying filter which the glottal wave form is passed through. The parameters `F1` through `F5` and `B1` through `B5` control the frequencies and the bandwidths of five of the poles of the filter. There are two more poles and two zeros whose frequencies are controlled by the parameters `FNP`, `FNZ`, `FTP` and `FTZ` and whose bandwidths are controlled by the parameters `BNP`, `BNZ`, `BTP` and `BTZ`.

3.1.2 The Conventional Method

The first method which was used to do synthesis, called the conventional method, used only the first five poles. This is what is typically done in most formant synthesizers.

For an utterance such as /ɑm/, the synthesis is done in the following way. First, the formants are placed so as to get a good spectral match with the vowel; this can be done by measuring resonance locations from real speech, or by consulting a table such as the one given by Peterson and Barney in [7] (TABLE II). Some time before the consonant closure, the bandwidths are increased to mimic the effect of the lowering of the velum. The formants are then varied to follow the formant transitions of a stop consonant. These transitions are described by Stevens, in [9] chapter 7. At the time of the consonant closure, the bandwidths are rapidly increased, and the frequencies lowered to cause a sharp decrease in the spectrum amplitude above 500 Hz.

The actual frequency and bandwidth time functions used to synthesize the the words in the listening experiment are shown in figures 3-1 through 3-4. For the words where the nasal consonant is in the initial position (i.e. knit and mitt), the method is the same except that all functions are time reversed.

If the synthesis is to be done based on recorded speech, which is what was done in this case, the poles are varied in such a way as to get the best possible match between the synthesized and recorded speech. In other words, the poles were not just varied according to the rules for a stop consonant, but were based on the actual recorded utterance. This accounts for the unusual behavior of F2 seen in figures 3-3 and 3-4, where the vowel was slightly diphthongized by the speaker.

3.1.3 The Proposed Method

The second method, called the proposed method, attempts to use a more theoretically based approach. Specifically, it also makes use of one or two extra pole-zero pairs. It is known that the transfer function has zeros in addition to the poles, due to the nasalization. The idea is that a more natural sounding consonant can be produced using this method.

The following rules describe the proposed method.

1. Decide on the formant frequencies and bandwidths for the vowel. These can be obtained from real speech or published data (see [7] for frequency values).

2. Fix the consonant closure time, and allow the formant frequencies to vary such that the transitions are similar to stop consonant transitions (see [9] chapter 7).

3. Introduce a pole-zero pair, at around 750 Hz, sometime before the consonant closure and allow it to slowly rise and separate in frequency so that the zero is slightly higher than the pole right before the consonant closure. The zero should be somewhere around 1,200 Hz, while the pole should be around 1,000 Hz.

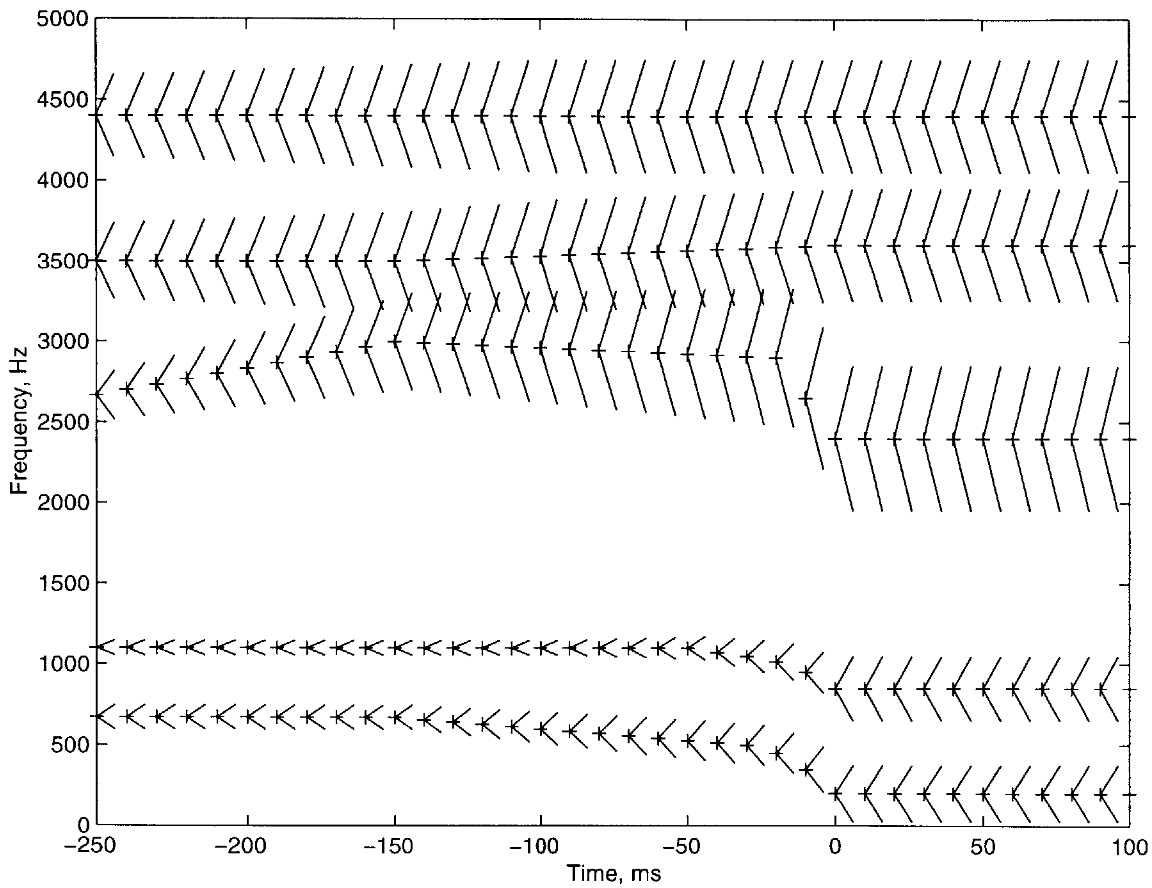


Figure 3-1: Formant frequencies and bandwidths used to synthesize the /ɑm/ in the word **Tom**, using the conventional method. The plus signs represent poles, and the slanted lines indicate the bandwidths of the poles.

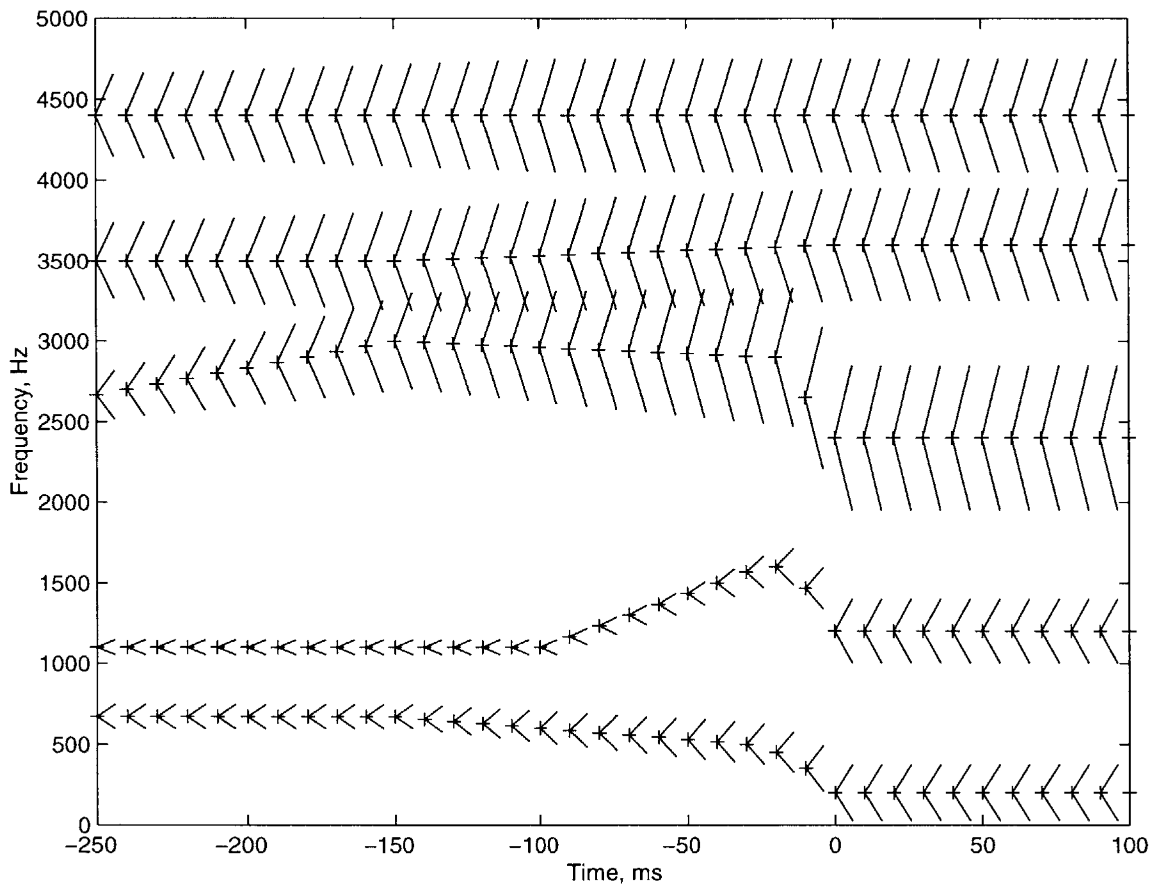


Figure 3-2: Formant frequencies and bandwidths used to synthesize the /ɑn/ in the word **tawn**, using the conventional method. The plus signs represent poles, and the slanted lines indicate the bandwidths of the poles.

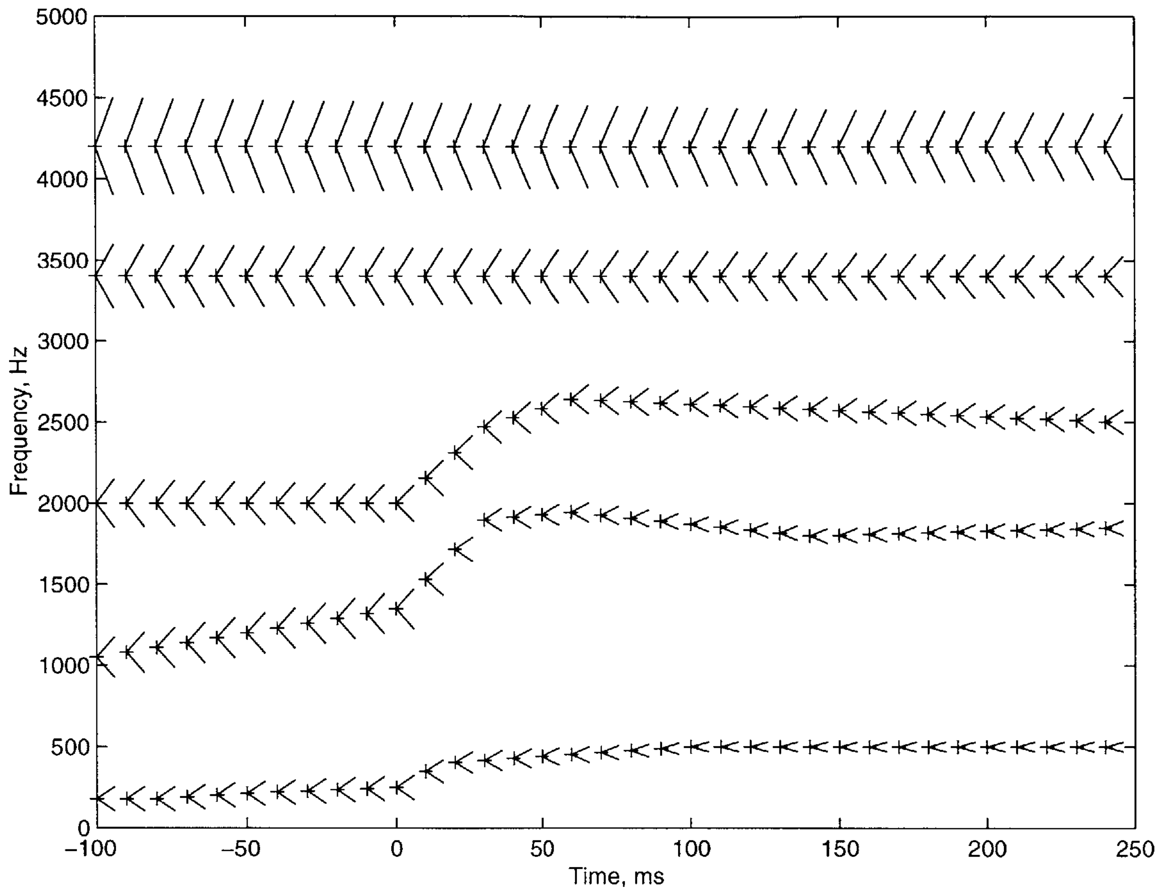


Figure 3-3: Formant frequencies and bandwidths used to synthesize the /mɪ/ in the word **mitt**, using the conventional method. The plus signs represent poles, and the slanted lines indicate the bandwidths of the poles.

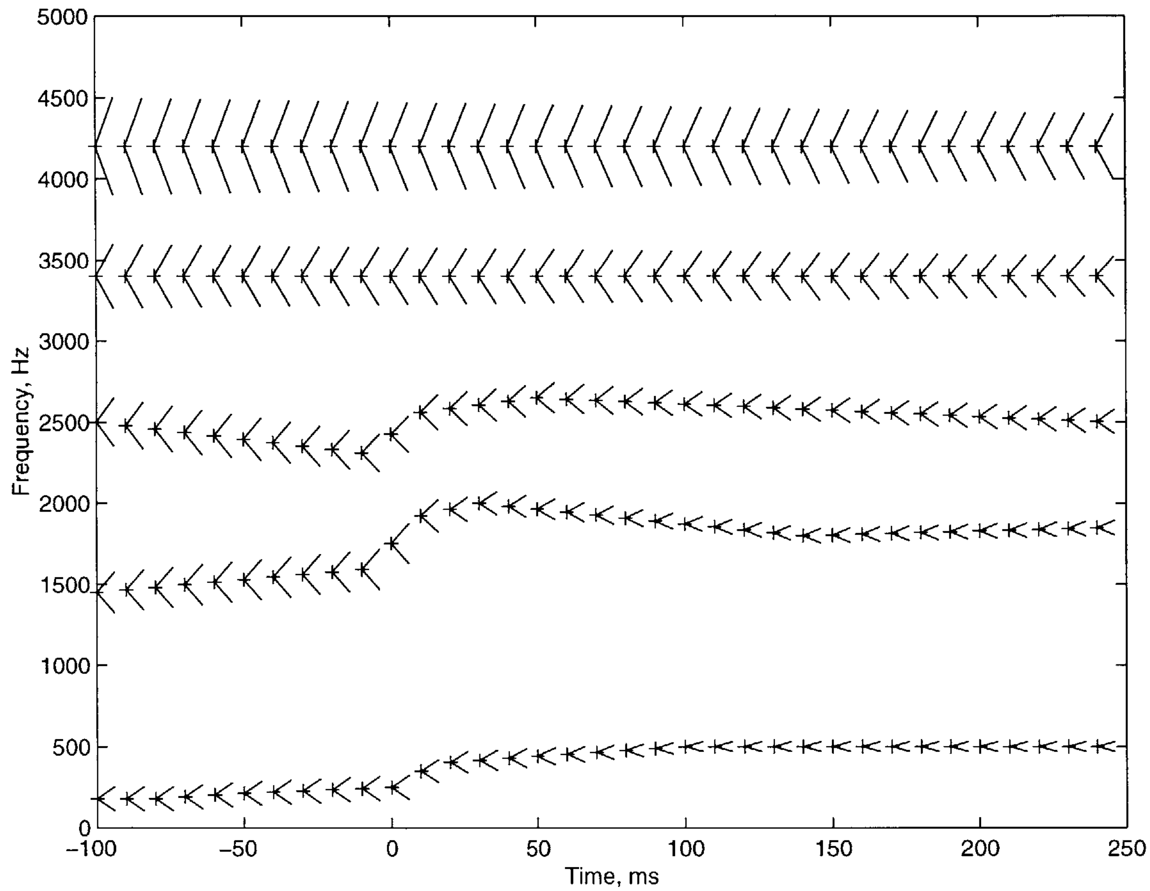


Figure 3-4: Formant frequencies and bandwidths used to synthesize the /nɪ/ in the word **knit**, using the conventional method. The plus signs represent poles, and the slanted lines indicate the bandwidths of the poles.

4. At the point of closure, the zero should make a sharp jump in frequency up to around 1,700 Hz for a /n/, and 1,400 Hz for an /m/. For the case of an /m/, the second formant should fall suddenly so that its frequency is almost equal to the frequency of the zero.

5. Optionally, for a low vowel, a second pole zero pair could be introduced near 250 Hz. At closure, this zero jumps up to cancel the first formant. See figure 3-10(b).

6. The bandwidth of the first and second formants should be increased by about 20 to 50 percent, while the bandwidths of the higher formants should be increased by about 150 percent. These should not be sudden increases, but should start before the consonant closure and should be gradual.

7. If the above movement of formants and the extra poles causes two poles to cross at any time, this should be corrected by only allowing the poles to approach each other and then separate. After this the roles of the poles are exchanged. Two poles should not be too close (within about 100 Hz of each other), unless one is being canceled by a zero.

These rules are based in part on observations of the simulated pole-zero tracks produced in chapter 2. For example, in rule number 3 the actual value of 750 Hz is chosen to be close to what was observed in figures 2-5 through 2-7. The rules are also based on empirical observations of real speech. For example, rule number 5 is based on observation of an extra pole-zero pair described in section 3.1.4. The pole-zero locations during the murmur given in rule number 4 are partially based on the simulations, and partially on the observations made by Fujimura [4] given in figure 1-4.

The actual frequency and bandwidth time functions used to synthesize the the words in the listening experiment are shown in figures 3-5 through 3-8. As before, for the words for which the nasal consonant is in the initial position, the method is the same except that all functions are time reversed. Also, the same diphthongization mentioned before in section 3.1.2 can be seen here in figures 3-7 and 3-8.

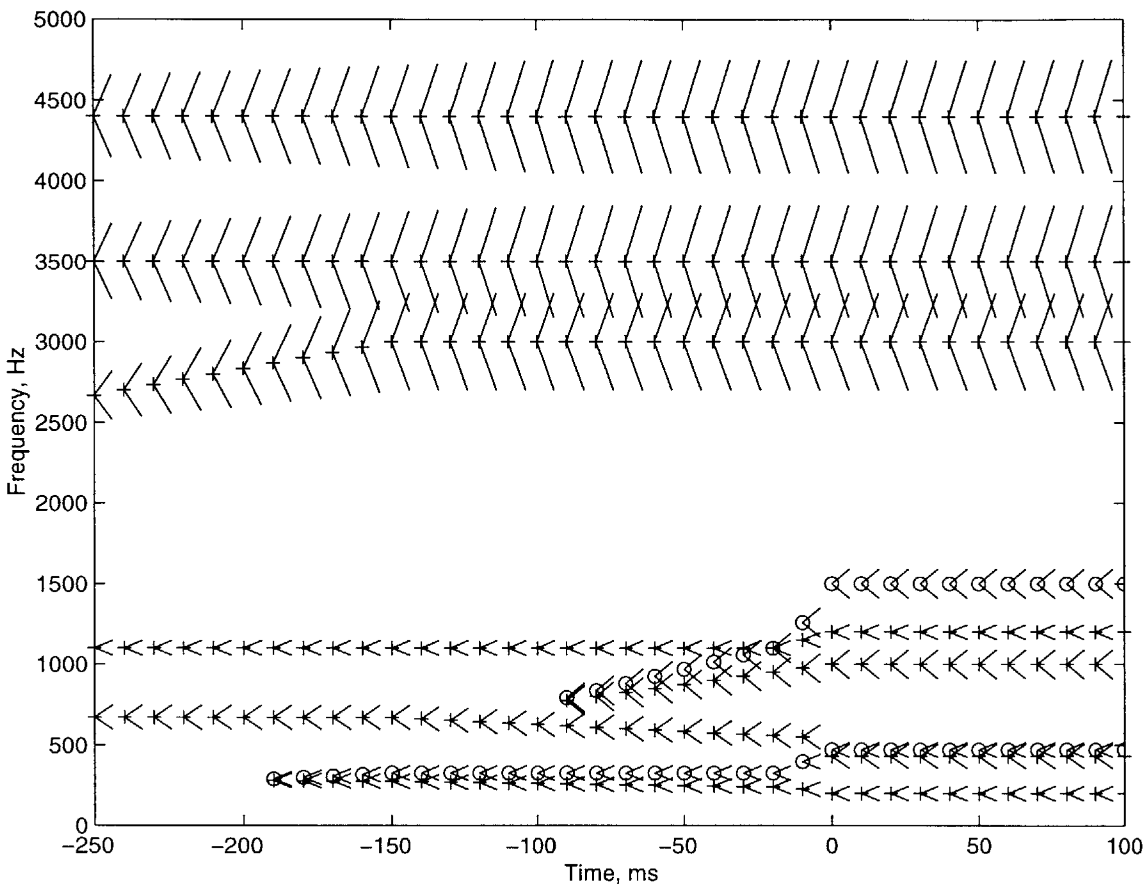


Figure 3-5: Formant frequencies and bandwidths used to synthesize the /ɑm/ in the word **Tom**, using the proposed method. The circles represent zeros and the plus signs represent poles. The slanted lines indicate the bandwidths of the poles and zeros.

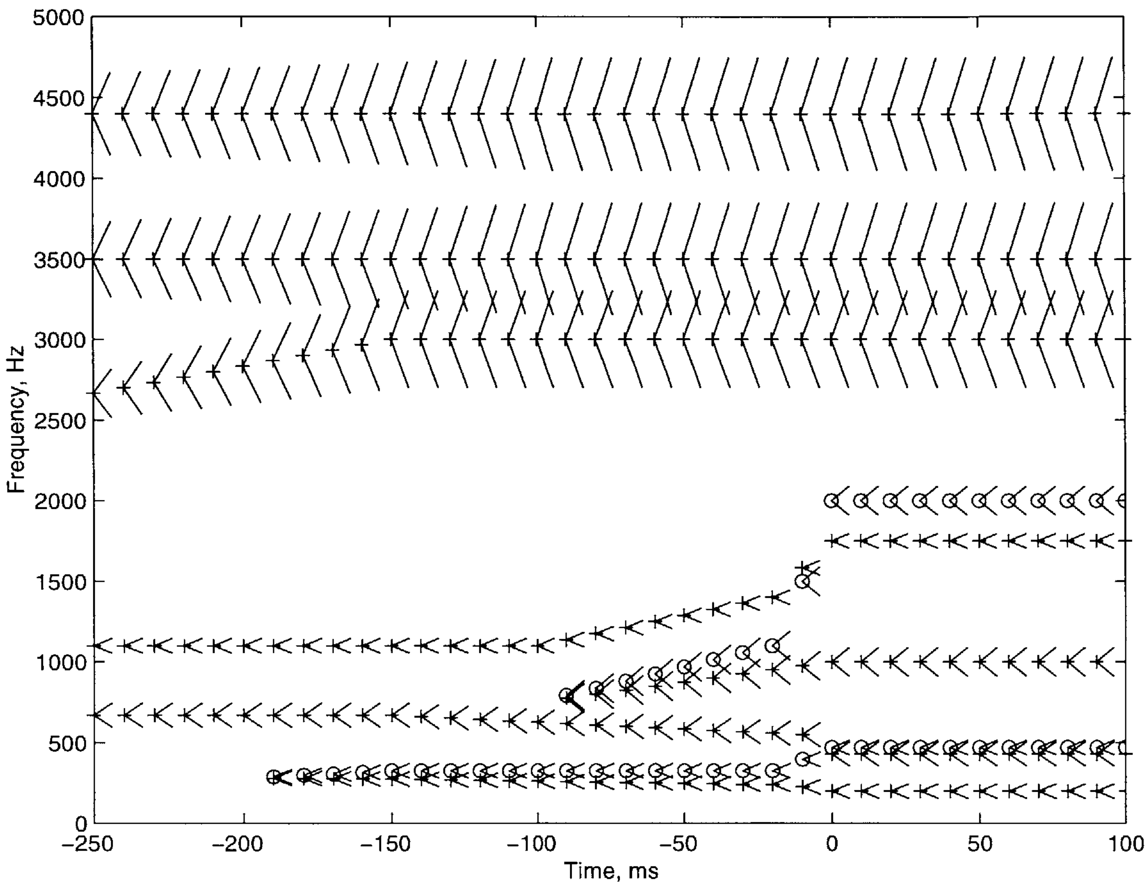


Figure 3-6: Formant frequencies and bandwidths used to synthesize the /ɑn/ in the word **tawn**, using the proposed method. The circles represent zeros and the plus signs represent poles. The slanted lines indicate the bandwidths of the poles and zeros.

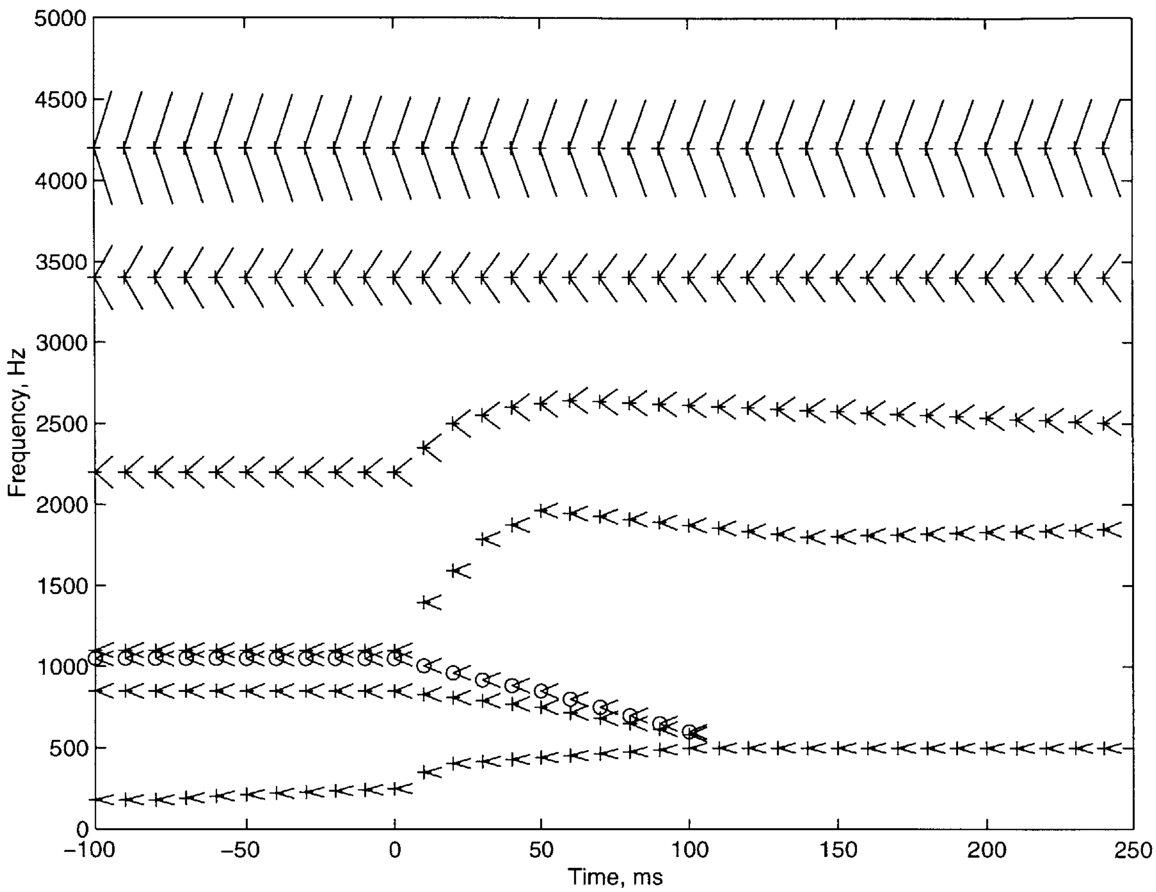


Figure 3-7: Formant frequencies and bandwidths used to synthesize the /mɪ/ in the word **mitt**, using the proposed method. The circles represent zeros and the plus signs represent poles. The slanted lines indicate the bandwidths of the poles and zeros.

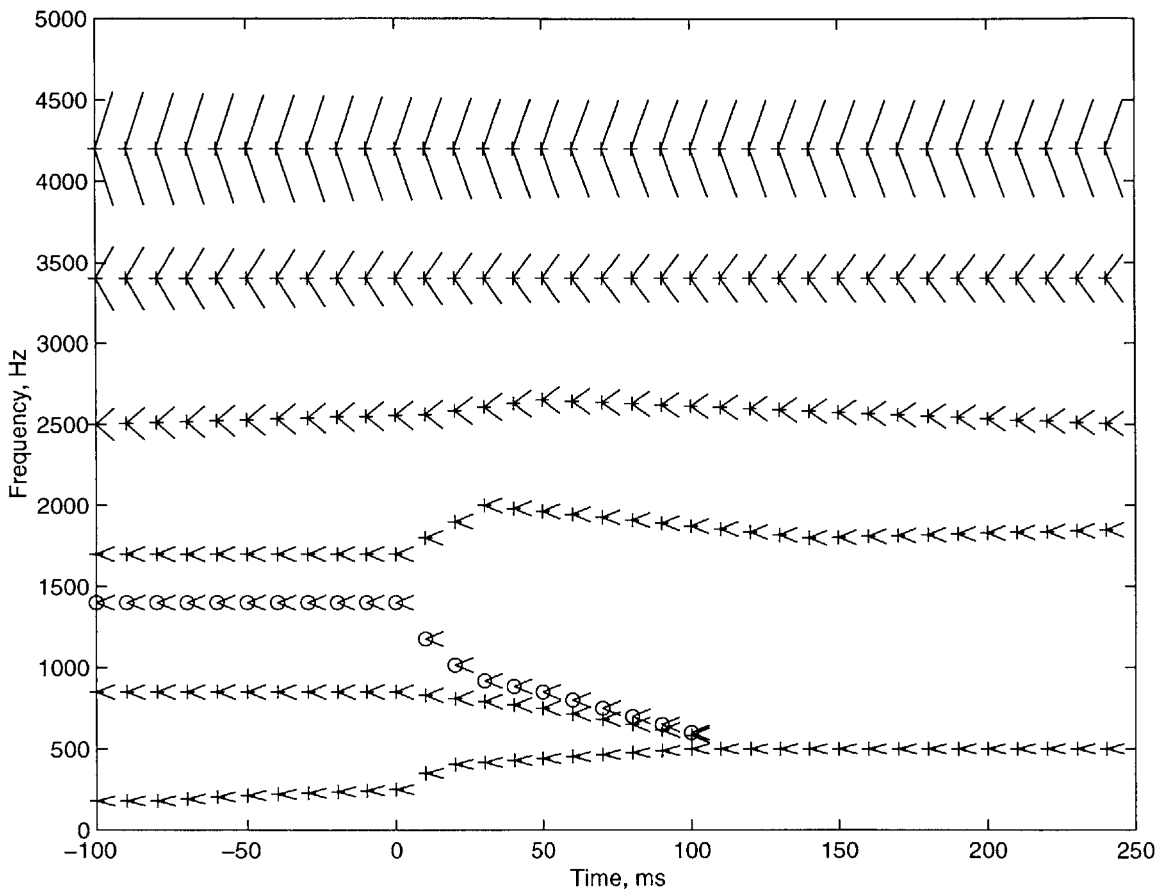


Figure 3-8: Formant frequencies and bandwidths used to synthesize the /nɪ/ in the word **knit**, using the proposed method. The circles represent zeros and the plus signs represent poles. The slanted lines indicate the bandwidths of the poles and zeros.

3.1.4 Observations

By looking at spectral slices of the recorded speech and attempting to match the spectra as closely as possible, it was seen that many more than two zeros would be required. Figure 3-9 shows an example of one of these spectra. Between 600 Hz

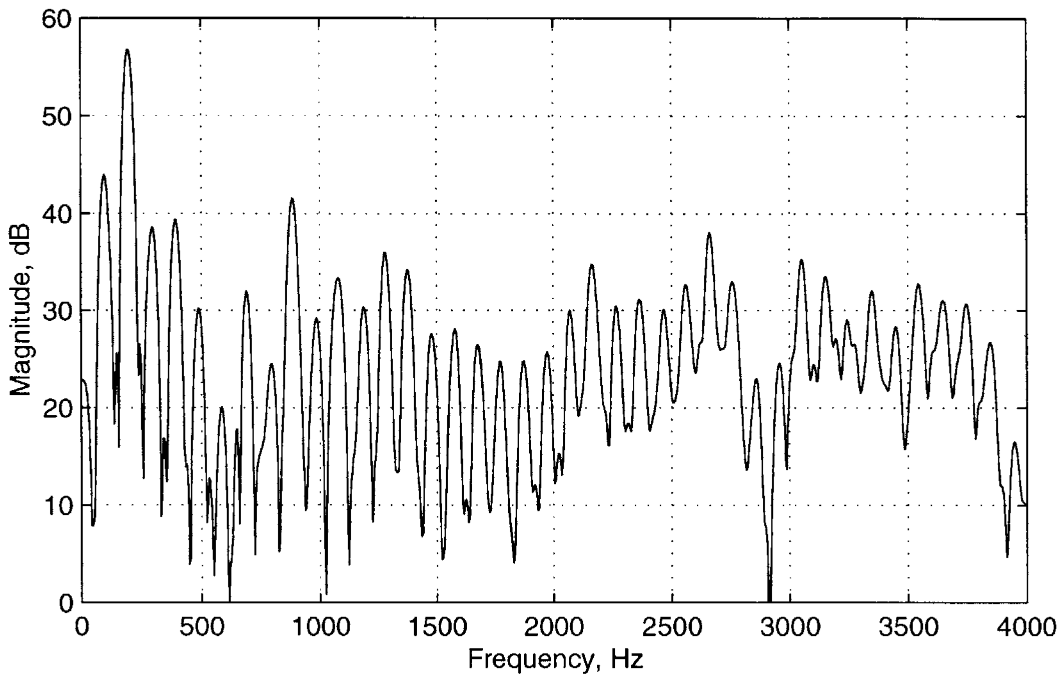


Figure 3-9: A spectrum taken during the murmur for /m/. This is the magnitude of the Fourier transform in dB. Two main zeros are at about 550 Hz and about 2900 Hz. Several small dips can be seen between 600 Hz and 1300 Hz.

and 1300 Hz, the amplitudes of the harmonics seem to alternate between increasing and decreasing. Many more pole-zero pairs appear in the transfer function. These are as a result of acoustic coupling with sinuses and possibly with the trachea. All of these pole-zero pairs have the effect of making the spectrum more bumpy on a highly resolved scale. However, on a much broader scale, the spectrum actually becomes more flat. The periodic nature of the glottal waveform has the effect of sampling the frequency response of the transfer function at integer multiples of the fundamental frequency or the pitch. Because of this sampling, the finely resolved shape of the frequency response is lost, and matching the locations of several pole-zero pairs becomes impossible.

It was also observed that for non-low vowels (i.e., those which do not possess the feature *low*), F1, the frequency of the lowest formant, would fall fairly smoothly from its location during the vowel to somewhere around 250 Hz. This is shown in figure 3-10 (a). However, if the vowel is low, then there is usually a pole-zero pair which facilitates the transition, as shown in figure 3-10 (b).

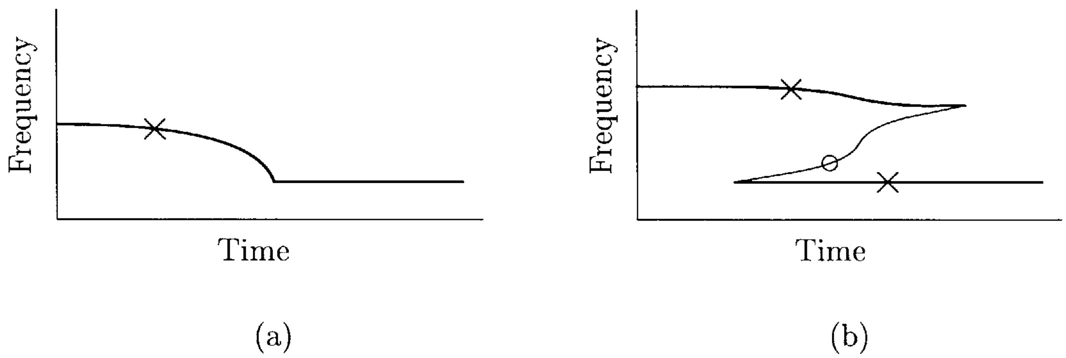


Figure 3-10: The schematic plots show how F1 becomes low at the time of the consonant closure. (a) In the case of a high vowel, F1 just falls. (b) In the case of a low vowel, the transition is made by means of a pole-zero cancellation. The thin line represents the movement of the zero.

This extra zero at such a low frequency has not been fully understood yet. It is thought that it may be due to a sinus resonance. However, attempts made to find the resonant frequencies of the sinuses did not discover a sinus with such a low resonance (see Dang et al. [3]).

3.2 Perceptual Tests

A simple listening experiment was performed to determine whether or not synthesis done using the proposed method sounded more natural than synthesis done using the conventional method.

3.2.1 Procedure

The synthesized words were played through a loudspeaker and the subject was asked to listen and make judgments based on naturalness (i.e., the extent to which the synthesized word sounded like it was spoken by a human). Words were presented to the subject in groups of three. For each group, the same word from table 3.1 was presented three times. The first and last words were synthesized in exactly the same way, and the middle word was synthesized using a different method. Half of the time the proposed method was used to synthesize the middle word, and the conventional method was used for the first and last words. The rest of the time, the opposite was true. For each A-B-A triad, the subject was asked to choose the one which sounded more natural (i.e. A or B). The instructions and the response form with which the subject indicated their choice is given in appendix B.

The words were presented to the subject in eight sets of four triads each. The words used for the four triads in each set were the four bold faced words from table 3.1.

3.2.2 Results

Table 3.2 shows the percentage of the total number of presentations of a particular word for which the proposed method was chosen. From this, we see a clear preference for the proposed method for **knit**, **Tom** and **mitt**.

WORD	Proposed Method
knit	76%
Tom	72%
mitt	63%
tawn	45%

Table 3.2: This is the percent of times for which the proposed method was chosen across all subjects for each word.

These numbers are somewhat difficult to interpret, because they incorporate data from subjects who were just guessing. In order to better capture the fact that subjects who could distinguish between the methods were usually consistent in there answers, the data was also analyzed using a category scheme as follows.

For each of the four words, subjects generally were in one of the following three categories: (category C) they preferred the conventional method, (category P) they preferred the proposed method, or (category I) they were indifferent or were unable to distinguish between the two methods.

Subjects were classified into one of the three categories for each of the four words. A subject was classified in category P if they indicated that the proposed method was preferred more than 75 percent of the time (i.e. for seven or eight of the eight sets). They were classified in category C if they indicated that the conventional method was preferred more than 75 percent of the time. They were classified in category I otherwise. Table 3.3 shows the number of subjects classified into each category for each of the four words.

WORD	Category		
	P	I	C
knit	9	10	0
Tom	8	10	1
mitt	7	10	2
tawn	2	14	3

Table 3.3: Number of subjects classified into each category for each of the four words.

This categorization was done in order to focus on the subjects who were actually able to hear a difference between the two methods, and who demonstrated a preference.

A number of the subjects who could not tell the difference between the two versions for a particular word simply put the same answer for all eight sets. Since the proposed method was used for version A four out of the eight times, as seen from table 3.4, these subjects were classified into category I.

3.2.3 Observations

As indicated by the data, the difference between the two methods was only just noticeable to the untrained listener. It was interesting that some subjects were more sensitive to the differences between the two versions for certain words, while other subjects were more sensitive for other words. For example, one subject always chose

WORD	Prac.	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8
knit	B	A	B	A	B	A	B	B	A
Tom	A	B	B	A	A	A	A	B	B
mitt	A	A	A	B	A	B	B	A	B
tawn	A	B	B	A	B	A	B	A	A

Table 3.4: This table shows which version used the proposed method for each A-B-A triad. This is what was actually presented to the subjects in the listening experiment.

the proposed method for **Tom**, but was unable to distinguish between the two methods for **mitt**, while another subject always chose the proposed method for **mitt**, but was indifferent for **Tom**. This seems to indicate that each person is trained to be perceptually sensitive to different acoustic characteristics of speech.

From informal listening and comparison , it was difficult to distinguish between the two methods. It was also noticed that if the intensity at which the synthesized utterances were played at was set too high, it was more difficult to hear the difference. It seems that at high levels, the listener is less sensitive to the subtle differences. Also, because the experiments were not performed in an anechoic chamber, the room acoustics may have made certain parts of the spectrum more important depending on where in the room the subject and the loudspeaker were placed. This effect is probably not very significant, however.

Given that the number of subjects in category P was larger than for category C (which is some cases was zero), it can be inferred that the proposed method produced more natural sounding synthesized consonants. **Tawn** seems to be an exception, and was consistently the most difficult case to distinguish. Perhaps because it is not a common English word, untrained subjects were unable to make a judgment as to which version sounded more natural. Thus most subjects were in category I and about the same number were in category P as in category C. However, based on informal listening, it seems that the proposed method does sound slightly more natural than the conventional method.

For all cases, an overall preference for the proposed method was shown. However, for some words, there were some subjects who indicated a strong preference for the conventional method. These cases seem to be due to the subjects' personal taste relating to some other aspect of the synthesis.

Chapter 4

Conclusions

Linear predictive coding (LPC) of speech, described in [1], works in much the same way as the Klatt synthesizer. An impulse train is filtered by a linear filter. The most common form of LPC uses an all-pole filter. This raises the question of whether or not LPC does a good job of representing nasal consonants. It seems that it does do a good job, based on listening to the recorded examples accompanying [1]. There are two reasons why this is the case. First, the filter used generally has more poles than is expected for a particular sampling frequency. Below 5,000 Hz, one would generally expect there to be five formants. However, LPC generally uses seven or eight poles and therefore can create a more precise representation of the spectrum. Secondly, it seems that listeners are more sensitive to peaks in the spectrum, rather than dips. This is because of the phenomenon of *masking*. Frequencies containing a lot of energy tend to mask frequencies which are close by. Therefore, spectral peaks tend to be more important, because dips in the spectrum are masked by areas of high energy on either side, while peaks tend to do the masking.

If it is possible to produce a good sounding nasal consonant using an all-pole filter in the model, then why is it important to use theoretically predicted zeros? LPC attempts to match the spectrum of speech in a least-square-error sense. By doing this, there is no guarantee that the poles used in the LPC correspond to the actual poles of the transfer function imposed by the vocal tract. The poles are not even guaranteed to vary continuously as we would expect for the actual transfer function.

We would prefer to have a model for which the poles correspond to something more meaningful, such as physical resonances of the cavities. Well established theory about the locations of these poles and zeros exists.

It is easier and more natural to make rules for synthesis which are based on theoretically predicted poles and zeros than to make rules which govern how poles should be manipulated to match a frequency response which is known to have zeros as well. It may be that rules could be developed based on an all-pole transfer function, but these would have to be somewhat arbitrary, and not based on actual physical resonances of the vocal tract. Also, these rules would cause the poles to move is such a way as to be inconsistent with what established theory says about the movement of the actual poles. Since the poles are related to actual resonances of physical cavities, they must vary as slowly as the cavities can change, and so a very abrupt movement of a pole would violate this expected behavior.

4.1 Summary

An attempt was made to describe the transfer function from the glottis to a point outside the lips and nostrils. By simulating the acoustics of the vocal tract and the nasal cavity, the locations of the extra pole-zero pairs of the transfer function were found. Some predictions of these locations have been made before (by Stevens [9] chapter 6), but a more precise estimate was necessary.

These simulations, in combination with estimates based on analysis of recorded speech, served as a guide for formulating rules for theoretically based synthesis. Synthesis was done by making use of these rules, as well as by doing simple spectral matching. The results of using these two methods were compared by means of a listening experiment. It was found that the synthesis done by using the rules generally sounded more natural. This is in addition to the fact that the rules help to resolve the differences between the established theory and practice. These rules are given in section 3.1.3.

4.2 Further Research: New Idea for Synthesis

As mentioned earlier, the Klatt synthesizer only allows two zeros in the transfer function, but in actuality there are many more due to sinuses and side branches of the vocal tract. Even if we corrected this problem by using a different synthesizer, the complexity of the rules which govern the placement of the poles would grow as the number of poles increases.

A better way to approach this problem is to calculate the locations of the pole-zero pairs directly, instead of relying on rules. To do the synthesis, the user would specify the resonances of the vocal tract only (`F1` and `B1` through `F5` and `B5`), as well as the amount of nasalization with a parameter such as `AN` which would be the area of the opening to the nasal cavity. From the vocal tract resonances, the shape of the vocal tract can be reconstructed. Then using the simulation method described in chapter 2, a new set of poles and zeros can be obtained. The simulations could be more complex, and could include the dual-tube model with sinuses used by Dang et al. [3] described in section 1.3.1.

A more general approach to synthesis which used the same idea, is to specify acoustic features of the speech, such as locations of formants, and pitch, and then to use an articulatory model matched to these features to do the actual synthesis. In this way, the features which are most important perceptually are maintained and controlled, while the details are taken care of by the articulatory model.

This idea of using higher-level (HL) parameters has been suggested by Stevens and Bickley [10], but the transformations from the HL parameters to the Klatt parameters which are used are all rule based. A number of the HL parameters used are supposed to be the cross-sectional areas of different constrictions and ports. It would therefore be more natural to do the transformations via an articulatory model since there is a very straightforward correspondence between the HL parameters and the physical parameters in the articulatory model.

Appendix A

Matlab Code

A.1 Simulation Functions

A.1.1 File lossynaf2pz.m

This function takes the area functions of the pharynx (afv), the mouth (afm) and the nasal passage (afn) as its input, and returns the complex locations of the poles (p) and the zeros (z). It includes losses at the lips, nostrils and glottis, as well as a series loss in the line.

```
function [p,z]=lossynaf2pz(afv,afm,afn,l0)
% [p,z]=lossznaf2pz(afv,afm,afn)
% takes the area functions of the nasal cavity, afn,
% the mouth cavity, afm, and the other part, afv, as
% vectors containing the area sampled at l0 cm intervals
% in cm^2. and returns the poles, p, and zeros, z, as
% complex values in the zplane. It impliments some of
% the losses used by Liljencrants


sf = sqrt(1000);        %square root of the frequency at which
                        %we aproximate the series losses

lv = length(afv);       %lengths of the different tubes
lm = length(afm);
ln = length(afn);
```

```
vsf = 2;              %shape factors of the different tubes
msf = 3;
nsf = 5;

sp = 35000;            %speed of sound in cm/s

%note: (lm minus ln) must be even

%radiation at the nostrils from JL 3-2

tema = .0779 + (.2373 * sqrt(afn(ln)));
temb = -.8430 + (.3062* sqrt(afn(ln)));
en = mypol(-1,[2 -2*temb]);

a = mypol(-.5,[tema+1 -(tema+temb)]);
b = mypol(-1.5,[tema-1 temb-tema]);



for j = (ln-1):-1:1

     aold = a;
     bold = b;

     if afn(j+1) == 0

          a = mypol(.5,1);
          b = mypol(-.5,1);
          en = mypol(0,1);

     else

          dee = 3.626e-5 * sf * nsf * 10 / sqrt(mean(afn([ ...
               j j+1])));

          a = mypolplus(mypolmul((afn(j+1)+afn(j))*(1+dee) ...
               /(2*afn(j+1)),.5,aold) , mypolmul(((-afn(j+ ...
               1)*(1+dee))+(afn(j)*(1-dee)))/(2*afn(j+1)), ...
               .5,bold));
          b = mypolplus(mypolmul(((-afn(j+1)*(1-dee))+(afn ...
               (j)*(1+dee)))/(2*afn(j+1)),-.5,aold) ,      ...
               mypolmul((afn(j+1)+afn(j))*(1-dee)/(2*afn(j ...
               +1)),-.5,bold));
```

```
    end

end

%radiation at the lips

tema = .0779 + (.2373 * sqrt(afm(lm)));
temb = -.8430 + (.3062* sqrt(afm(lm)));
em = mypol(-1,[2 -2*temb]);

c = mypol(-.5,[tema+1 -(tema+temb)]);
d = mypol(-1.5,[tema-1 temb-tema]);



for j = (lm-1):-1:1

    cold = c;
    dold = d;

    if afm(j+1) == 0

        c = mypol(.5,1);
        d = mypol(-.5,1);
        em = mypol(0,1);

    else

        dee = 3.626e-5 * sf * msf * 10 / sqrt(mean(afm([ ...
            j j+1])));

        c = mypolplus(mypolmul((afm(j+1)+afm(j))*(1+dee) ...
            /(2*afm(j+1)),.5,cold) , mypolmul(((-afm(j+ ...
            1)*(1+dee))+(afm(j)*(1-dee)))/(2*afm(j+1)), ...
            .5,dold));
        d = mypolplus(mypolmul(((-afm(j+1)*(1-dee))+(afm ...
            (j)*(1+dee)))/(2*afm(j+1)),-.5,cold) ,       ...
            mypolmul((afm(j+1)+afm(j))*(1-dee)/(2*afm(j ...
            +1)),-.5,dold));

    end

end
```

```
if afm(1)==0 | afn(1)==0

     if afm(1) ==0

          e = en;
          fold = a;
          gold = b;
          af = afn(1);

     elseif afn(1) ==0

          e = em;
          fold = c;
          gold = d;
          af = afm(1);

     end

     f = mypolplus(mypolmul((af+afv(lv))/(2*af),.5,fold) , ...
          mypolmul((-af+afv(lv))/(2*af),.5,gold));
     g = mypolplus(mypolmul((-af+afv(lv))/(2*af),-.5,fold) ...
          , mypolmul((af+afv(lv))/(2*af),-.5,gold));

else

     if any(afm==0)

          e = mypolmul(2*afn(1),0,mypolplus(c,d),en);

     elseif any(afn==0)

          e = mypolmul(2*afm(1),0,mypolplus(a,b),em);

     else

          e = mypolplus(mypolmul(2*afm(1),0,mypolplus(a,b) ...
               ,em),mypolmul(2*afn(1),0,mypolplus(c,d),en));

     end

     f = mypolplus(mypolmul(afn(1)+afm(1)+afv(lv),.5,a,c), ...
          mypolmul(afn(1)-afm(1)+afv(lv),.5,a,d),mypolmul( ...
          -afn(1)+afm(1)+afv(lv),.5,b,c),mypolmul(-afn(1)- ...
          afm(1)+afv(lv),.5,b,d));
```

```
    g = mypolplus(mypolmul(-afn(1)-afm(1)+afv(lv),-.5,a,c ...
        ),mypolmul(-afn(1)+afm(1)+afv(lv),-.5,a,d),        ...
        mypolmul(afn(1)-afm(1)+afv(lv),-.5,b,c),mypolmul ...
        (afn(1)+afm(1)+afv(lv),-.5,b,d));


end


for j = (lv-1):-1:1

    fold = f;
    gold = g;

    dee = 3.626e-5 * sf * vsf * 10 / sqrt(max(afv([j j+1])));

    f = mypolplus(mypolmul((afv(j+1)+afv(j))*(1+dee)/(2*  ...
        afv(j+1)),.5,fold) , mypolmul(((-afv(j+1)*(1+dee ...
        ))+(afv(j)*(1-dee)))/(2*afv(j+1)),.5,gold));
    g = mypolplus(mypolmul(((-afv(j+1)*(1-dee))+(afv(j)*( ...
        1+dee)))/(2*afv(j+1)),-.5,fold) , mypolmul((afv( ...
        j+1)+afv(j))*(1-dee)/(2*afv(j+1)),-.5,gold));


end


if any(afv==0) | (any(afn==0) & any(afm==0))
    h = mypol(0,1);
else

%now for the glotis loss from JL 2-10 and rab 87

    rg = (1-(.496/afv(1)))/(1+(.496/afv(1)));
    h =mypolmul(2/(1+rg),0,mypolplus(f,mypolmul(-rg,0,g)));
end

p = roots(h.pol);
z = roots(e.pol);
```

A.1.2 File pz2formband.m

This function takes the pole and zero complex locations and returns frequencies and bandwidths for a specific range.

```
function [f,bf,af,baf] = pz2formband(p,z,l0)
% [f,bf,af,baf] = pz2formband(p,z)
% takes a list of poles, p, and a list of zeros, z, as
% complex values, and the length of one tube , l0.
% it returns the formant and anti-formant
% frequencies and bandwidths
% f,   formant frequencies
% bf,  formant bandwidths
% af,  anti-formant frequencies
% baf, anti-formant bandwidths
%
% c assumed at 35000 cm/s and a delay of z^-1 is 2*l0 cm

f = log(p(find((angle(p)>0) & (angle(p)<pi) & abs(p)>.5))) ...
     *35000/(4*l0*pi);

af = log(z(find((angle(z)>0) & (angle(z)<pi) & abs(z)>.5)) ...
     ) *35000/(4*l0*pi);

bf = -2 *real(f);

f = imag(f);

baf = -2 *real(af);

af = imag(af);
```

A.2 Helper Functions

These files are used to create and manipulate the special kind of polynomial data-type used by the simulation functions.

A.2.1 File `mypol.m`

```
function p=mypol(del,pol)
% p=mypol(del,pol) makes a structure p which is made up of
% the two parts
% pol, polynomial co-efficients in decending order
% del, power of z for the last co-efficient in pol
%
% See also MYPOLPLUS, MYPOLMUL

p=struct('del',del,'pol',pol);
```

A.2.2 File `mypolmul.m`

```
function p = mypolmul(sc, pow, varargin)
% p = mypolmul(sc, pow, p1, p2, p3, ..., pn)
% multiply a scalar, sc, and a power of z, pow, by a seris
% of mypols, p1 through pn

p = mypol(0,1);

for j = 1:length(varargin)

     p = mypol(p.del+varargin{j}.del,conv(p.pol,varargin{j ...
          }.pol));

end

p = mypol(p.del + pow, sc * p.pol);
```

A.2.3 File mypolplus.m

```
function p = mypolplus(varargin)
% p = mypolplus(p1, p2, p3, ..., pn)
% adds any number of mypols together

ps =  [varargin{:}];

del = min([ps.del]);

for j = 1:length(varargin)

     pols(j,(1:length(ps(j).pol))+ps(j).del-del) = fliplr( ...
          ps(j).pol);

end

pol = fliplr(sum(pols,1));

p = mypol(del,pol);
```

Appendix B

Experiment Response Form

Words will be played in A-B-A triads. For each triad, all three words are the same, but there are two different versions of the word. The first and last versions are always exactly the same, and each is called A. The middle version is slightly different, and is called B. For each A-B-A triad, you are asked to choose which version sounds more natural, A or B. A version is more natural if it sounds more like it was produced by a person rather than a computer. Pay close attention to the nasal consonants (ie. the 'm's and 'n's). Indicate which you prefer by marking A or B in the appropriate box.

You will listen to nine sets, each with four triads. Each set will consist of the same four words, which are, in order; knit, Tom, mitt and tawn. (Tawn sounds a lot like Tom.) Even though the same four words are used, the triads may be different for different sets. The first set is for practice, and your answers for this set will not count.

WORD	Prac.	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8
knit									
Tom									
mitt									
tawn									

Table B.1: Mark A or B in each box.

Bibliography

[1] B. S. Atal and Suzanne L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50(2 (Part 2)):637–655, August 1971.

[2] T. Baer, J.C. Gore, L.C. Gracco, and P.W. Nye. Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *Journal of the Acoustical Society of America*, 90(2):799–828, August 1991.

[3] Jianwu Dang, Kiyoshi Honda, and Hisayoshi Suzuki. Morphological and acoustical analysis of the nasal and paranasal cavities. *Journal of the Acoustical Society of America*, 96(4):2088–2100, October 1994.

[4] Osamu Fujimura. Analysis of nasal consonants. *Journal of the Acoustical Society of America*, 34(12):1865–1875, December 1962.

[5] Dennis H. Klatt and Laura C. Klatt. Analysis, sinthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2):820–857, February 1990.

[6] Johan Liljencrants. *Speech Synthesis with a Reflection-Type Line Analog.* PhD dissertation, Royal Institute of Technology, Stockholm, Sweden, 1985. Speech Communication and Music Acoustics.

[7] Gordon E. Peterson and Harold L. Barney. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2):175–184, March 1952.

[8] Lawrence R. Rabiner and Ronald W. Schafer. *Digital Processing of Speech Signals*, section 3.3, pages 82–98. Prentice-Hall Signal Processing Series. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978.

[9] Kenneth Stevens. *Acoustic Phonetics.* MIT Press, 1999.

[10] Kenneth N. Stevens and Corine A. Bickley. Constraints among parameters simplify control of klatt formant synthesizer. *Journal of Phonetics*, 19(1):161–174, January 1991.