Production of Realistic Vowel Sounds Using a Neural Network Model of Speech Production and Acquisition

Dave Johnson

July 8, 1995

Abstract

The vocal tract model used by DIVA, a model of speech acquisition and production, is examined and found to be inadequate for the production of realistic vowel sounds. Various vocal tract models are considered and the model proposed by Shinji Maeda is chosen for implementation in a new simulation of DIVA. Formant frequencies and amplitudes are computed and used to drive a parallel formant synthesizer. Realistic vowel sounds are generated in real time during the DIVA babbling phase.

Introduction

A realistic model of speech production and acquisition is faced with the daunting task of explaining a large and growing body of speech production data. For example, Lindblom, Lubker, and Gay (1979) found that Swedish talkers compensated for the effects of a bite block during the production of four vowels, even on the first glottal pulse, and suggest, therefore, that normal speech is compensatory. MacNeilage and DeClerk (1969) demonstrated that both anticipatory and carry-over coarticulation occur in CVC syllables. Kent and Murray (1982) studied the comfort-state vocalizations of 3, 6, and 9 month old human infants and found that a wide range of speech-like sounds were produced. They suggest that these sounds and the corresponding articulatory gestures may play a role in the infant's development of speech. Fowler (1980) surveys a number of approaches to the modelling of speech production and argues that coordinative structures that encompass the respiratory, laryngeal, and supralaryngeal systems exist in speech production. How can all of this data be reconciled?

The DIVA model (Guenther, 1993, 1994) makes substantial contributions to our understanding of the speech process. DIVA (Directions Into Velocities of Articulators) is a model of speech production and learning which has been used to explain a wide variety of speech data including anticipatory and carry-over coarticulation, compensation for perturbations using motor equivalence, speech acquisition via babbling, speaking rate effects such as vowel reduction, and the emergence of coordinative structures.

The DIVA model learns two coordinate transformations or "mappings". One such transformation maps an auditory field of neural activity which encodes the current phoneme into a vector in orosensory space. The directions in orosensory space represent acoustically important degrees of freedom. These degrees of freedom are more abstract than the articulators themselves, but less abstract than the phonemes. The DIVA model posits that targets of the articulators are represented as convex hulls in orosensory space. (See Rabiner, 1966, for a similar proposal.)

A second coordinate mapping from orosensory space to articulator velocities is also learned. By mapping to articulator velocities rather than positions, learning is simplified.

These mappings self-organize during a sequence of action perception cycles known as babbling. In the DIVA model, an endogenous random generator of articulator velocities drives the articulators to produce output. Feedback then drives the learning process as shown in Figure 1.

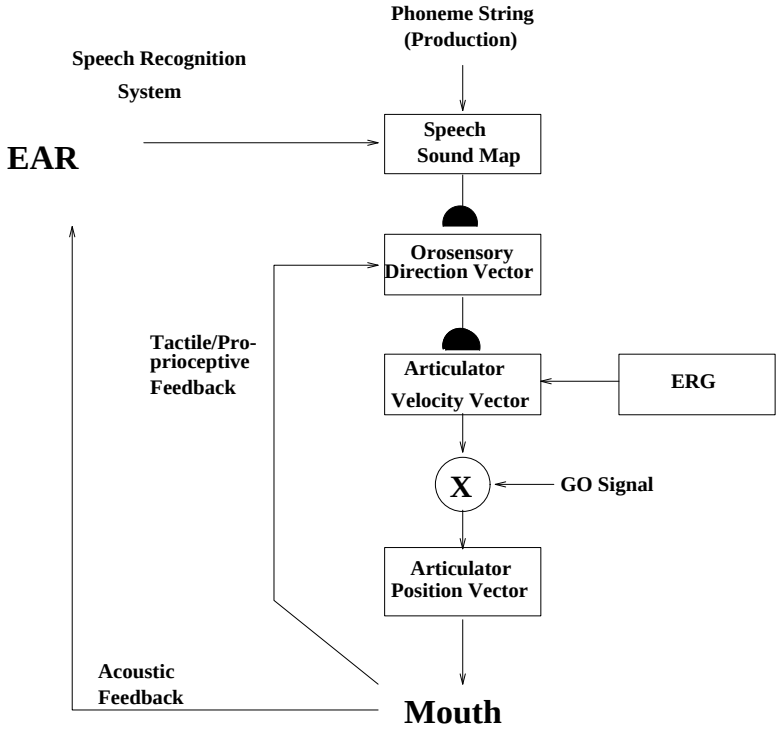

Figure 1: Block diagram of DIVA

The simulations of the DIVA model (Guenther, 1993, 1994) do not produce acoustic output. A graphics display and numerical output from the simulation serve to illustrate the behavior of the model. In this paper, we report work done to enhance the DIVA model and the corresponding computer simulation in order to produce acoustic output of acceptable quality in real time during the babbling phase. This enhancement adds articulatory speech synthesis capabilities to the DIVA model. Other articulatory synthesizers exist (Rubin, Baer, & Mermelstein, 1981; Maeda, 1990), but none with the capabilities inherent in the DIVA model, namely the ability to learn phonemes during a babbling phase and the ability to explain the speech data mentioned above.

Is it really necessary to add speech synthesis capability to the DIVA simulation? Many questions about speech acquisition and production can be answered by a model that produces no acoustic output. However, there are many questions that cannot be answered without such a capability. These questions derive from the simple observation (often completely lost in the volumes of literature on speech production and acquisition) that the human speech system produces high quality natural speech. Therefore, any correct and accurate model of the human speech system must be capable (at least, in principle) of

producing high quality and natural speech.

It must be admitted that artificial speech synthesizers are currently not capable of producing speech at the level of quality of that of a human talker. Some of the reasons for this technological limitation will emerge from the discussions which follow. However, the goal of human speech production modelling must be the production of natural speech.

Given such a capability, it would be possible to answer the following questions: What (if any) are the coordinative structures of speech? What gestures are important during the babbling of infants for the acquisition of natural speech? What acoustic features are important for the self-organization of the speech system?

Moreover, real time generation of speech by the simulation is important for two reasons. First, any model of speech acquisition must have access to the acoustic speech output. To the extent that the learning takes place in real time, the generation of acoustic speech must also take place in real time. Second, placing some of the focus of the research on speed performance forces a close examination of the mechanisms and trade-offs in the speech production process.

Out of this research, it is hoped that we will 1) provide a more rigorous test of the strengths and weaknesses of the DIVA model, 2) allow exploration of the relationship between neural models and speech quality, 3) allow faculty and students to obtain expertise in speech synthesizer techniques, 4) provide a platform for further research in the neural basis of speech production, 5) provide a platform for research on speech perception, and 6) explore realistic models of the vocal tract for the production of vowels.

The DIVA simulation reported here produces digitized speech output and is implemented to run on a SPARC-10 workstation running SunOS 4.1. These workstations support the UNIX /dev/audio device which is adequate for speech synthesis. A UNIX manual page describes this device in detail.

The following sections will discuss technical issues in achieving the speech synthesis goals for DIVA, vocal tract models, methods for calculating the formant frequencies, and trade-offs in various configurations of formant synthesis.

Technical Issues

Several technical issues had to be addressed. Among them were:

1 Speech quality - Since the DIVA model supports all English phonemes, it might be expected that any enhancement to add audible speech must support all of these phonemes. However, we have chosen to synthesize non-nasalized vowels only during this phase of the research. It is likely that nasalized vowels and nasal consonants will be synthesized in the next phase. Then fricatives would be treated, and the last phase would treat stop consonants.

2 Computational efficiency - Ideally, audible speech of sufficient quality can be computed in real time (simultaneous with the neural computations and the graphical display of the vocal tract articulators). However, very high quality speech synthesis requires time consuming parameter selection, an impossibility in real time. Therefore, sampling of the vocal tract configurations is performed. This should be adequate for steady state vowel production.

   Another performance issue concerns the method by which the speech samples are computed. We have chosen to generate the digitized speech using a formant synthesizer (from Sensimetrics Corp). The formants are computed using software from Shinji Maeda which accounts for acoustic losses in the vocal tract. The rationale for this decision is discussed at length below.

3 Ease of implementation and maintainability of the software - Several approaches to the speech synthesis problem will be outlined below. The viability of this project demands a simple and elegant implementation 1) to ensure its usefulness and 2) to facilitate its development in the near term. To this end, existing software will be used where possible.

4 Ease of usability - The realization of the above-mentioned benefits assumes that user access to the simulation features is maximized. The user interface of the original simulation is written in Motif. This standard is also followed here.

5 Realism of the articulatory model - The quality of the digitized speech and the relevance of this research to the problem of speech production and acquisition in humans hinges on the realism of the vocal tract model. A model of vowels proposed by Shinji Maeda (Maeda, 1990) was selected to replace the DIVA articulator model, at least for the generation of the digitized speech.

DIVA's model of the vocal tract

The purpose of the original simulation of DIVA was to demonstrate the behavior of the neural network model and show that it exhibited compensatory behavior and carry-over coarticulation effects (Guenther, 1993). Later simulations demonstrated the model's ability to explain speaking rate effects and anticipatory coarticulation (Guenther, 1994). Neither of these efforts required that a realistic model of the vocal tract be used. An effort was made to use articulatory degrees of freedom that were known to affect acoustic properties of speech (such as those suggested by Ladefoged, 1964), but acoustic output was not produced. For the initial purpose of the simulation, it was sufficient that the number of degrees of freedom be sufficiently large to permit compensatory and coarticulation effects to be observed.

The original DIVA model of the vocal tract has the following articulatory degrees of freedom:

upper lip (vertical and horizontal)

lower lip (vertical and horizontal)

jaw (vertical)

tongue tip (vertical and horizontal)

tongue body (vertical and horizontal)

velum (open and closed)

voicing (on and off)

noise (on and off)

Because of the artificial shapes of these articulators and their scaling in pixel coordinates, the model's ability ever to produce realistic vowel sounds was doubted. It was decided that a realistic model of the vocal tract should be chosen and substituted for the DIVA vocal tract model.

Before elaborating on the process of choosing such a vocal tract model, it is necessary to discuss the process of speech synthesis in general and the techniques used in this research project.

Speech synthesizers

Speech synthesis by computer has a long history. (For a concise discussion of this history, see, for example, Klatt, 1987.) The main methods of speech synthesis today are 1) formant synthesis, 2) articulatory synthesis, and 3) synthesis by rule (commonly used in text-to-speech systems). There is much overlap among these methods. Synthesis by rule will not be considered further in this paper.

In order to properly motivate the selection of the vocal tract model, the rudiments of speech synthesis will be presented. We will begin with formant synthesis which is the simplest technique for generating realistic vowel sounds, and then we will consider the techniques of articulatory synthesis.

Formant synthesis

A formant synthesizer is a computer program or other mechanism that produces digitized speech output given only the necessary formants. Formants arise in the context of source filter theory. In the synthesis of vowel sounds, the speech system can be modelled using source filter theory in which the cavities between the glottis and the lips act as an acoustic filter which filters the source signal generated by the vocal folds at the glottis. In this conception, a source of sound energy produced at the glottis (the space between the vocal cords or folds) is a volume velocity waveform (measured in cubic centimeters per second of air flow). The output at the lips is also a volume velocity. The sound wave which is heard at some distance from the face is customarily represented as a pressure waveform and is related to the volume velocity at the lips by the radiative characteristic. Figure 2 depicts such a system.

A formant is a peak or maximum of the transfer function relating the glottal volume velocity waveform to the lip volume velocity waveform. In the case of adult male speech, usually five formants exist below 5kHz. Each formant is characterized by its frequency, its bandwidth (the width in Hz of the peak at 3dB below its maximum value), and its amplitude (its value in dB at the maximum).

In the frequency domain, the pressure wave at a sufficient distance from the lips is given by

$$P(f) = S(f)T(f)R(f)$$

where $S(f)$ is the spectral representation of the glottal source signal, $T(f)$ is the transfer

function relating lip volume velocity $U_L(f)$ to glottal volume velocity $S(f)$, and $R(f)$ is the radiative impedance relating the pressure $P(f)$ to the lip volume velocity $U_L(f)$. This expression is a reasonable approximation for the steady state production of speech sounds such as nasalized and non-nasalized vowels, and nasal consonants. With appropriate modification for source of excitation, the expression may also describe fricatives (Maeda, 1982; Stevens, 1971). However, plosives or stops cannot accurately be represented by steady state methods (Fant, 1970; Rabiner & Schafer, 1978).

Before proceeding to derive the transfer function, a comment about source filter theory is in order. Source filter theory assumes a linear model of the vocal tract. This assumption is a reasonable approximation of the speech system only for low frequencies and energies, and for certain configurations of the vocal tract. There are many instances of nonlinearity that may arise. One is the interaction between the subglottal and supraglottal cavities which increase with the size of the glottal aperture (Klatt & Klatt, 1990). For this reason and others, Maeda (1982) has suggested that an aerodynamic treatment of the entire respiratory system and the supraglottal region is called for. Such a treatment goes beyond the scope of this paper.

In order to understand how a formant synthesizer works, it is necessary to mathematically analyze the transfer function of the vocal tract for vowels. Assume that the transfer function can be represented by poles in the complex plane. Thus it has the form of the quotient of a constant and a polynomial which can be factored(Klatt, 1980):

$$T(f) \cong T(z) = \prod_{i=1}^{N} \frac{A_i}{1 - B_i z^{-1} - C_i z^{-2}}$$

where

$$C_i = -e^{-2\pi\delta_i T}$$

$$B_i = 2e^{-\pi\delta_i T} cos(2\pi F_i T)$$

$$A_i = 1 - C_i - B_i$$

and

$T \equiv$ period (1/sampling rate)

$\delta_i \equiv$ bandwidth of the i^{th} formant

$F_i \equiv$ frequency of the i^{th} formant

Each factor can be thought of as a digital resonator. The z-transform representation is used because of its direct connection to the discreet-time speech output. The z variable can be written

$$z = e^{j\omega T}$$

where

$$\omega = 2\pi f$$

The discreet-time output of the digital resonator is given by

$$y_i(nT) = A_i x_i(nT) + B_i y_i(nT - T) + C_i y_i(nT - 2T)$$

The significance of the z-transform of the transfer function is that it has a (sampled) impulse response identical to a corresponding analog resonator at sampling times nT.

Cascade versus Parallel

If the transfer function can be modelled using only poles, then Flanagan (1957) showed that the formant frequencies and bandwidths are sufficient to determine the shape and amplitudes of the frequency response curve when the digital resonators are cascaded. This works fine for vowels and sonorants, but not for fricatives. This is true because accurate modelling of the vocal tract for fricatives requires introduction of zeros into the transfer function. These zeros correspond to frequencies at which an infinite acoustic impedance is seen looking backwards from the fricative noise source. Zeros insert notches and affect the amplitudes of formants. The notches are not perceptually important, but the formant amplitudes *are* perceptually important. In the case zeros are necessary, the frequencies and bandwidths of the formants are not sufficient to deduce the amplitudes of the formants and, hence, the overall form of the transfer function. However, a parallel scheme using only poles works as long as the formant amplitudes are known.

Simple rules are known (Klatt, 1980) for calculating the formant amplitudes given the frequencies and bandwidths under the assumption that zeros are not present in the transfer function. The resulting transfer function is approximately the same as the corresponding cascade response. We are free to use either a parallel or a cascade configuration of a formant synthesizer. Since our goal is to model non-nasalized vowels only, we may use either. However, on the expectation that we will eventually want to model vocal tract configurations having zeros in the transfer function, we will use the parallel configuration instead as it gives us more flexibility.

The Source

A simplified discreet-time expression for the glottal source is given by

$$\begin{aligned} g(n) &= \tfrac{1}{2}[1 - cos(\pi n/N_1)] && 0 \leq n \leq N_1 \\ &= cos(\pi(n - N_1)/2N_2) && N_1 \leq n \leq N_1 + N_2 \\ &= 0 && \text{otherwise} \end{aligned}$$

This expression is only an approximation which does not capture the complexity of the glottal source in natural speech. For example, the fundamental frequency (which is implicit in the choices of N_1 and N_2) undergo minute random fluctuations. Another departure from simplicity is the presence of tracheal poles and zeros (a simplified tool for dealing with the nonlinear interactions between the tracheal and pharyngeal cavities). Another is the aspiration noise which accompanies "breathy" voice, common to both males and females and serves to widen the first formant and steepen the role-off of the glottal signal in the frequency domain. Another is the presence of diplophonia, a form a aperiodicity in the glottal source which occurs when pulses are attenuated or entirely omitted. Klatt and Klatt (1990) discuss all of these effects and others, and serve to justify our choice of the Sensimetrics formant synthesizer (based on the Klatt88 synthesizer) which may allow us to model these effects in future simulations.

The Radiation Characteristic

In order to calculate the pressure wave $p(nT)$ far from the lips, it is necessary to know the acoustic impedance at the lips. However, an approximation of $p(nT)$ can be obtained by taking the first difference of the lip volume velocity. In the discreet-time representation,

$$p(nT) = u(nT) - u(nT - T).$$

However, this expression is not accurate enough for high quality speech and must be modified by modelled or measured radiative impedances and is implemented using one or more digital resonators such as those found in the formant synthesizer discussed above.

Another approach is to fold the radiative impedance into the vocal tract transfer function. This is the approach taken by the Maeda synthesizer. Therefore, the formant frequencies and amplitudes computed by the Maeda synthesizer already take the radiation characteristic into account. As it turns out, the Sensimetrics formant synthesizer does not provide a way to account directly for the radiation.

Acoustic loss

Maeda (1982) describes an algorithm to calculate the frequency response of the vocal tract when loss is taken into account. In a lossless model, peak amplitudes are infinite and bandwidths are zero. However, when losses are considered, the formant peaks are broadened and have finite amplitude. Also, the amplitudes are not all of the same magnitude. It is desirable that the DIVA vocal tract model include losses to increase the realism of the vowel sounds.

Most scientists are more familiar with electrical transmission line theory than with the theory of acoustics. Therefore, many presentations of the theory of the acoustics of speech use the electrical analog as a starting point (Fant, 1970, "pages 91-92"). In this formulation, the Telegrapher's equations are first presented. The Telegrapher's equations are the electrical transmission line analog of the Horn equations which describe the propagation of sound waves in space. In one dimensional propagation, the Telegrapher's equations are given by

$$\frac{dV}{dx} = -IZ$$
$$\frac{dI}{dx} = -VY$$

where V is the voltage and I is the current as a function of position x along the transmission line. Z, the complex impedance per unit length, and Y, the complex admittance per unit length, are found to be frequency dependent and are given by

$$Z = R + j\omega L$$

$$Y = G + j\omega C$$

where R is the series resistance per unit length, L is the series inductance per unit length, G is the parallel conductance per unit length, and C is the parallel capacitance per unit length, and $j = \sqrt{-1}$. The telegrapher's equations can be combined to yield the second order wave equation which can be solved to obtain the rightward and leftward traveling waves corresponding either to the voltage wave or to the current wave.

The corresponding Horn equations hold where P (pressure) corresponds to V and U (volume velocity) corresponds to I. In addition, there are defined complex impedance and complex admittance which are functions of the geometry of the enclosure in which the sound propagates and the acoustical properties of the media. These terms are important because frequency-dependent losses account for variations in the shape of the transfer function which are perceptually important.

The complex acoustic impedance may be written (Rabiner & Schafer, 1978):

$$Z(x,\omega) = \frac{S(x)}{[A_0(x)]^2}\sqrt{\omega\rho\mu/2} + j\omega\frac{\rho}{A_0(x)}$$

and the acoustic admittance may be written

$$Y(x,\omega) = \frac{S(x)(\eta-1)}{\rho c^2}\sqrt{\frac{\lambda\omega}{2c_p\rho}} + j\omega\frac{A_0(x)}{\rho c^2}$$

where

x = position along the direction of propagation

$S(x)$ = circumference of the tube

$\omega = 2\pi f$ where f = frequency

c = velocity of sound

$A_0(x)$ = nominal area

c_p = specific heat (at constant pressure)

c_v = specific heat (at constant volume)

μ = coefficient of friction

ρ = density of air

λ = coefficient of heat conduction

$\eta = c_p/c_v$

Computation of the formant frequencies and amplitudes

A number of methods exist for calculating the formants corresponding to a vocal tract configuration as specified in terms of its area function. An area function is the cross sectional area of the vocal tract as a function of distance from the glottis. Since in some models the position of the glottis is not fixed with respect to the skull (see for example Maeda, 1990), the position dimension of the area function is sometimes measured with respect to the upper teeth.

In the simplest computational models (Rabiner & Schafer, 1978), the vocal tract is approximated with a cascade of lossless uniform cylindrical tubes which provides a best

fit to the area function. Each tube is treated as a two-port network. Various formalisms have been applied to the problem of characterizing each two-port such as using reflection and transmission coefficients, scattering parameters, or impedances. In linear two-port network theory, it is always possible to convert between any of these formalisms and, therefore, the choice of formalism is partly a matter of taste and convenience (Fant, 1970). The most widely used formalism for lossless networks is based on reflection coefficients.

In general, the transfer function is constructed from these two-ports to obtain the quotient of polynomials which may then be solved for the maxima or may be evaluated at the frequencies of interest (Rabiner & Schafer, 1978). Using a peak-detection algorithm, it is possible to find the peaks, bandwidths, and amplitudes of the formants. These may then be used by a formant synthesizer to synthesize digitized speech as described in the preceding section.

Articulatory synthesis

Articulatory synthesis is speech synthesis in which the input to the synthesizer is a set of articulator positions instead of the standard formants or, possibly, the area functions. This method presumes that a model of the vocal tract exists in which articulators define the shape of the vocal tract. Articulatory synthesizers may assume that true articulators define the vocal tract or that shape factors derived from a factor analysis define the shape of the vocal tract. Shape factor analysis is defined and discussed in detail below. For the present discussion, articulatory synthesis may be based on either physical articulators or on shape factors.

In articulatory synthesis, the shape of the vocal tract is specified in terms of a set of articulator values or parameters (or shape factor values). This set of parameters is transformed into an area function. One of the techniques discussed above is then used to compute the digitized speech samples.

Articulatory synthesizers are important in the study of speech production for several reasons (Rubin et al., 1981). They allow detailed study of the role of articulators and their effects on the acoustic properties of speech and they allow the study of models of coarticulation and compensation.

Specific Approaches

The following basic approach to synthesizing the digitized speech was used. The articulator positions predicted by the DIVA model were transformed into Maeda shape factor (articulatory) parameters. These parameters are then converted into an area function. This area function is then transformed into digitized speech samples.

This last step may be done in a number of ways. These really boil down to the following three:

1 Direct synthesis - This technique uses a digital filter in which approximately 20 stages perform a shift operation. This is perhaps the simplest to implement.

2 Formant synthesis with lossless model - This technique uses a minimum number of lossless tubes and ignores source and radiative impedances and ignores formant amplitudes and bandwidths.

3 Formant synthesis with lossy model - This technique requires that realistic values of the losses be used and involves a more complicated computational model. Since we have the software from Maeda which performs this computation, this would seem to be the best technique.

The decision was made to use formant synthesis based on a lossy model. There were several reasons for this decision. First, the software to do this was already available and was part of the Maeda software package. Initial testing proved that the software worked as expected, though the computation is slow, even on a SPARC-10. Second, it was judged that the computational simplifications resulting from removing the loss-dependent calculations would not significantly speed up the computation, but would, most likely, degrade the quality of the resulting speech. Third, some form of formant synthesis would be necessary in order to produce speech in real time, thus ruling out direct synthesis altogether. Direct synthesis is too slow compared to formant synthesis because at least twenty stages would be necessary in a direct synthesis digital filter while five (or fewer) efficient digital resonators (with two shift stages per resonator) would be sufficient for formant synthesis. In addition, we already had a very good formant synthesizer (the Sensimetrics Klatt88-based formant synthesizer) and we did not have an implementation of a direct synthesis digital filter.

Because of the computational expense involved in calculating the formant frequencies, it is necessary to sample the articulator positions at a rate sufficient to track the relevant

shapes of the vocal tract. On a SPARC-10, it was determined that vowel production could be simulated adequately by sampling the articulator positions every 10-20 configurations. The formant frequencies are computed at these points and interpolated between. Wright and Elliott (1990) showed that interpolation of articulatory parameters produced unacceptable speech (due to lack of bandwidth control), but suggested that vowel transitions were characterized by linear regions in the formant versus time plots. It was found that linear interpolation of formant frequencies between the articulator sampling points produced acceptable speech.

Direct synthesis

Formant synthesis is a practical method for generating digitized speech only if the formants are available. Since it is likely that formants will play a role in the simulation of acoustic feedback in the DIVA model and because we have a very good formant synthesizer, it is worth the trouble to compute the formants. However, for vowel production, direct synthesis is probably more efficient and is worthy of some consideration.

Direct synthesis of digitized speech is based on the assumption that the vocal tract can be approximated by sufficiently many cascaded linear two-port networks. Suppose that the vocal tract is sectioned into equal-length tubes. It can be shown (Fant, 1970) that if the tube length is sufficiently small (a length which is determined by the sampling frequency $1/T$), then the resulting individual two-ports will be *linear*. Then the techniques of linear two-port networks may be used to compute the output of the network given any discreet-time input. In particular, a digital filter may be constructed in which the propagating volume velocity undergoes only a delay when in a two-port and only a complex reflection at the interface between two adjacent two-ports. The delay can be approximated by multiplication by the z-transform variable z^{-1}. The reflection coefficient seen from tube i looking toward tube $i+1$ is given by:

$$\Gamma_{i,i+1} = \frac{Z_{i+1} - Z_i}{Z_{i+1} + Z_i}$$

where Z_i is the characteristic impedance of tube i and Z_{i+1} is the characteristic impedance of tube $i+1$.

In the lossless case, the characteristic impedance of a tube depends only on its cross sectional area (or circumference) as is evident from the limiting case of the lossy complex impedance and admittance given above.

Therefore, a simple direct synthesizer can be constructed using only delays and multiplications assuming that the area function is known. Using graph-theoretic techniques, it is possible to optimize the digital filter in order to minimize the number of multiplications(Rabiner & Schafer, 1978). It can be shown that the optimal design uses only one multiplication per two-port network.

To restate, a direct synthesizer was not used to simulate the DIVA model because a good formant synthesizer was available and it is believed that the formant frequencies will be required during learning. However, direct synthesis remains an option should performance become a critical issue and it is determined that formants are not required for learning.

Vocal Tract Models

DIVA was not designed to accurately represent the human vocal tract, but used, instead, an idealized representation of the articulators. As you can see, they bear some relation to the human vocal apparatus, but are not suitable for the accurate acoustic modelling necessary for high quality speech synthesis. Even though the resulting shapes of the vocal tract produced by DIVA might be qualitatively correct, the resulting acoustic parameters using these shapes will be wrong. Sounds would be produced, but those sounds wouldn't sound like human speech. If we want to learn to produce phonemes from the speech output during babbling, then the babbled speech must accurately reflect the shape of the vocal tract or we will have no chance to learn the necessary articulator velocities.

How may we solve this problem? A number of possibilities present themselves. Before we can proceed, we need to look in the literature at previous attempts to parametrize the vocal tract configuration in a way that accurately models humans.

Models of the Vocal Tract

Stevens and House (1955) used parabolic approximation of area function to show that reasonably intelligible vowel sounds could be produced.

Ladefoged (1964) proposed a simple physiological model using 10 parameters. The parameters included 1) air pressure at the trachea, 2) position of the vocal cords, 3) tension of the vocal cords, 4) degree of velo-pharyngeal stricture, 5,6) coordinates of the center of the body of the tongue, 7,8) coordinates of the tip of the tongue, 9,10) protrusion and

opening of the lips. At the time he presented this model, he had not yet built a synthesizer which used it.

Lindblom and Sundberg (1971) were the first to treat the jaw as a separate speech articulator. They found that by so doing, they were able to explain why the jaw was partly open during speech. Under this condition, the tongue undergoes minimum deformation. Their data was obtained from Swedish talkers.

Liljenkrants (1971), cited in Harshman et al. (1977), proposed a Fourier decomposition of tongue shape. Though recognized as probably non-biological, the model does allow simple specification of a wide variety of tongue shapes.

Shinji Maeda (1972) suggested that shape factors derived by the method of Analysis by Synthesis could be used to capture more accurately the shapes of the vocal tract. Factor analysis is discuss in more detail in a later section.

Mermelstein (1973) proposed a simple geometric model of the vocal tract which he used for both vowels and consonants. It is one of the most widely cited quantitative models of the vocal tract. The model assumes that the tongue body conforms to a circular arc of constant radius. The jaw is allowed to undergo only angular displacements. The tongue blade is a straight line. However, Mermelstein was able to fit a wide variety of acoustic data with this model. Rubin, Baer, and Mermelstein (1981) used this model in an articulatory speech synthesizer with good results.

Most recent models of vocal tract shape are derived using factor analysis.

Harshman, Ladefoged, and Goldstein (1977) performed Factor analysis of tongue shapes.

Jackson (1988) studied Cross-linguistic factors of shape to propose that such factors might be related to coordinative structures of speech articulation. He suggests that some languages may favor one gesture over another and thinks that factor analysis may be able to decide this question.

Types of Vocal Tract Models

From this survey of the vocal tract modelling literature, we can see that several different approaches have been used. Three broad categories of approaches can be identified as follows:

Area function

Articulators

Shape factors

An area function gives the cross sectional area of the vocal tract as a function of distance from the glottis (or from the teeth which are fixed). From the area function, acoustic parameters such as formant frequencies and amplitudes (or bandwidths) may be calculated. Area functions are still commonly used to model the vocal tract and the other types of models can always be transformed into their equivalent area function for the purposes of acoustic processing.

Early models of the vocal tract consisted of an area function. It was found that the most important parameters were 1) place of constriction, 2) size of constriction, and 3) aspect ratio of the lips. Using ever more accurate area functions, higher quality of speech can be obtained, limited by such problems as 1) multimodal acoustic propagation, 2) errors due to lumped element simulation, 3) nonlinear interactions between the glottis and the pharynx, and 4) thermal and mechanical losses. However, study of the area function does not teach us very much about the underlying speech production processes, especially from a motor control point of view. At some point, we must begin to model the human speech articulators and use these to derive the acoustic output.

Articulator-based models were introduced to more accurately model the biomechanics of speech and to permit the study of coarticulation and coordinative structures which are difficult to study in terms of the area function. In these models, the important questions are What is the minimum set of articulators and how do they affect each other? Usually, the shapes of the articulators are fixed or are constrained to vary in a simple way.

The third type of vocal tract model is the shape factor model. Shape factors are determined from the statistical analysis of many tracings of the vocal tract. The next section discusses the method of factor analysis.

Factor Analysis

In order to motivate the discussion of factor analysis, assume that many vocal tract shapes are encoded in vector form. In this analysis, a shape vector has a large number of components, each corresponding to a point in some suitably chosen coordinate system. The underlying assumption of factor analysis is that (under certain conditions) there exist a set of *shape factors* such that each vocal tract shape vector can be written as a linear combi-

nation of the shape factors. Factor analysis is a statistical method for deriving these shape factors from a sufficiently large collection of vocal tract shapes.

Factor analysis may be applied to problems outside the scope of vocal tract shape and assumes only that the factors are statistically independent and significant. Factor analysis is actually a broad class of techniques. Principal component analysis, perhaps the best known of these, is just one example of a factor analysis technique.

Maeda used principle component analysis in combination with extraction and subtraction methods to arrive at his collection of shape factors (Maeda, 1990). The hierarchy of techniques in factor analysis can be summed up by the following definitions:

> Factor analysis: The statistical analysis of correlated data into orthogonal components or "factors".
>
> Arbitrary Factor Analysis: Generalized version of the extraction- subtraction procedure which may also use principal component analysis.
>
> Principal component analysis: A general statistical procedure which finds an optimal set of orthogonal basis vectors which minimizes mean square error.

Mathematics of principal component analysis

Two-way principal component analysis decomposes a data set into factors and weights or *loadings*. In this technique, it is assumed that all of the data share the same factors. This may not be a reasonable assumption if the data is taken from several sources (such as different human talkers). In this case, three-way factor analysis is available and computes the degree to which a factor is used by a subject as well as the degree to which a factor is used in the formation of a particular vowel (Harshman et al., 1977).

We will be concerned only with two-way factor analysis. According to Zahorian and Rothenberg (1981), the following analysis applies: Given a covariance matrix $[C]$ with

$$C_{ij} = \frac{1}{K}[\sum_{k=1}^{K}(x_{ki} - \overline{x_i})(x_{kj} - \overline{x_j})],$$

for $i, j = 1, 2, \ldots, n$ where

K = # data frames

$x_{ki} = i^{th}$ data sample of the k^{th} frame

$\overline{x_i}$ = average over k of i^{th} data sample

n = # data elements in each frame

The basis vectors are the m eigenvectors ($m \leq n$) of $[C_{ij}]$ corresponding to the m largest eigenvalues.

Let $\{x_j\}$ represent a shape vector and $\{y_i\}$ represent the shape factor loadings. Then

$$y_i = \sum_{j=1}^{n} A_{ij} x_j$$

for $i = 1, 2, \ldots, m$, and

$$x_j^{'} = \sum_{i=1}^{m} A_{ij} y_i + M_j$$

for $j = 1, 2, \ldots, n$, where the prime on the x_j indicate that the values are reconstructed estimates of the original data. In these expressions, A_{ij} = j^{th} component of the i^{th} eigenvector of $[C]$ with the eigenvectors ranked in order of decreasing eigenvalues, and

$$M_j = \sum_{i=m+1}^{n} A_{ij} \sum_{p=1}^{n} A_{ip} \overline{x_p}$$

for $j = 1, 2, \ldots, n$

Maeda Shape Factors

Maeda defines 7 shape factors or articulators. They are:

jaw height

tongue-body position

tongue-body shape

tongue-tip position

lip height (aperture)

lip protrusion

larynx height

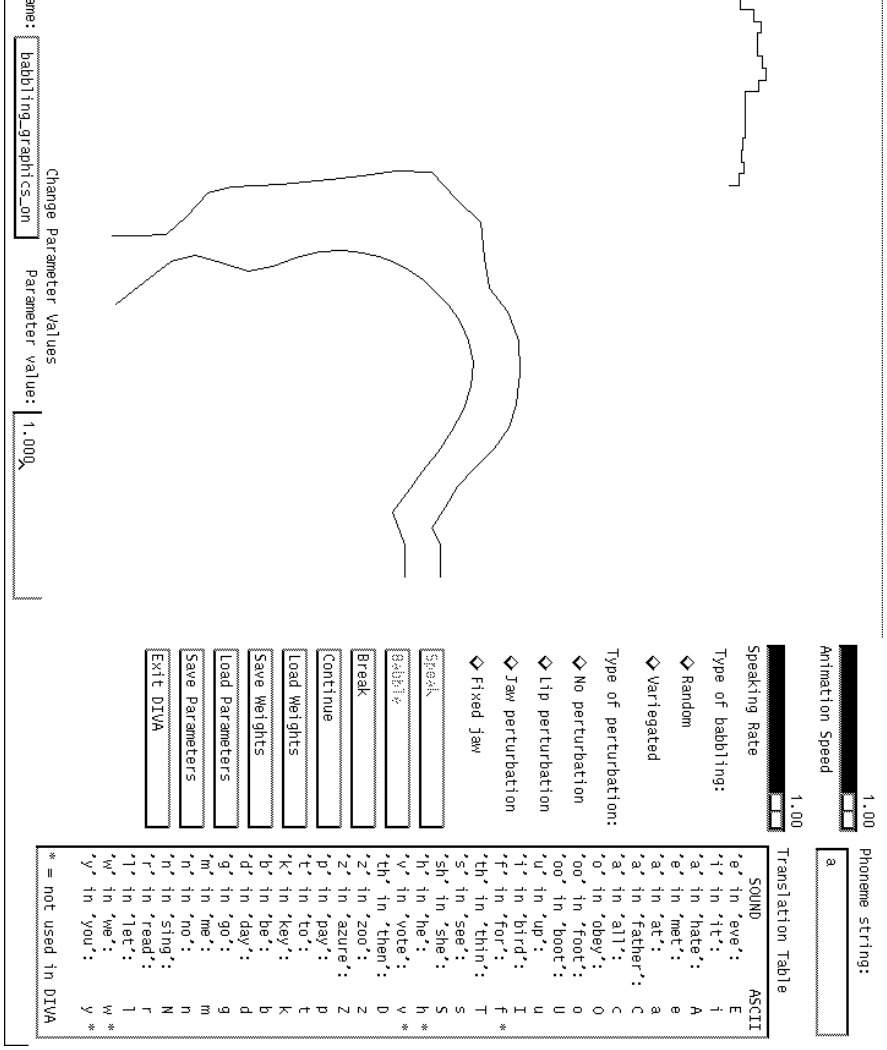

Figure 2: Uniform vocal tract: all Maeda parameters are zero

He usually uses the term "parameter" to refer to them, but, strictly speaking, they are shape factors. Because of the manner in which these factors are derived, they can take on the value of approximately -3 to $+3$ where zero represents the mean shape or position corresponding to that factor.

Figures 2 - 6 show the shape of the human vocal tract for different values of the Maeda parameters. The parameter values of -3, 0, and $+3$ are plotted.

It must be pointed out that shape factor analysis works well when there are shape factors in the data. It is too early to decide whether shape factor analysis will work for vocal tract shapes corresponding to consonants. In particular, if the articulators deform in a nonlinear manner when moved and brought into contact with each other, it is unlikely that a linear decomposition of the data into shape factors will occur. However, shape factor analysis is useful for deciding whether gestures are common across individuals

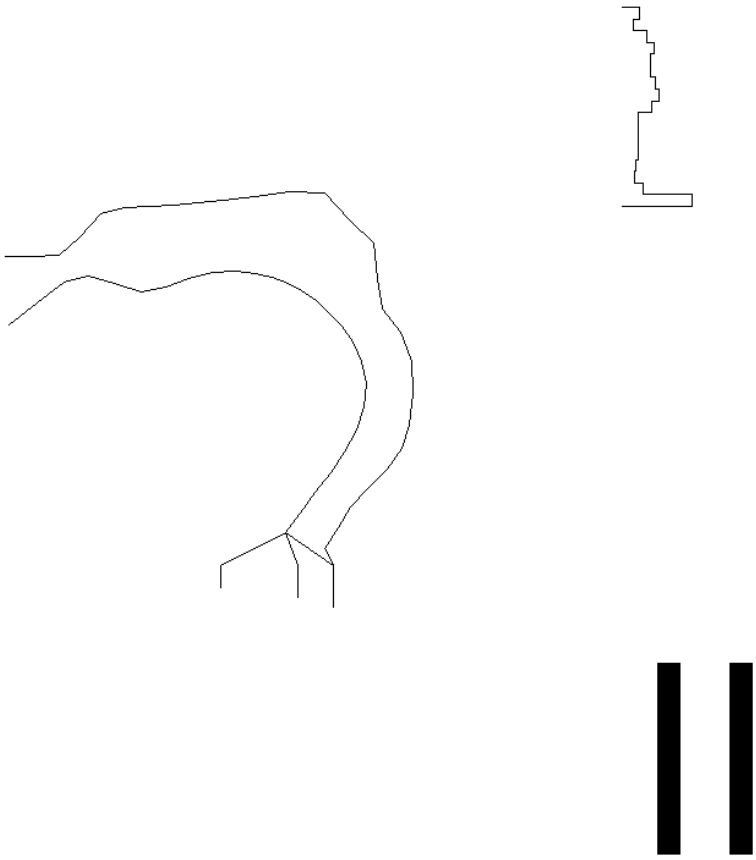

Figure 3: Variations of the Maeda lip aperture parameter

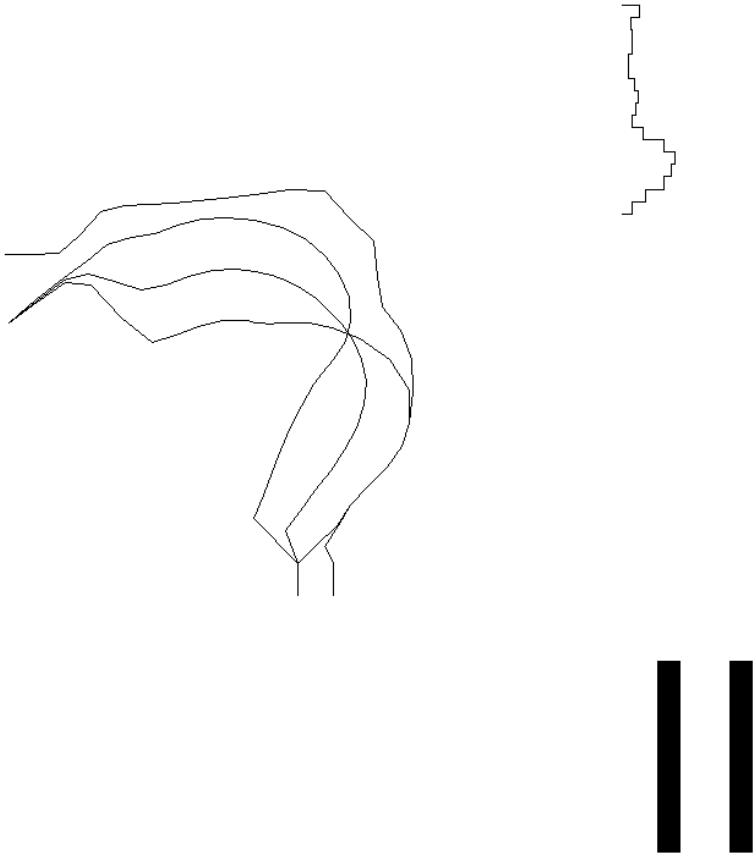

Figure 4: Variations of the Maeda tongue body position parameter

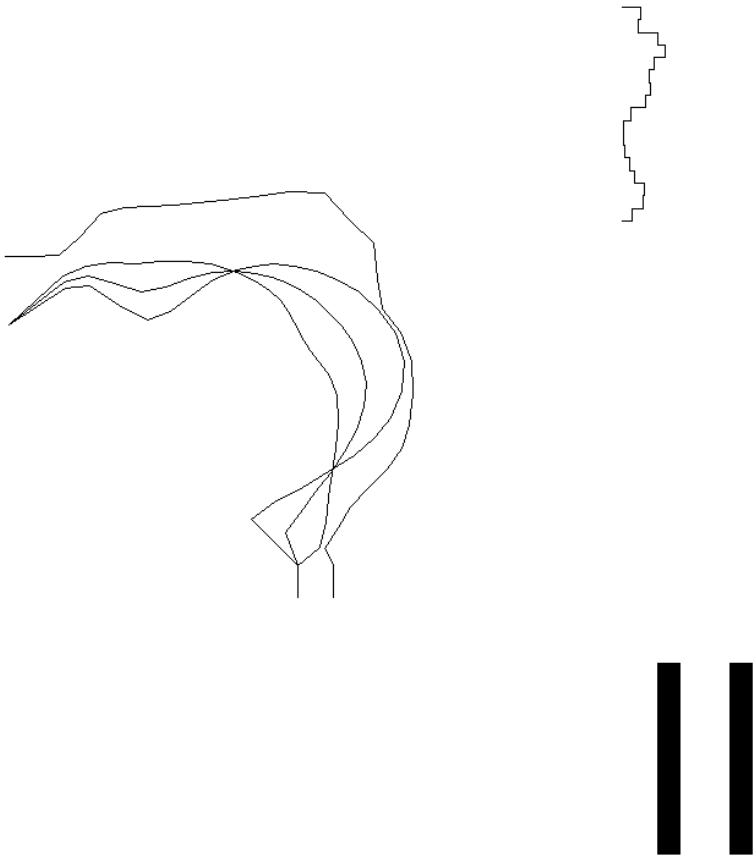

Figure 5: Variations of the Maeda tongue shape parameter

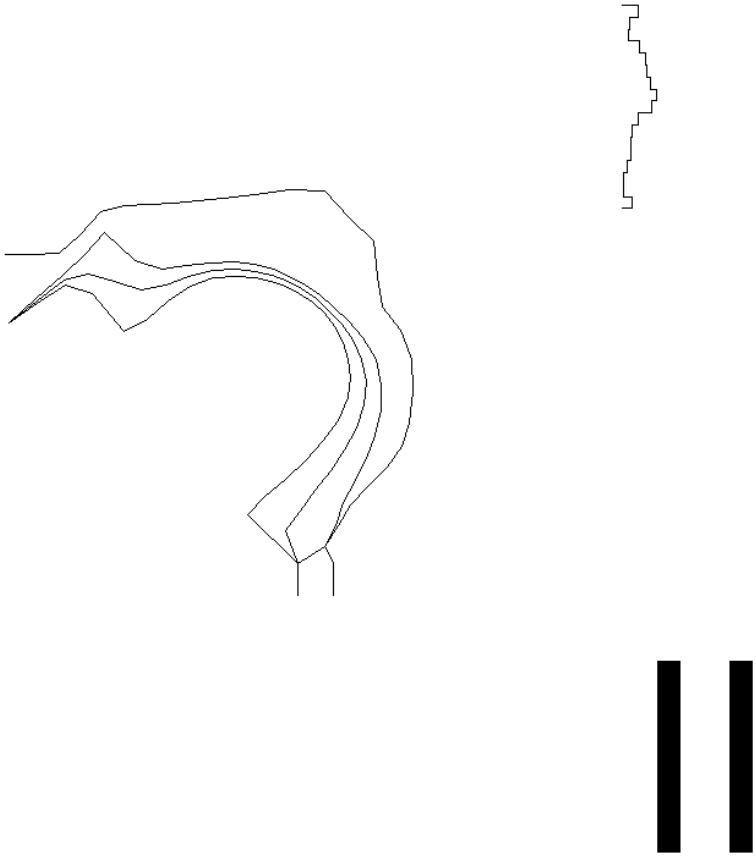

Figure 6: Variations of the Maeda tongue tip position parameter

or different languages (Jackson, 1988). Moreover, it may still turn out that shape factor analysis will work for consonants and shouldn't be abandoned too soon.

Putting it all together

Simple assumptions were made about the relation between the Maeda shape factor parameters and the DIVA articulator positions. In particular, a linear transformation relating the two was implemented, allowing a realistic vocal tract to be derived corresponding to every DIVA articulator configuration.

Figures 7 and 8 clarify the computational stages in the generation of digitized speech during the DIVA simulation. The flowchart presented in Figure 7 represents the steps performed by the Maeda synthesizer. Similarly, the flowchart presented in Figure 8 represents the steps performed by the Sensimetrics formant synthesizer.

The necessary software modifications were made to both synthesizers to allow them to be integrated by a simple function call with the DIVA simulation software. A considerable amount of effort was invested in this activity and to ensure that the synthesizers still functioned as expected.

Conclusions

All of these modifications were implemented on a SPARC-10 and the output speech samples are generated and written to /dev/audio in real time at 8000 samples per second. The result is quite acceptable sound which represents the babbling phase of DIVA.

Although the performance of the resulting simulation is slower than hoped, there is reason to believe that this problem can be solved. However, the speech quality is very good, judging subjectively, compared to other speech synthesizers, and under the circumstances. Of course, only non-nasalized vowels are produced, but these vowels sound recognizable and seem to track the DIVA and Maeda vocal tracts that are displayed concurrently.

Suggestions for further study include the following:

1 Clean up loose ends in current implementation (of which there are many).

2 Investigate other Articulatory and Shape factor models of the vocal tract.

3 Increase speed of the transfer function calculation (if possible).

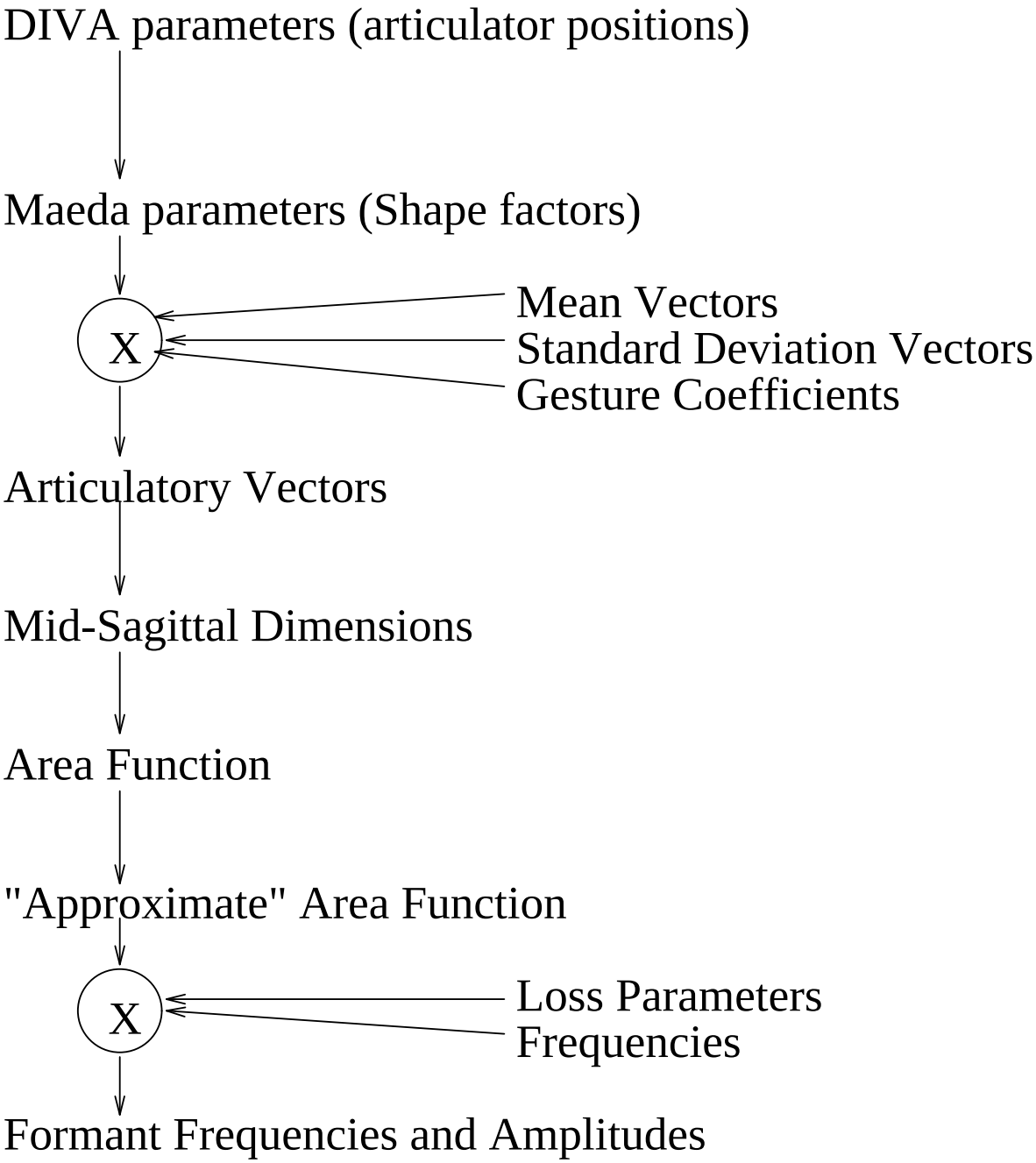

Figure 7: Flow chart of formant frequency calculation

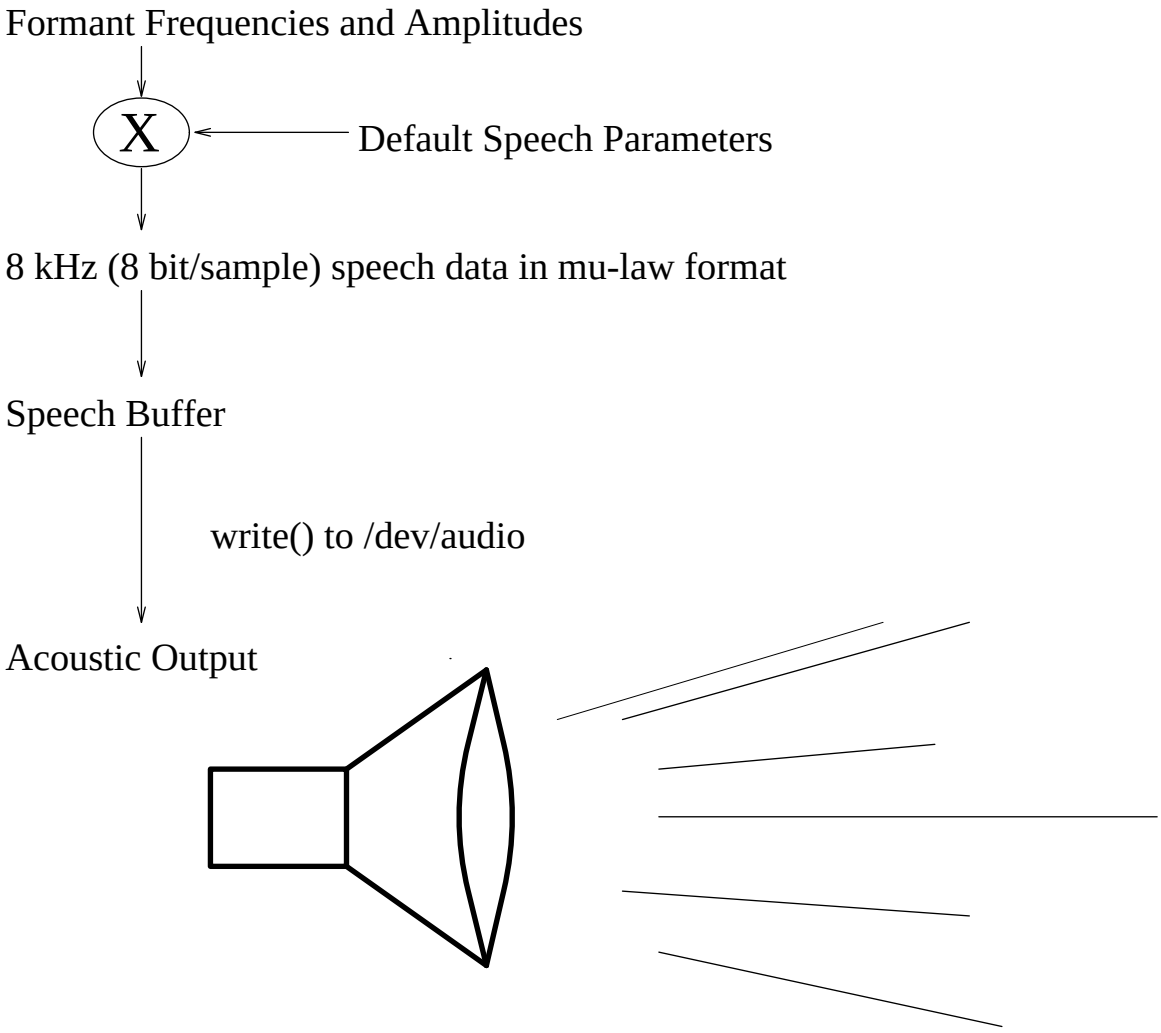

Figure 8: Flow chart of digitized speech production

4 Read further on consonant production.

5 Upgrade Maeda model to support consonant production.

6 Explore AudioFile. According to a recent posting to comp.speech, AudioFile (AF) was developed at DEC and permits /dev/audio on one workstation to be shared by other workstations (and indeed, other versions of UNIX and other types of audio devices) on a network. The source code is distributed publicly and reportedly has already been used to implement speech synthesis across a network. Perhaps this capability can be used to allow speech synthesis on one machine and speech perception (e.g., for babbling) on another machine.

7 Characterize the quality of the speech. Objective measures of the quality of speech produced during babbling by the new DIVA simulation need to be investigated. Ideally, acoustic features of infant babbling should be matched to those predicted by DIVA. One shortcoming that will become apparent is that we have used an adult male model of the vocal tract derived by Maeda. This is probably not appropriate for infant babbling.

Acknowledgements

I would like to thank Frank Guenther for his assistance on this project. I would like also to thank Joseph Perkell for making available a DOS version of the Maeda vowel synthesizer and Sensimetrics Corporation for its UNIX version of the Klatt-based formant synthesizer. In addition, I wish to thank Seth Cameron and Jonathan Chey for valuable discussions regarding the Maeda vocal tract model and speech synthesizer fundamentals.

Reference

Fant, G. (1970). *Acoustic Theory of Speech Production, 2nd Ed*. Mouton, The Hague, The Netherlands.

Flanagan, J. L. (1957). Note on the design of terminal-analog speech synthesizers. *J. Acoust. Soc. Am.*, *29*, 306–310.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, *8*, 113–133.

Guenther, F. H. (1993). A neural network model of speech acquisition and motor equivalent speech production. Tech. rep. CAS/CNS-93-054, Boston University, Center for Adaptive Systems, Boston.

Guenther, F. H. (1994). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Tech. rep. CAS/CNS-94-012, Boston University, Center for Adaptive Systems, Boston.

Harshman, R., Ladefoged, P., & Goldstein, L. (1977). Factor analysis of tongue shapes. *J. Acoust. Soc. Am.*, *62*, 693–707.

Jackson, M. T. T. (1988). Analysis of tongue positions: Language-specific and cross-linguistic models. *J. Acoust. Soc. Am.*, *84*, 124–143.

Kent, R. D., & Murray, A. D. (1982). Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *J. Acoust. Soc. Am.*, *72*, 353–365.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.*, *67*, 971–995.

Klatt, D. H. (1987). Review of text-to-speech conversion for english. *J. Acoust. Soc. Am.*, *82*, 737–793.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.*, *87*, 820–857.

Ladefoged, P. (1964). Physiological parameters of speech. *J. Acoust. Soc. Am.*, *38*, 1037.

Lindblom, B., Lubker, J., & Gay, T. (1979). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics*, *7*, 147–161.

Lindblom, B. E. F., & Sundberg, J. E. F. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. *J. Acoust. Soc. Am.*, *50*, 1166–1179.

MacNeilage, P. F., & DeClerk, J. L. (1969). On the motor control of coarticulation in cvc monosyllables. *J. Acoust. Soc. Am.*, *45*, 1217–1233.

Maeda, S. (1972). Conversion of midsagittal dimensions to vocal tract area function. *J. Acoust. Soc. Am.*, *51*, 89–90.

Maeda, S. (1982). A digital simulation method of the vocal-tract system. *Speech Communication*, *1*, 199–229.

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W.J., H., & Marchal, A. (Eds.), *Speech Production and Speech Modelling*, pp. 131–149. Kluwer Academic Publishers, The Netherlands.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *J. Acoust. Soc. Am.*, *53*, 1070–1082.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, New Jersey.

Rabiner, L. R. (1966). Speech synthesis by state simulation. *J. Acoust. Soc. Am.*, *40*, 1272.

Rubin, P., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *J. Acoust. Soc. Am.*, *70*, 321–328.

Stevens, K. N. (1971). Airflow and turbulence noise for fricative and stop consonants: Static considerations. *J. Acoust. Soc. Am.*, *50*, 1180–1191.

Stevens, K. N., & House, A. S. (1955). Development of a quantitative description of vowel articulation. *J. Acoust. Soc. Am.*, *27*, 484–493.

Wright, R. D., & Elliott, S. J. (1990). Parameter interpolation in speech synthesis. *J. Acoust. Soc. Am.*, *87*, 383–391.

Zahorian, S. A., & Rothenberg, M. (1981). Principal-components analysis for low-redundancy encoding of speech spectra. *J. Acoust. Soc. Am.*, *69*, 832–845.