Articulatory Speech Synthesis and Speech Production Modeling

Jun Huang

Thesis Proposal
Ph.D. Preliminary Examination
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

February 9 2001
3269 Beckman Institute

Committee:
Prof. Stephen E. Levinson (Chairperson)
Prof. Mark A. Hasegawa-Johnson
Prof. Thomas S. Huang
Prof. William D. O'Brien
Dr. Don W. Davis

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 History of Speech Synthesis

Speech is the most important form of human communication. It consists of a sequence of sounds which are generated by the vocal apparatus and used as a vital tool for interaction among human beings. People have long been fascinated by the possibility of enabling inanimate machines with the power of speech, or at least something resembling it.

Early attempts to construct talking machines can be traced back to the eighteenth century. Interest in speech synthesis began during the Greek and Roman civilizations when clever deception gave voice to inanimate statues and gods. In 1779, Kratzenstein constructed a set of acoustic resonators which, when activated by a vibrating reed, produced a reasonable imitation of the steady-state vowels [79]. Twelve years later, Von Kempelen built a more elaborate machine that could generate connected utterances of speech. The Von Kempelen speaking machine used a bellows to supply air through a reed acting as the vocal cords. Voiced sounds, consonants and nasals could be generated by exciting a resonator whose shape was determined by the operator's one hand and

controlled by the fingers of the other hand. An elaborate version of this machine was later built and demonstrated by Sir Charles Wheatstone in 1879. The influence of tVon Kempelen's machine also motivated Alexander Graham Bell in the late eighteenth century to model many of the articulators using a cast from human skull and modeling the interior parts of the vocal tract in guttapercha. Interesting enough, Bell and his brother were able to imitate many of the vowels and nasals, and their creative work was later acknowledged as a U.S. patent, dated 1876. The construction of mechanical machines to simulate voice sounds was not an easy task due to the numerous variations in pitch, intonation, etc., and the complex movements of the vocal apparatus. As an article in *Scientific American* published in 1871 [33] puts it, "Machines which, with more or less success, imitate human speech, are the most difficult, to construct, so many are the agencies engaged in uttering even a single word - lungs, larynx, tongue, palate, teeth, lips - so many are the inflections and variations of tone and articulation, that the mechanician finds his ingenuity taxed to the uttermost to imitate them."

The first electrical synthesizer which attempted to generate connected utterances was the Voder [20] developed by Dudley *et al.* in 1939. The device was driven manually and consisted of ten parallel bandpass filters that span the entire frequency range of speech and a set of resonances and anti-resonances were applied to the excitation source. The resonance control box was excited by either a noise source or a buzz oscillator. The Voder achieved enormous success in producing reasonably intelligible speech quality and was later demonstrated at the world's fairs of 1939 in New York and 1940 in San Francisco. The evolution of electronic technology and the greater interest in human-machine communications sprang new and exciting dimensions in speech synthesis. The first was the development of the Vocoder machine by Dudley [21] based on the principles of speech analysis and synthesis. With the recent growth in high speed digital computers and the progress encountered in the development of integrated circuitry, analysis-synthesis tech-

niques began to play an important role in speech communication to reduce the transmission bandwidth used for speech signals. The application of analysis-synthesis techniques extends to many areas where memory storage is limited. Synthetic speech in devices for human-machine communication has proved successful in producing resonable human utterances and reducing message storage space significantly.

As an alternative to analysis-synthesis methods, speech can be synthesized according to certain linguistic and acoustic rules which convert a string of discrete symbols into speech. This is known as *Text-To-Speech* (TTS), or synthesis by rule. Prosodical information such as duration, pitch and stress is necessary to describe the manner and strategy by which the utterance is to be spoken. The history of speech synthesis by rule extends back to the late 1950s when acoustic signals started to be recognized as individual speech sounds [65]. However, the first automatic TTS system was generated in 1961 by Kelley and Gerstman [56] which enabled direct transformation of linguistic units, supplemented with pitch and timing information, into segments of speech. Their rules for formant synthesis were later elaborated for British English by Holmes *et al.* in 1964 [43].

1.2 Speech Production

In the previous section, a brief history of speech synthesis was outlined. Any kind of speech synthesis method is based on a certain type of speech production model. They differ in their complexity and in the methods employed for achieving natural quality synthetic speech. Speech can be described as a sequence of sounds which are generated when a flow of air is disrupted or perturbed by the vocal apparatus, namely, lips, jaw, tongue, velum and larynx. These are also known as the vocal articulators. A simplified diagram of the human vocal system for speech production is shown in Fig. 1.1. The vocal

tract is considered as a non-uniform acoustic tube which extends from the vocal cords to the lips. The vocal cords are essentially shelflike ligament-covered muscles embedded in the edges of the larynx. The region between the larynx and the velum is referred to as the *pharyngeal tract* and that between the velum and the lips as the *oral tract.* The cross-sectional area of the vocal tract varies from 0 to 20 cm^2 depending on the position of the articulators, which also determine the overall length of the vocal tract. The vocal tract length varies between 15 cm and 20 cm for different speakers and may change according to the sound produced. For an average adult male, the vocal tract is considered to be about 17 cm long in its rest position. For the production of nasal sounds, the nasal tract is acoustically coupled to the vocal tract by lowering the velum and forming a constriction at some point along the oral cavity. Under these conditions, the air flows through the pharynx into the nasal cavity and sound is radiated at the nostrils. For an average adult male, the overall length of the nasal tract is about 12 cm long and its cross-sectional area changes mainly in the region close to the velum.

Speech sounds can be roughly classified into three categories, namely *voiced, fricatives* and *plosives* depending on their methods of excitation. *Voiced* sounds are produced if the air passing through the larynx causes vibration of the vocal folds. Consequently, quasi-periodic pulses of air are produced whose period is determined by the mass and tension of the cords, as well as the subglottal pressure. Examples of voiced sounds are /*i*/ (as in *beet*) , /*u*/ (as in *boot*) and /*a*/ (as in *hot*). *Fricative* sounds are generated by forcing air through a constriction formed at some point in the vocal tract which results in a turbulent flow of air in that region. Constrictions can be labio-dental, dental, alveolar, palatal or glottal. For example, the sound /θ/ (as in *thick*) is produced by a constriction in the dental region. The third class of sounds is referred to as *plosive.* These are produced when a build-up of pressure behind a complete closure in the vocal tract is suddenly released by the articulators. Closures can be labial, alveolar, palatal or velar.

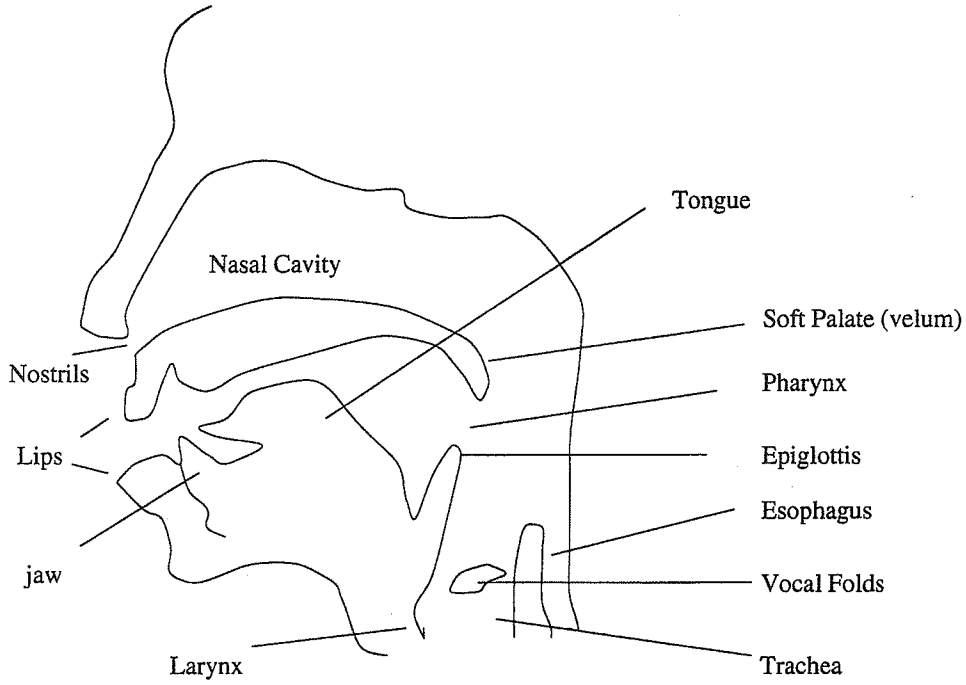


Figure 1.1 A simplified diagram of the human vocal system.

A typical example is the sound /*p*/ (as in *pick*), produced when a closure in the labial region is formed. All the above speech sounds are considered as basic linguistic units and are often referred to as *phonemes*. Details of the physiology of speech production and a phonetic study of English sounds are described in great detail by Fant [27].

Currently, the most widely used speech production model assumes that the excitation source (vocal cords) and the "filter" (vocal tract) are separable; the vocal tract can be approximated as a series of abutting lossless acoustic tubes and the sound propagation inside the vocal tract can be modeled as a one-dimensional plane wave. These assumptions lead to the source-filter speech production which can be formulated mathematically as the linear prediction coding (LPC) model. Despite of its simplification of the complicated speech production mechanism, the LPC model has resulted in tremendous success in a wide range of speech processing scenarios, such as feature extraction in speech recog-

nition, speech production modeling in speech synthesis and in speech coding. The details of the source-filter model and LP analysis will be presented in the next chapter.

1.3 Contributions

During the last decades, progress in computational power, pattern recognition, signal processing, statistical modeling *et al.* have opened a new phase in the area of speech signal processing. For example, the concatenative speech synthesis technique can synthesize highly intelligible speech sounds. For some specific domain applications, it can synthesize almost natural sounding speech by carefully concatenation of word strings in a large database. However, there still remain specific setbacks that prevent current systems from achieving the goal of completely human-sounding speech. Furthermore, even if the current systems might have a lot of commercial success, they can't help us understand the basics and some unsolved, yet important problems in human speech and language.

This dissertation addresses the problem of articulatory speech synthesis based on fluid dynamic principles. The overall strategy of our investigation is to devise a refined speech production model based on the most fundamental physics of the human vocal apparatus. Unlike the conventional "source-filter" model which assumes the independence of the excitation and the acoustic filter, we treat the entire vocal apparatus as one system consisting of a fluid dynamic aspect and a mechanical part. We model the vocal tract by a three-dimensional moving geometry. We also model the sound propagation inside the vocal apparatus as a three-dimensional non-plane wave propagation inside a viscous fluid described by Navier-Stokes equation. The primary contribution of this thesis can be summarized as follows. First, we integrated the whole articulatory speech system including vocal cord modeling, vocal tract geometry estimation and Navier-Stokes equation solving to synthesize both speech phonemes and continuous speech sentences, which we believe

has never been done before. Second, we proposed a signal approximation technique in the Hilbert space framework and applied it to the estimation of dynamic articulatory movement. Third, we will simulate and analyze the velocity and vorticity fields in the speech production and investigate the interaction between the sound mode and the flow mode and their influences in speech production. If time permits, we will also construct a physical articulatory speech production model based on the magnetic resonance imaging (MRI) data of the human vocal apparatus.

1.4 Organization of the Proposal

This proposal is organized as follows. In Chapter 2, we review the relevant literature on speech synthesis and speech production modeling. The algorithms of estimation of dynamic articulatory parameters and the error analysis are developed in Chapter 3. In Chapter 4, we present different kinds of vocal cord excitation models. The theoretical development of the articulatory model construction based on MRI data is introduced in Chapter 5. In Chapter 6, the velocity and vorticity fields in speech sound production are analyzed and their relation to a refined speech production model is stated. We present the evaluation results on speech synthesis and analysis in Chapter 7. Finally, the results of this thesis work and the topics for future research are discussed in Chapter 8.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview of Speech Synthesis Techniques

Speech synthesis technology can been classified into three broad categories: concatenative synthesis, formant synthesis and articulatory synthesis.

Concatenative synthesis is currently the most popular technique in commercial and research text-to-speech (TTS) systems. This method relies on extracting model parameters from speech data (or often just storing raw waveforms) and concatenating these to create new utterances. There are several classes of speech models that allow varying degree of control over the different characteristics of the voice quality: the linear predictive coding (LPC), based on a source-filter model using a synthetic source; pitch-synchronous time-domain models; and sinusoidal models that represent the speech waveform as a sum of time-varying sinusoidal waves. The concatenative synthesis holds the promise of representing ill-understood fine details in the voice and replacing hand-optimization of rules with a data-driven approach to waveform generation. The intelligibility of the concatenative speech synthesis systems is good. However, they are not flexible in producing

characterized sounds of different speakers and the naturalness of the synthesized speech is not satisfactory.

Formant synthesis is a parametric approach which applies a set of rules for controlling the frequencies and amplitudes of the formants and the characteristics of the excitation source. Although certain isolated phonetic units can be characterized almost solely by their formant frequencies and motions, the formant locations in continuous natural speech are heavily influenced by context. Because of this fact, the rules necessary to control a formant synthesizer are rather complex. Another feature of parametric formant synthesis is that the input parameter list can grow to be fairly large. A Klatt synthesizer can have over fourty input parameters that need to be tracked over time. The shortcomings of formant speech model are the fundamental limit to the naturalness of speech produced by rule, and the large number of parameters that need to be tracked.

Articulatory synthesis attempts to produce speech by first understanding how the vocal apparatus changes shape during speech production, then understanding the acoustic problem of how those movements translate into sounds. The input parameters to an articulatory synthesizer used to create speech include the positions of the model articulators, such as the tongue body height, and these parameters are specified as trajectories through time. Physical models of aerodynamic processes and of wave propagation inside a tube are used to convert the input parameters of the synthesizer into sound. Articulatory synthesis is potentially the most efficient way to generate speech waveforms with natural sounding, customized voices. However, many important problems need to be solve before articulatory synthesizer can be used to produce better quality speech. One of the major impediments to the use of articulatory synthesis in creating natural sounding speech has been a lack of knowledge of the articulatory movement patterns. Second, a good vocal fold model need to be coupled with the fluid model inside vocal apparatus to generate the excitation signals for high quality speech synthesis. Third, there is a problem in

identifying the articulatory degrees of freedom which are most salient to the production and propagation of sound. It is necessary to know the salient components of articulation because there are simply too many degrees of freedom to hope to make a practical articulatory synthesizer without such knowledge. Fourth, It is well-known that the sound propagation inside the vocal tract is a three-dimensional non-plane wave propagation inside a viscous fluid described by the governing Navier-Stokes equation. How can we investigate the energy exchange between the convective and propagative components of the fluid flow and the effect of nonlinear fluid flow on the speech production? Finally, although articulatory synthesis is efficient in terms of number of controlled parameters, the computational cost is very high. A new computational model which can capture the nonlinear effects in speech production with resonable computational complexity is very important for the practical usage of articulatory synthesis.

2.1.1 Concatenative Synthesis

A widely used method for converting a string of phonemes into an acoustic signal is the concatenation of segments (such as diphones) of naturally spoken utterances. In concatenative synthesis, segments of speech are excised from spoken utterances and are connected to form the desired speech signal. Currently, concatenative synthesis is the most comercially successful technique for speech generation. In December 2000, SpeechWorks announced "Speechify 1.0, a product of the strategic partnership between SpeechWorks and AT&T, is the first to capitalize on 30 years of AT&T Labs research developing human sounding, synthesized speech with sophisticated language analysis capability." "With Speechify1.0, weather updates, traffic reports, search engine results, email, stock quotes and a wide range of other kinds of dynamic information from databases or websites can be accessed in real-time and easily understood by callers. The Speechify engine

retrieves information from a database, synthesizes the text and outputs it in audio format over any phone."

There are two main problems in concatenative synthesis: *unit selection* and *speech model*. *Unit selection* involves defining the inventory of units as well as selecting the appropriate unit for a given phonetic (and prosodic) context. *Speech models* in speech synthesis include linear predictive coding (LPC), pitch-synchronous time-domain models such as pitch-period-sized waveform samples (PSOLA), and sinusoidal models that represent the speech waveform as a sum of time-varying sine waves. Although each of these models has its own merit, none have yet proven to satisfy all the desired properties of a TTS waveform model.

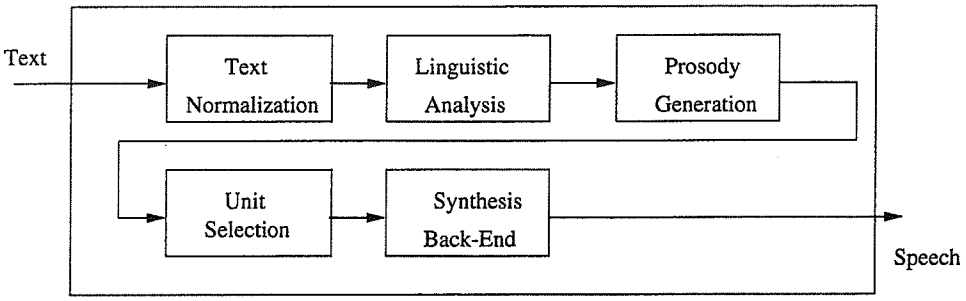


Figure 2.1 System architecture of the AT&T Next-Gen TTS.

In this section, we will discuss the AT&T NextGen speech synthesis system, which is a typical example of the state-of-art contatenative synthesis system. Fig. 2.1 shows the system diagram of AT&T Next-Gen TTS system. The text normalization, linguistic processing such as syntactic analysis, word pronunciation, prosodic prediction and prosody generation (translation between a symbolic representation to numerical values of fundamental frequency F_0, duration and amplitude) is done by a Flextalk object that borrows heavily from AT&T Bell labs' previous TTS system [8]. In the following, we will discuss in certain detail about the unit selection and synthesis back-end of the AT&T NextGen TTS system.

Diphone systhesis has been popular for a number of years, due to the high intelligibility that such system provide. They have the ability to preserve some of the coarticulation effects that are present at phoneme boundaries. However, such systems are handicapped by having a large number of distance units (in the range of 1000 - 3000 diphones, depending on language and phone-set chosen), for which it is not easy to create sufficiently large database that capture all relevant coarticulation effects. Statistics become even worse if we demand that all relevant prosodic variations are covered.

The online unit selection of AT&T's NextGen TTS was adopted from ATR's CHATR system. In CHATR system, it requires a set of speech units that can be classified into a small number of categories such that sufficient examples of each unit are available to make statistical selection viable. Hence, the original CHATR system uses phonemes as units. To avoid problems of concatenation at phoneme boundaries, a flexible join technique is employed that allows moving unit boundaries. In order to arrive at a robust paradigm, the AT&T's NextGen TTS has chosen to use half phones as the basic units of synthesis in a way that allows synthesis from units ranging in size from diphones and phones to whole words and even phrases.

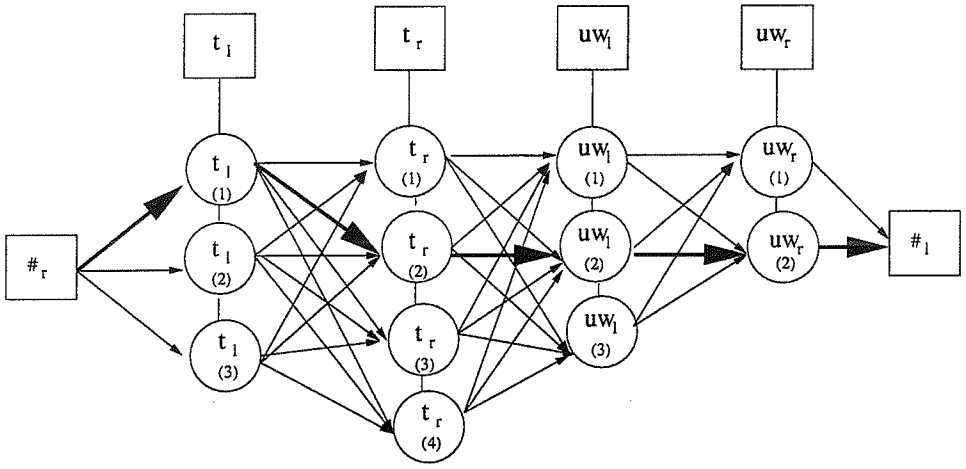


Figure 2.2 Viterbi search based on an inventory of multiple instances of each half-tone needed for synthesizing silence-/t/-/uw/-silence (the word "two")

Most of the concatenative TTS system use the statistical modeling technique from speech recognition field to dynamically search for the sequence of synthesis units that minimizes context mismatch and concatenation costs. In this way, units with size varying from a fraction of a phone to many words can be used together in a dynamic way. Fig. 2.2 shows the viterbi search based on an inventory of multiple instances of each half-tone needed for synthesizing silence-/t/-/uw/-silence (the word "two"). More details can be found in [15].

In the context of speech synthesis based on concatenation of acoustic units, speech signals may be encoded by speech models. These models are required to ensure that the concatenation of selected acoustic units results in a smoothed transition from one acoustic unit to the next. One option in AT&T's NextGen TTS is the Harmonic plus Noise model (HNM). HMM has the capability of effective waveform modifications such as controls of prosody, emotional stress, *et al.* Combining HNM to efficiently represent and modify speech signals with a unit selection algorithm may alleviate difficulty of prosodic modification problem in conventional concatenative speech synthesis system.

In HNM, the speech spectrum is divided into two bands: a low band, which is represented by harmonically related sinusoids with slowly time-varying amplitudes and frequencies, and a high band that is instantiated by a time-varying auto-regressive (AR) model that is excited by Gaussian noise. HNM analysis consists of three steps. First, fundamental frequency (F_0) and maximum voiced frequency that determine the number of harmonics used are set using a time-domain approach. Then, harmonic amplitudes and phases are estimated by minimizing a weighted time-domain least squares criterion. Finally, the AR filter for the high band is estimated by the autocorrelation approach. the analysis windows are set at a pitch-synchronous rate during voiced portions of speech and at a fixed rate during unvoiced portions. Note that the length of local pitch epochs in HNM are estimated internally. Analysis windows are two pitch epochs long. For

HNM synthesis, inter-unit phase mismatches are eliminated using the center-of-gravity approach [117]. Prosody may be altered as desired. Around unit concatenation points, the HNM parameters are smoothed in order to minimize residual discontinuities by employing a linear interpolation over a small number of frames. The actual synthesis is done by following the overlap-and-add paradigm. For each frame, the noise part is high-pass filtered according to the maximum voiced frequency found during analysis. Furthermore, the noise part is modulated by a parametric triangular envelope synchronized in time with the pitch period. Details of HNM and its application to concatinative synthesis can be found in [118]

2.1.2 Formant Synthesis

The efficient representation of the speech signal in the spectral domain led to the development of formant (or resonance) synthesizer. In formant synthesis, the acoustic characteristics of the vocal tract are modeled directly in the frequency domain by a set of resonators. It is based on the assumption of source-filter separation and attempts to model the acoustics of the vocal tract using poles (formants). Formant synthesizer may also introduce additional anti-resonances to accommodate zeros in the transfer function for the production of nasal and unvoiced sounds. They are also suitable for including effects of radiation at the lips and nostrils.

Formant synthesizers have essentially two implementation structures. The first structure is based on the orator verbis electric (OVE) synthesizer developed by Fant [26]. It includes a set of resonators connected in cascade which models the vocal tract transfer function in the frequency domain. A simplified structure of a cascaded formant synthesizer is shown in Fig. 2.3 (a). Five resonators are selected to cover the range up to 5 kHz and are adjusted by a set of control parameters representing formant frequencies

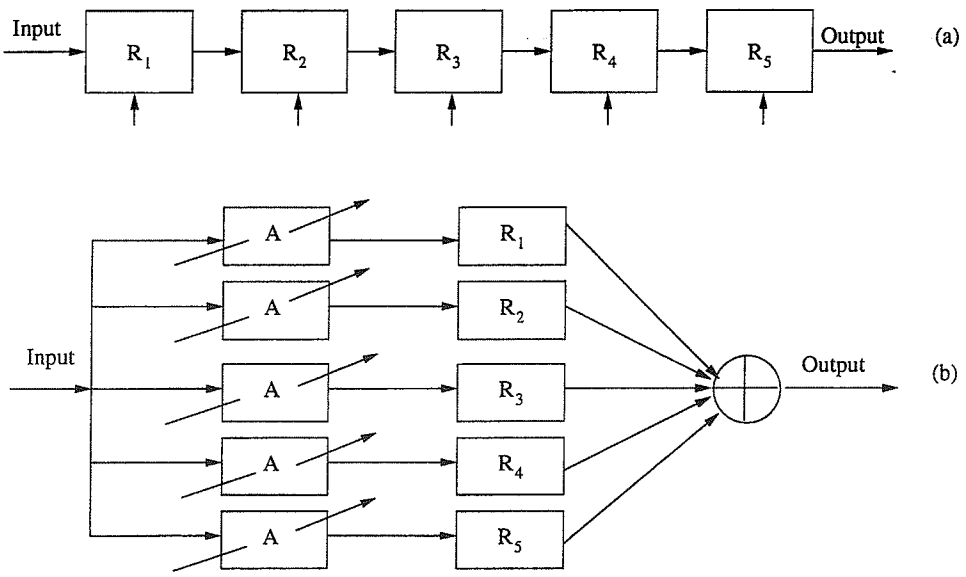


Figure 2.3 Block diagrams illustrating different structures of formant synthesizers (a) cascade; (b) parallel.

and bandwidths. The second structure of formant synthesizers is based on the notion of Lawrence's parametric artificial talking (PAT) device [63], [1]. The resonators which simulate the vocal tract transfer function are connected in parallel as shown in Fig. 2.3 (b). This structure enables formant peaks to be adjusted individually for the production of voiced and unvoiced sounds. A refined version of this model was reported by Holmes [44] which incorporated four resonators to model the acoustics of the vocal tract and an additional one to accommodate nasalization. This was later extended by the same author to include four more filters to cover frequencies up to 8 kHz [46].

Comparison of the cascaded and parallel configurations is presented rigorously by Holmes [45]. In theory, both structures should be able to produce similar vocal tract transfer functions once the driving parameters are suitably adjusted. The main advantage of cascaded formant synthesizers is their ability to produce natural-sounding vowels without the need to control the relative amplitude of the individual formant resonators. These synthesizers, however, cannot easily accommodate changes of vocal effort. Parallel

formant synthesizers, on the other hand, are capable of compensating for spectral variations due to the excitation and can produces fricatives and plosives adequately by simple control over the spectral levels in critical band regions above 3 kHz. To take advantage of both structures, Klatt introduced a hybrid synthesizer (*Klattalk*) [59] which requires 39 control parameters and employs cascaded resonators for generating voiced sounds and parallel resonators for producing fricatives. The block diagram of the Klattalk formant synthesizer is shown in Fig. 2.4.

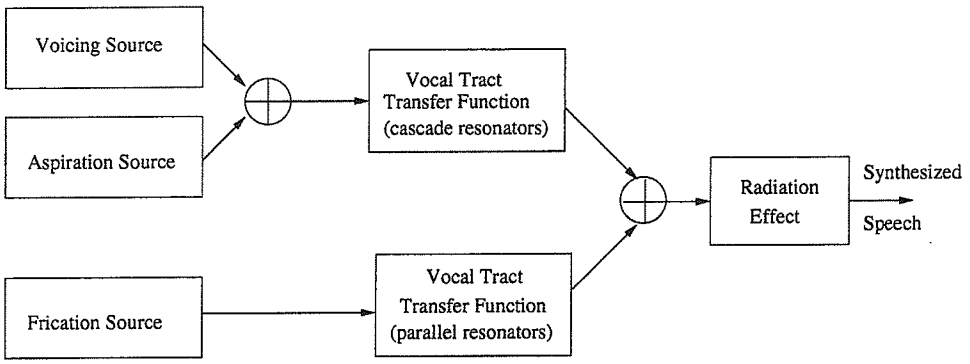


Figure 2.4 Block diagram of the Klattalk formant synthesizer.

Several models have been proposed for voiced excitation of formant synthesizers. Holmes [44] examined variations of glottal waveforms for different speakers by inverse filtering the speech signal. His finding showed that the low frequency harmonics of different excitation waveforms are similar and, perceptually, only the fundamental frequency ,the *open quotient* (ratio of pulse duration to pitch period) of the pulse shape and the *speed quotient* (ratio of rising pulse duration to falling pulse duration) are important parameters. He then recommended a glottal pulse produced by a spectral flattening procedure from the speech signal of a specific speaker. Other more sophisticated models have been reported by Fant *et al.* [29] and by Klatt [59]. For voiceless excitations, friction noise is introduced at the selected frequencies [59]. The ability to control the acoustic properties

of the vocal tract directly by formant synthesis is important for many psycho-acoustic studies. Their application to formant vocoders and synthesis by rule has been successful.

2.1.3 Articulatory Synthesis

An alternative solution to generating synthetic speech is to model the physical human vocal apparatus. This is known as *articulatory synthesis*. Articulatory synthesis essentially contains two parts: a vocal fold model which represents the excitation source and a vocal tract model which describes the positions of the articulators such as tongue, lips and jaw. It is expected that in articulatory synthesis high quality speech will be generated if the positions of the vocal apparatus are defined explicitly and an excitation source is provided which can interact appropriately with the acoustic input impedance of the vocal tract, thus providing a more exact model of the speech production mechanism. Digital synthesizers that model the speech production mechanism directly have been extensively studied in the past decades. They are essentially different in the method employed for modeling the excitation source and the vocal tract. Flanagan *et al.* [35] and Maeda [68] used a set of linear and non-linear differential equations to characterize the wave propagation in the vocal system. Models which employ types of wave-digital filters have been described by Kelly and Lochbaum [57], Liljencrants [66] and Meyer *et al.* [72]. Sondhi and Schroeter developed a hybrid synthesizer that models the glottis in the time domain and the vocal and nasal tracts in the frequency domain [110]. Levinson and Schmidt proposed an unconstrained optimization technique to estimate the static articulatory parameters of English phonemes based on minimizing the difference between the natural speech spectra and the model speech spectra computed from the lossy Webster equation [64]. Hasegawa-Johnson has developed a low-complexity finite element model

of the vocal folds, and a multi-speaker MRI database of three-dimensional vocal tract shapes of vowels [39].

The advantages of articulatory synthesis are mainly the following.

1. Nonlinear interaction can be employed between the excitation source and the acoustic input impedance of the vocal tract. Conventional speech synthesizers based on linear prediction of the speech waveform assume source-tract separability. This is unrealistic for the production of consonants such as fricatives and plosives since the position and amplitude of the excitation signals are governed by the shape of the vocal tract. In our study, we even find out strong source-tract interaction in the production of voiced sounds.
2. Since the control parameters of an articulatory speech synthesizer are expected to change very slowly in time to model the movements of the human articulators, they form potential candidates for very low bit rate speech coding applications. Interpolation of these parameters over several frames of speech generates reasonable vocal tract shapes, whereas, interpolation of LPC coefficients can give rise to unrealistic speech sounds.
3. For TTS applications, articulatory synthesis use rules which naturally relate to articulator movements to deal with the problems of coarticulation and unit concatenation, which are difficult problems in formant synthesis and concatenative synthesis.
4. Theoretically speaking, articulatory synthesis has the potential of producing human-like speech sounds although the computational complexity of constructing accurate mechanical vocal folds model and the human-like vocal tract shape and the solution of highly-nonlinear differential equa-

tions governing the sound propagation inside vocal apparatus is still too high even for the currently available super-computers.

5. Articulatory synthesis offers a more direct and elegant way to synthesize personalized speech sounds. For example, in order to synthesize the speech sounds of different speakers such as a tall male, an average male and a young boy or sounds under different stress conditions, conventional synthesizers rely on some empirical rules and certain minimum error criterion to find a set of sub-optimal parameters for the waveform modification for a specific speaker. Since most of the parameters in conventional speech synthesizer are not directly relate to the physical parameters involoved in human speech production, the synthesized speech waveform is unsatisfactory in many cases. In the case of articulatory synthesis, we can easily change the value of several articulatory parameters such as vocal tract length and expect to synthesize personalized speech sounds very close to human sounds under different scenarios.

An articulatory synthesis system has been proposed by Rahim *et al.* [84], [85], [86], [88] essentially based on artificial neural network (ANN). Our work differs from his work in the following aspects.

1. In Rahim's work, a multi-layer perceptron (MLP) neural network was extensively used to model the non-linear relationship between the acoustic representation of speech signal and the vocal tract shapes. An acoustic-to-articulatory codebook was generated by the training data to cover an adequate span of the entire articulatory space of voiced and nasal sounds. Then, a dynamic programming (DP) approach was used to select an "optimal" trajectory of vocal tract shapes from the codebook by minimizing

a sum of acoustical and geometrical cost components over several frames of speech. On the other hand, in our work, an advanced digital signal processing (DSP) approach is used to estimate the trajectory of articulatory parameters from the static articulatory parameters. Instead of generating a very large codebook which trie to cover the whole range of articulatory space, our approach just needs the estimation of static articulatory parameters. Then the trajectory of articulatory parameters is estimated using advanced digital signal processing (DSP) technique. Finally, the estimated dynamic articulatory parameters were mapped into the moving vocal shapes.

2. Rahim's articulatory synthesizer assumes one-dimensional plane-wave propagation inside a lossy vocal tract. In our work, sound propagation inside the vocal apparatus is considered to to be a non-plane wave propagation inside a viscous fluid with the governing Navier-Stokes equations.

3. The glottal excitation in Rahim's synthesizer was represented by a parametric model. In our work, the excitation signal is generated by a multimass mechanical model of the vocal fold and a closer and more realistic source-tract interaction is investigated in our approach.

2.2 Overview of Speech Production Model

2.2.1 Source-filter Speech Production Model

One of the most popular models in speech signal processing is source-filter speech production model In this model, the source waveform is filtered by a time-varing linear filter with its spectrum peaks corresponding to the vocal tract resonace. The source

filter model is computationally fast and provides a vital tool for estimating basic speech parameters such as pitch, formants and pseudo vocal tract areas. Fig. 2.5 shows the system diagram of source-filter model [81].

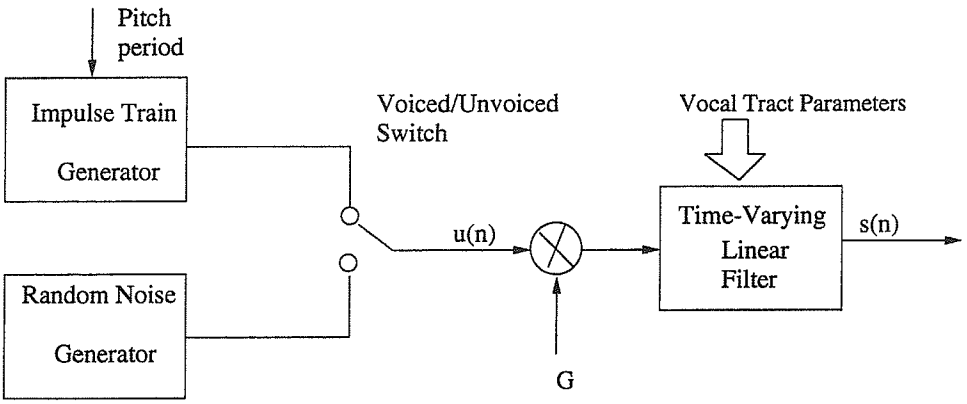


Figure 2.5 Diagram of the Source-filter speech production model

In signal processing scenario, the source-filter model can be described by linear prediction (LP) analysis. The basic idea of LP analysis is to perform prediction of a speech sample from knowledge of past samples. The prediction error is the difference between the predicted sample and the actual one, and when minimized, leads to an optimized set of predictor coefficients. In an LP vocoder, speech is classified as either voiced or unvoiced. During voiced regions, the filter is excited by quasi-periodic pulses generated at intervals of averaged pitch periods. While for voiceless sounds, the excitation signal is provided by a white noise generator. The excitation is scaled by a gain factor G which is determined by matching the energy of original and synthetic segments of speech [81]. A complementary feature in LP analysis is that efficient and simple procedures are available for estimating the AR coefficients a_k and the gain factor G. There are two major implementations of linear prediction: the *auto-correlation* method [70] and the *covariance* method [2]. The difference of these two methods lies in their segmentation of the speech signal.

Linear prediction methods usually assume that the signal being processed is stationary within the analysis interval. It is therefore necessary to perform linear prediction over a segment of speech where the vocal tract movement is negligible. The quality of the output speech varies according to the position and size of the analysis window as well as the order of the predictive filter and the analysis method employed for estimating the prediction coefficients. A study of the variation of the prediction error with the position of the analysis window is discussed in [81]. There is usually a trade-off between the quality of the synthesized speech and the update rate of the filter parameters. For example, during regions of unvoiced speech, the filter parameters are changed at regular intervals (typically 10 ms). While for voiced segments, these parameters are updated at the beginning of each pitch period (*pitch-synchronous*) or per fixed frame size (*pitch-asynchronous*). Pitch-synchronous analysis has been found to be a more effective synthesis method, and is commonly used for estimating pseudo vocal tract area functions.

2.2.2 Fricative Model

In his Ph.D. thesis [105], Sinder proposed a fricative model which combines an existing model for acoustic propagation with a reduced model for the structure of turbulent jets. This model is based on physical principles from aeroacoustics and unsteady aerodynamics and is specifically applicable to articulatory speech synthesis.

As described in Chapter 1, Fricatives are produced when air from the lungs is forced through a tight constriction in the vocal tract forming a high speed jet. At the constriction exit, the jet becomes turbulent and generates flow induced noise which excites resonances of the vocal tract. Sinder's fricative model is based on Howe's aeroacoustics and aerodynamics theory of sound generation [49]. In this theory, the information re-

quired to model the generation of aerodynamic sound and its propagation inside a vocal tract can be parameterized by three component models as following.

1. *Jet Model:* A description of the formation and convection of vorticity, including vortex strengths, speeds and trajectories.
2. *Mean Potential Flow Model:* A model for the direction of the irrotational mean flow.
3. *Acoustic Propagation Model:* A model which solves for acoustic wave propagation in the duct given a description of acoustic source present.

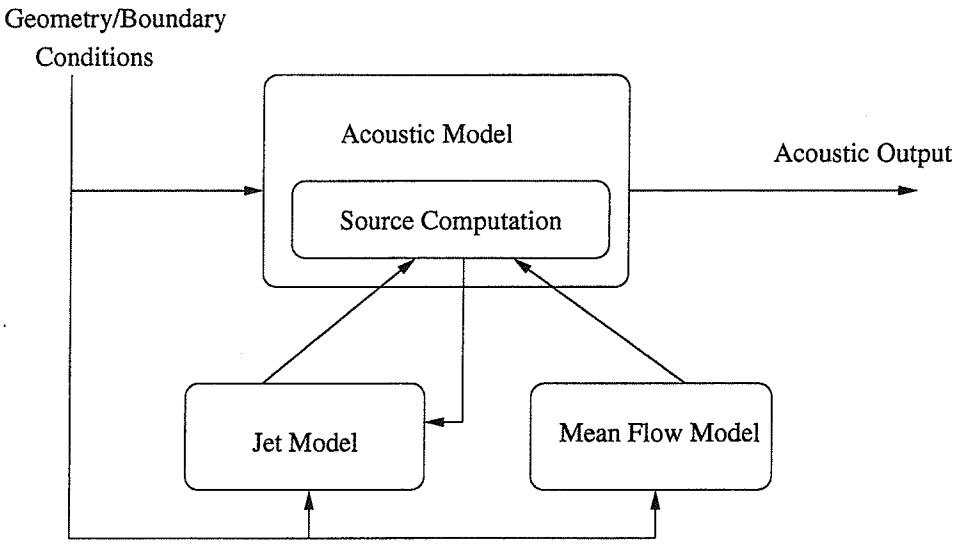


Figure 2.6 System diagram of Howe's aerodynamic sound generation model.

Fig. 2.6 illustrates the components for computing aerodynamic sound generation using Howe's source term. In the following sections, we will briefly discuss the core components of Sinder's fricative model based on aeroacoustics and aerodynamics [105].

2.2.2.1 Acoustic Model for Unsteady Potential Flow

The acoustic model computes the small perturbations of velocity and pressure associated with acoustic resonance of the vocal tract. The conventional transmission line

analogy is used to model plane wave propagation in the vocal tract. The transmission line T-networks include resistors representing *viscous loss*, inductors representing *fluid inertia*, and capacitors representing *fluid compliance*. A series pressure source is included in each T-network to allow the introduction of source pressure computed either from Howe's source term or a random process generator. In series, a supplemental resistance is also included to incorporate losses due to flow separation where a jet is formed. Such losses will be discussed later. Omitted are additional impedance to account for heat conduction through the wall and wall vibration. These elements generally result in slight increases in formant bandwidths, especially at low frequencies due to the wall vibration. Contributions to the acoustic field outside the mouth due to sound radiation through the vocal tract wall are neglected. Also, a nasal tract branch is not included since the nasal tract is not involved in fricative production. The transmission line equations are solved in the time domain in a fashion similar to the approach described in [54].

The inlet boundary condition specifies a volume source at the inlet. In speech production, lungs serve as a reservoir of air which is supplied to the vocal tract by increasing subglottal pressure. This mass source can be modulated by the valving action of the vocal folds during phonation. In the case of voiced fricatives, the inlet mass supply acts as a monopole acoustic source which excites the vocal tract. For unvoiced fricatives, a steady flow of air is supplied from the lung and the vocal tract is acoustically excited by a downstream acoustic source which can be modeled by a dipole acoustic source. Two inlet boundary conditions are used in Sinder's model. The first one supplies a steady volume source at a constant duct area by fixing the volume velocity there. This inlet condition is used for fixed vocal tract shapes when simulating sustained, unvoiced fricatives. The second inlet condition, which models the monopole source due to phonation, is Ishizaka and Flanagan's two-mass vocal fold model [54]. The two-mass model provides broad band excitation of the vocal tract when simulating voiced speech. When the glottis rest

area is set sufficiently wide open, the folds stop vibrating and a steady volume velocity is supplied to the vocal tract. Thus, this model is used to simulate fricatives in a vowel context.

The outlet boundary conditions is applied by terminating the transmission line with a radiation impedance which approximates the radiation characteristics of a piston in an infinite plane baffle [33]. Specifically, a parallel combination of a resistance R_{out} and inductance L_{out} is used in the following expressions.

$$R_{out} = \frac{128\rho c}{9\pi^2 A_{out}} \tag{2.1}$$

$$L_{out} = \frac{8\rho}{3\pi A_{out}} \sqrt{\frac{A_{out}}{\pi}} \tag{2.2}$$

where A_{out} is the area of the outlet opening.

2.2.2.2 Mean Flow Model for Steady Potential Flow

The potential flow is that portion of the flow which is irrotational. In Howe's source term, the relative directions of vortex trajectories and potential flow are critical. Thus, the magnitude of the mean flow alone is not sufficient. Its direction as it flows around obstacles is a key component, which is why the mean flow model discussed below is important. For speech synthesis, a solution for arbitrary geometries is desirable, as is a solution which is computationally reduced. Therefore, rather than numerically solve a set of partial differential equations for potential flow in a fine spatial grid, a simplified model was developed. This simplified mean flow model has two major components, namely approximation of the flow direction and the computation of streamlines.

The directional information is computed by assuming that the flow uniformly expands and contracts through changes in the cross-sectional area of the vocal tract. Consider a short length of the tract as shown in Fig. 2.7 [105] with endpoints x_n, and x_{n+1}, radii

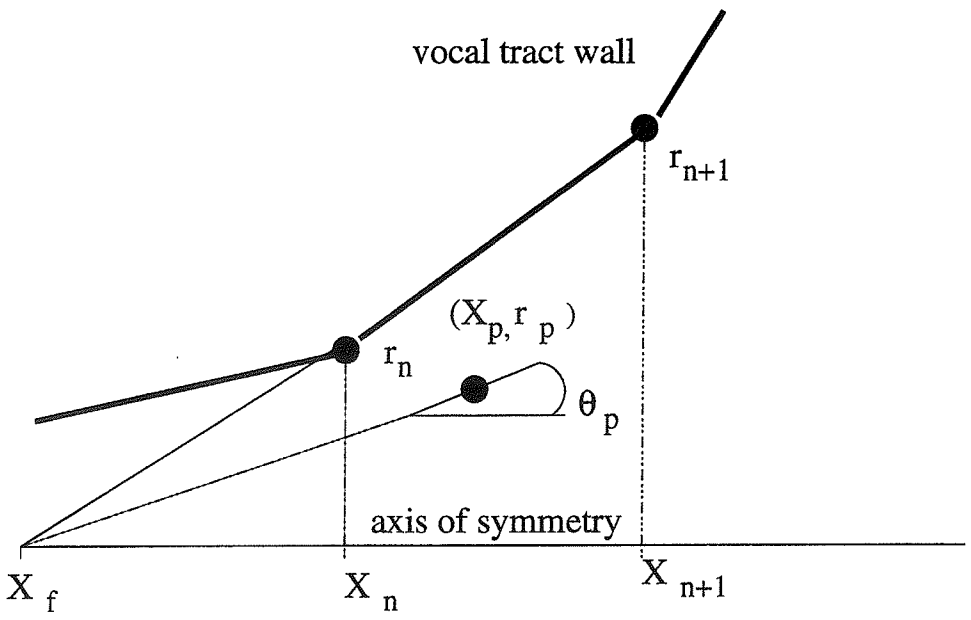


Figure 2.7 Illustration of flow direction computation for flow from left to right.

r_n and r_{n+1}, and flow from left to right. The flow direction at an arbitrary point P at (x_p, r_p) within the small section is determined to be the same as the vector from the focus point x_f to point P. Specifically, angle θ_p between the flow direction and the axis of symmetry is given by:

$$\theta_p = tan^{-1}(\frac{r_p}{x_p - x_f}) \tag{2.3}$$

$$x_f = x_n - r_n \frac{(x_{n+1} - x_n)}{(r_{n+1} - r_n)} \tag{2.4}$$

In fluid mechanics, lines which are everywhere tangential to the velocity vector are called *streamlines* [17]. The approximate streamlines can be computed by assuming that all streamlines have the same focus point within the section. Note that the distance between streamlines can be interpreted in terms of flow velocities. That is, denser streamlines indicate higher particle velocities.

2.2.2.3 Jet Model for Rotational Flow

Fluid flow can be decomposed into irrotational and rotational components. We will briefly describe Sinder's model for the rotational component, its structure and evolution in this section. Note that this model is a highly reduced description which still remains faithful to the physical principles important for noise generation.

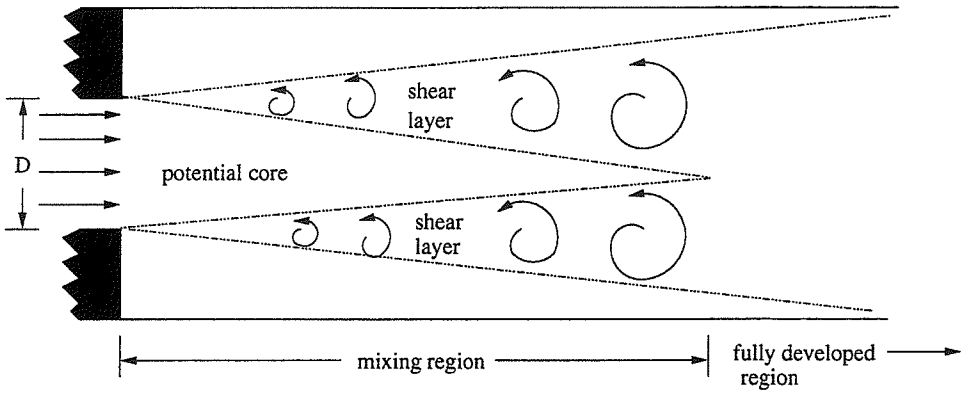


Figure 2.8 Illustration of streamline computation for flow from left to right.

At a sudden expansion in a duct, high speed flow can separate from the relatively stagnant fluid on the downstream side of the expansion wall, resulting in a turbulent jet. Vorticity, initially present in the wall boundary layer at the jet exit, causes the jet to mix with the surrounding fluid reservior through the action of viscous and turbulent diffusion. At first, this mixing occurs along the boundary of the jet in a shear layer which grows approximately linearly with distance from jet formation. In an axisymmetric jet, the shear layer is annular in shape. Eventually, the shear layer merges with itself, closing the annulus. At this point, the potential core of the jet ends, and the jet is fully developed. Fig. 2.8 illustrates this structure [105].

The shear layer is dominated by the largest vortices (also called *coherent structure*) which control the mixing rate and are nearly as wide as the shear layer itself. These large eddies shed from a location near the jet exit at nearly regular intervals forming an array

of vortex rings which convect downstream. This arrangement is referred to as a *vortex street.* The jet model is based on the structure described above, with particular focus on large eddy structures of the vortex street concept which play an important role in sound generation. This model has three major components, namely vortex formation, vortex strength and velocity, and vortex motion and dissipation.

The separation points for vortex formation are determined though automatic shedding criteria. The criteria contain three conditions, i.e. (1). the minimum area (between 0.001 cm^2 and 0.4 cm^2); (2). the Mach number (greater than 0.01); (3). the first two conditions exists for at least 2 ms. The time interval between shede vortices is modeled as a random variable T_{shed} with its mean value given by:

$$E\{T_{shed}\} = \frac{D}{StU_j} \tag{2.5}$$

where D is the diameter of jet exit, St is the Strouhal number and U_j is the jet velocity, respectively.

The vortex strength and the velocity are given by:

$$\Gamma = 0.5U_j^2 T_{shed} \tag{2.6}$$

$$\vec{v}_{new} = (1-\alpha)\vec{v} + \alpha\frac{v_x}{\hat{w}_x}\hat{\vec{w}} \tag{2.7}$$

where Γ denotes the vortex circulation, $\vec{v}_{new}$ denotes the velocity of the new vortex, $\vec{v}$ denotes the velocity of the previous vortex at axial location x, $\hat{\vec{w}}$ denotes an unit vector with its direction as the vocal tract wall at axial location x, v_x and w_x denotes the components of $\vec{v}$ and $\hat{\vec{w}}$ in the axial direction, respectively. The parameter α is the cosine of the angle between the local wall and the direction of vortex velocity $\vec{v}$.

Vortices in the flow are managed in a module called vortex tracker which associate each vortex with its own location, velocity, strength and lifetime. The velocity and

strength of vortex are computed in Eqs. (2.6) - (2.7). The lifetime is specified as the axial distance the vortex travels and is experimentally determined to be $4D$. Since the vortices move much slower than the sound propagation, a first order approximation is sufficient for updating their positions. The vortex location $\vec{x}_n$ at iteration n is given by:

$$\vec{x}_n = \vec{x}_{n-1} + \vec{v}_{n-1}T \tag{2.8}$$

Finally, the aeroacoustic source contribution to the acoustic field is computed at each iteration by computing the contribution due to each vortex contained within the vortex tracker. The sound pressure p due to each individual vortex is given by:

$$p = \frac{-\rho_0}{A}\int_0^{2\pi}[\Gamma \times \vec{v} \cdot \hat{\vec{U}}]d\theta = \frac{-2\pi r_\omega \rho_0}{A}[\Gamma \times \vec{v} \cdot \hat{\vec{U}}] \tag{2.9}$$

where A is the vocal tract area at the vortex location, r_ω is the radius of the vortex ring, $\hat{\vec{U}}$ is an unit vector with the mean flow direction, ρ_0 is the ambient density of air. The symbols $\times$ and $\cdot$ denotes the outer product operation and inner product operation, respectively. As vortices convect, they excite different tubelets along the vocal tract. therefore, the source is both spatially and temporally distributed.

2.2.3 Unvoiced Speech Sound Production Model

In his recent paper [62], Krane extend Howe's solution of convective wave propagation and proposed a general unvoiced speech production model. The physics of sound production by vocal tract airflow is not only the primary mechanism of unvoiced sound production, but also a secondary source of sound in voicing. The sound propagation inside vocal apparatus consists of two modes, namely the *sound mode* (irrotational and compressible) and the *flow mode* (rotational and slightly compressible). The sound or propagation mode has been extensively studied using the linear acoustics and signal pro-

cessing techniques and result in successful speech production model such as the source-filter model described in previous section. On the other hand, the study of the flow or convective mode and its contribution to sound production have been neglected or avoided in the speech science literature for a long time. Traditional approach to analyze air motion in vocal tract has following misunderstandings. First, the flow is assumed to be irrotational and quasi-steady, and therefore the Bernoulli equation is sufficient to describe the relationship between pressure and particle velocity. Second, the acoustic excitation due to the flow is simply described by band-limited white noise. On the other hand, it is shown that in general, the characteristics of the aeroacoustic source are governed by the strength and spatial distribution of the jet velocity field, the convection speed and temporal spacing of vortical disturbances and the axial extent and shape of the vocal tract. For turbulent jets, the shape of the vocal tract is the dominant factor in determining the spectral content of the source.

Krane's model provides a complete framework for issues regarding aeroacoustic source strength, spatial distribution, frequency content, and impedance. Through the study of aeroacoustics, which describes the interaction between flow mode and sound mode, Krane's model leads to three major findings. First, the key ingredient of the flow mode is the *fluid vorticity*, which acts as a means of storing air inertia as rotational motion of fluid particles. Second, due the low-Mach number, high-Reynolds number nature of flows in unvoiced speech production, the production of sound by speech airflows is very inefficient and in general does not affect the behavior of the flow mode. The aeroacoustic source characteristics depend on the shape of the vocal tract and the airflow speed. Third, if the source region is acoustically compact, the source spectrum can be written as the convolution of two signals, namely the waveform of a single vortex passing through the source region and a function describing the arrival of vortices in the source region, scaled by the strength of the particular vortex. It's the vortex arrival statistics that determine

whether the source spectrum is dominated by the vorticity field or the wall shape. If the vortex arrival time series is highly coherent periodic, it will dominate the source spectrum. On the other hand, if the vortex arrival time series is broadband, the vocal tract shape will dominate the source spectrum. More details can be found in [62]

2.3 Overview of Articulatory Speech Model

2.3.1 Coker's Model

In 1976, Coker proposed an articulatory model for unlimited vocabulary speech synthesis [14]. Fig. 2.9 shows the representation of the midsagittal section of the human vocal tract. The articulator was described by eight physical parameters : tongue body height (Y); anterior/posterior position of the tongue body (X); pharyngeal opening (P); tongue tip height (B); tongue tip curliness (R); lip opening (W); lip roundness (C); and vocal tract length (L). In 1983, Levinson and Schmidt [64] proposed an unconstrained optimization technique to estimate the Coker's articulatory model parameters based on minimizing the difference between the natural speech spectra and the model spectra computed from the Webster equation. We implemented this technique on a Sun workstation to get the stationary articulatory model parameters for fifty-two English phone units.

...

2.3.2 Mermelstein's Model

In 1973, Mermelstein proposed an articulatory model as shown in Fig. 2.10. This model permits simple control over a selected set of articulatory parameters, namely, tongue body center (C), velum (V), tongue tip (T), jaw (J), lips (L) and hyoid (H) [71].

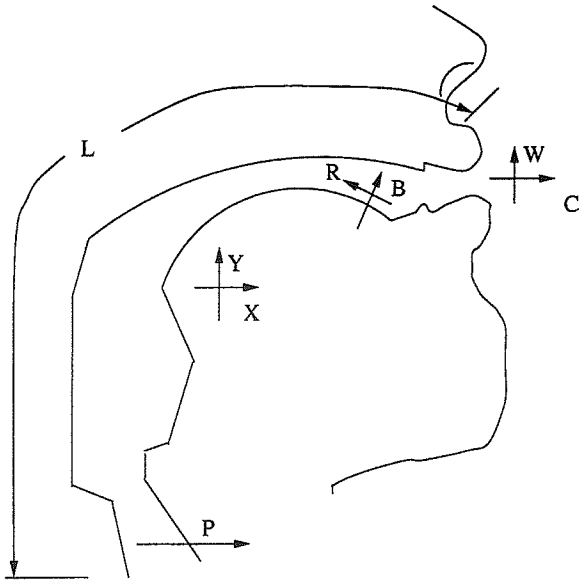


Figure 2.9 Coker's articulatory model.

The positions of these articulators determine the vocal tract length and cross-sectional areas, following the estimation of the vocal tract profile in the midsagittal plane.

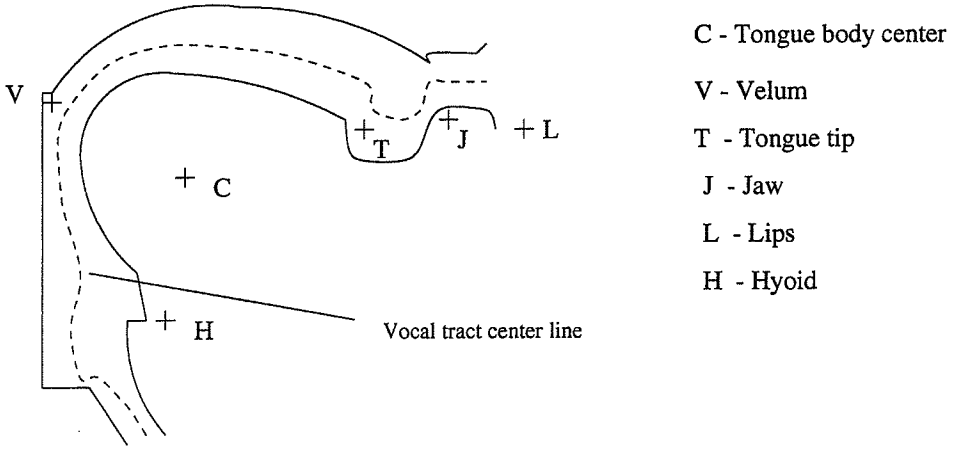


Figure 2.10 Mermelstein's articulatory model.

Mermelstein's model consists of fixed and movable structures to enable it to articulate movements of the vocal tract during speech production. Fixed structures are approximated by straight segments and circular arcs, and represent the rear pharyngeal wall (GR), the soft palate (RM), the hard palate (MN) and the alveolar ridge (NU). These

structures form the posterior-superior vocal tract outline which is considered fixed with the exception of the region among the upper lips. Movable structure approximate the inferior-anterior vocal tract outline and are classified into two categories: those whose movements are independent of other articulators (velum, hyoid bone and jaw), and articulators whose position is a function of other articulators (tongue body, tongue tip and lips). For example, the movement of the tongue tip is relative to that of the tongue body which itself depends on the position of the jaw. The jaw and the velum are considered movable structures with one degree of freedom. All other articulators have two degrees of freedom.

Table 2.1 Identification of parameters and imposed limits for Mermelstein's articulatory model [71].

No.	Name	Meaning	Lower limit	Upper limit
1	rtp	tongue radius $[cm]$	0.5	3.5
2	tejp	jaw angle $[deg]$	10.9	28.0
3	xcp	tongue center x $[cm]$	4.0	10.0
4	stp	tongue blade length $[cm]$	2.0	4.4
5	xclp	lip protrusion x $[cm]$	0.36	1.89
6	xhp	hyoid position x $[cm]$	5.20	7.60
7	ycp	tongue center y $[cm]$	1.48	7.48
8	telp	tongue elevation $[deg]$	68.2	96.8
9	cylp	lip height y $[cm]$	0.93	1.32
10	yhp	hyoid position y $[cm]$	7.43	9.83
11	vel	velum opening $[cm^2]$	0	2

By proper selection of the various positions of the articulators, Mermelstein's model is able to generate physiologically possible shapes that span the entire articulatory space. Table 2.1 shows the limits of the articulatory parameters in Mermelstein's model

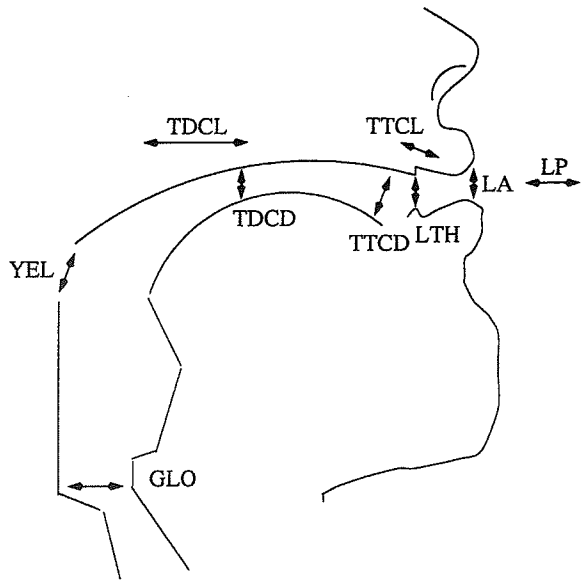


Figure 2.11 The task dynamic articulatory model, with tract variable degrees of freedom indicated by arrows.

2.3.3 Haskins Lab's Model

During the past decades, Saltzman and Rubin *et al.* in Haskins's lab have proposed a task-dynamic model for speech production [93], [95], [96]. Their model was constructed based on a midsagittal view of the vocal tract and a simplified kinematic description of the vocal tract's articulatory geometry. Fig. 2.11 shows the midsagittal view of the task-dynamic model. The presented model are associated with the control of bilabial, tongue-dorsum and lower-tooth-height constrictions. The creation and release of constrictions of differing degree in different regions of the vocal tract are described in two steps. The first step entails defining time-invariant dynamics at the level of tract-variable coordinates. The second step is to transform the tract-variable system into an explicitly articulatory set of coordinates to describe the articulatory details required for an adequate simulation of speech production. Table 2.2 shows the relationship between tract-variables and model articulators.

Table 2.2 Relationship between tract-variables and model articulators.

Name	Tract variables	Model articulators
LP	lip protrusion	upper and lower lips
LA	lip aperture	upper and lower lips, jaw
TDCL	tongue dorsum constrict location	tongue body, jaw
TDCD	tongue dorsum constrict degree	tongue body, jaw
LTH	lower tooth height	jaw
TTCL	tongue tip constrict location	tongue tip, tongue body, jaw
TTCD	tongue construct degree	tongue tip, tongue body, jaw
VEL	velic aperture	velum
CLO	glottal aperture	glottis

In a more recent reference, Rubin *et al.* have introduce three more tract variables, namely, subglottal pressure, transglottal pressure and delta virtual fundamental frequency to extend the previous task-dynamic model for better description of speech production. Details of the new model can be found in [94].

2.4 Overview of the Motor Control of the Articulator

The production of speech is portrayed traditionally as a combinative process that uses a limited set of units to produce a very large number of linguistically "well-formed" utterances [13]. Segmental speech units (such as phonemes) are usually seen as discrete, static and invariant across a variety of contexts. Putatively, such characteristics allow speech production to be generative, because units of this kind can be concatenated easily in any order to form new strings. During speech production, the shape of the vocal tract changes constantly over time. These changes in shape are produced by the movements of a number of relatively independent articulators (e.g., velum, tongue, lips, jaw *etc*).

Although the speech segmental units are discrete, the articulatory patterns movements of different articulators are interleaved into a continuous gestural flow.

Much theoretical and empirical evidences from the study of skilled movements of the limbs and speech articulators supports the hypothesis that significant informational units of action do not entail rigid or hard-wired control of joint and muscle variables. Rather, these units or coordinative structures must be defined abstractly or functionally in a task-specific, flexible manner. In next section, we want to investigate the motor control of the speech articulator according to some dynamical model or minimum cost hypothesis.

2.4.1 A Dynamical Model of Articulation

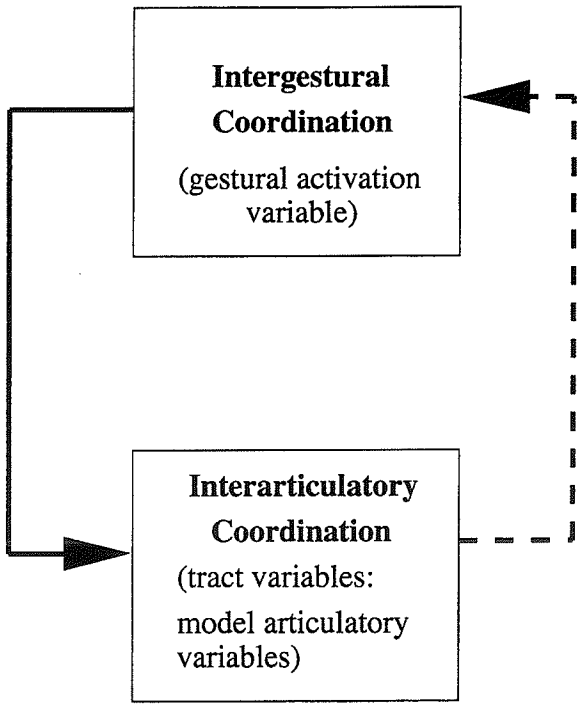


Figure 2.12 Schematic illustration of the two-level dynamical model for speech production.

In [95], Saltzman and Munfall proposed a dynamical model of articulation with two functionally distinct but interacting levels. The *intergestural* level is defined according to both *model articulator* and *tract variable* coordinates. Tract variables (e.g. bilabial aperture) are the coordinates in which context-independent gestural "intents" are framed; and model articulators (e.g. lips and jaw) are the coordinates in which context-dependent gestural performance are expressed. Fig. 2.12 shows the schematic illustration of the two-level model for speech production, with associated coordinate systems indicated. The solid arrow line from the intergestural to the interarticulator levels denotes the feedforward flow of gestural activation. The dashed arrow line indicates feedback of ongoing tract variable and model articulator state information to the intergestural level. Invariant gestural units are posited in the form of relations between particular subsets of these coordinates and sets of context-independent dynamical parameters (e.g. target position and stiffness). Contextually-conditioned variability across different utterances results from the manner in which the influences of gestural units associated with the utterances are gated and blended into ongoing processes of articulatory control and coordination. The activation coordinate of each unit can be interpreted as the strength with which the associated gesture attempts to shape vocal tract movements at any given point in time. The tract variable and model articulator coordinates of each unit specify the particular vocal-tract constriction (e.g. bilabial) and set of articulators (e.g. lips and jaw) whose behaviors are directly affected by the associated unit's activation. The intergestural level accounts for patterns of relative timing and cohesion among the activation intervals of gestural units that participate in a given utterance, e.g., the activation intervals for tongue-dorsum and bilabial gestures in a vowel-bilabial-vowel sequence. The interarticulator level accounts for the coordination among articulators evident at a given point in time due to the currently active set of gesture, e.g., the coordination among lips, jaw, and tongue during periods of vocalic and bilabial gestural coproduction [95].

These attributes of task-specific flexibility, functional definition, and time-invariant dynamics can be incorporated into a *task − dynamic* model of coordinative structures. Each gesture in an utterance is associated with a corresponding tract-variable dynamical system. In this approach, all such dynamical systems are defined as tract-variable point-attractors, i.e., each is modeled by a damped, second-order linear differential equation (analogous to a damped mass-spring). The tract-variable motion equations are defined in matrix form as follows:

$$\ddot{\vec{z}} = M^{-1}(-B\dot{\vec{z}} - K\triangle\vec{z}) \tag{2.10}$$

where $\vec{z}$ is the $m \times 1$ vector of current tract-variable positions, $\dot{\vec{z}}$ and $\ddot{\vec{z}}$ are the first and second derivatives of $\vec{z}$ respect to time; M is a $m \times m$ diagonal matrix of inertial coefficients; B is a $m \times m$ diagonal matrix of tract-variable damping coefficients; K is a $m \times m$ diagonal matrix of tract-variable stiffness coefficients; and $\triangle\vec{z} = \vec{z} - \vec{z}_0$ where z_0 is the target or rest position vector for the tract variables. The components of B, K, and $\vec{z}_0$ vary during the utterance according to the ongoing set of gestures being produced. For example, different vowel gestures are distinguished in part by corresponding differences in target positions for the associated set of tongue-dorsum point attractors. Similarly, vowels and consonant gestures are distinguished in part by corresponding differences in stiffness coefficients, with vowel gestures being slower (less stiff) than consonant gestures. Thus Eq. (2.10) describe a linear system of tract-variable equations with time-varying coefficients, whose values are functions of the current active gesture set.

Furthermore, a dynamical system for controlling the model articulators is specified by expressing tract variables ($\vec{z}$, $\dot{\vec{z}}$, $\ddot{\vec{z}}$) as functions of the corresponding model articulator variables ($\vec{\phi}$, $\dot{\vec{\phi}}$, $\ddot{\vec{\phi}}$). The tract variables of Eq. (2.10) are transformed into model articulator variables using the following direct kinematic relationships:

$$\vec{z} = \vec{z}(\vec{\phi}) \tag{2.11}$$

$$\dot{\vec{z}} = J(\vec{\phi})\dot{\vec{\phi}} \tag{2.12}$$

$$\ddot{\vec{z}} = J(\vec{\phi})\ddot{\vec{\phi}} + \dot{J}(\vec{\phi}, \dot{\vec{\phi}})\dot{\vec{\phi}} \tag{2.13}$$

where $\vec{\phi}$ is the $n \times 1$ vector of current articulator positions. $J(\vec{\phi})$ is the $m \times n$ Jacobian transformation matrix whose elements J_{ij} are partial derivatives $\frac{\partial z_i}{\partial \phi_j}$ evaluated at the current $\vec{\phi}$. The elements of J and $\dot{J}$ reflect the geometrical relationships among motions of the model articulators and motions of the corresponding tract variables. Using the direct kinematic relationships in Eqs. (2.11) - (2.13), the equation of motion derived for the actively controlled model articulators is given by:

$$\ddot{\vec{\phi}}_A = J^*(M^{-1}[-BJ\dot{\vec{\phi}} - K\triangle\vec{z}(\vec{\phi})]) - J^*\dot{J}\dot{\vec{\phi}} \tag{2.14}$$

where $\vec{\phi}_A$ is an articulatory acceleration vector representing the active driving influences on the model articulator; $J^* = W^{-1}J^T(JW^{-1}J^T)^{-1}$ is a $n \times m$ weighted Jacobian pseudoinverse where W is a $n \times n$ positive definite articulatory weighting matrix whose elements are constant during a given isolated gesture. The pseudoinverse is used because there are a greater number of model articulator variables than tract variables for this task. More specifically, using pseudoinverse provides a unique, optimal least square solution for the redundant differential transformation from tract variables to model articulator variables that is weighted according to the pattern of elements in the W matrix.

2.4.2 Motor Control Based on Minimum Cost Principles

In 1983, Nelson proposed several articulator motor control system based on some minimum cost principles [74]. For displacement of a mass m, along the dimension x with instantaneous velocity v, the equations governing the motion are:

$$\frac{dx}{dt} = v, \quad \frac{d(mv)}{dt} = f_a(t) - f_d(t) \tag{2.15}$$

where $f_a(t)$ is the net applied force along the x dimension and $f_d(t)$ is the net dissipative force (friction) resulting from the movement. We assume for the range of displacements and velocities remaining within some linearity region, L, that the mass is constant and the dissipative force is a linear function of the velocity. We further assume that there is some limit F_{max}, on the magnitude of the applied force. If we now define the control action $u(t)$ as applied force per unit mass (acceleration), we can get:

$$u(t) = \frac{f_a(t)}{m}, \quad |u(t)| \leq U_{max} \equiv \frac{F_{max}}{m} \tag{2.16}$$

The equations of motion can be written in the normalized form:

$$\dot{x}(t) = v(t), \quad x(0) = 0, \quad x(T) = D \tag{2.17}$$

$$\dot{v}(t) = u(t) - bv(t), \quad v(0) = 0, \quad v(T) = 0, \quad |u(t)| \leq U_{max} \tag{2.18}$$

Because they are invariant under translation in x, Eqs. (2.17) - (2.18) can describe each segment of a sequence of motions along the dimension x of various distances D and movement time (durations) T between successive equilibrium states ($x = x_i, \quad v = 0$). Each segment of the sequence depends only on the initial state ($x_i, \quad 0$) and the control force action $u(t)$ over the current movement segment. Thus the control strategy for each segment may be considered independently from that in the other segments.

The performance objectives can be expressed in terms of minimizing some measure of physical cost associated with accomplishing the movement. Five measures of cost might be considered in relation to skilled movements are:

$$min\{time \ \ cost\} = min\{T\} \tag{2.19}$$

$$min\{force\ \ cost\ = min\{A\} = min\{max_{t\in(0,T)}|u(t)|\} \tag{2.20}$$

$$min\{impulse\ \ cost\} = min\{I\} = min\{\frac{1}{2}\int_0^T |u(t)|dt\} \tag{2.21}$$

$$min\{energy\ \ cost\} = min\{E\} = min\{\frac{1}{2U_{max}}\int_0^T u^2(t)dt\} \tag{2.22}$$

$$min\{jerk\ \ cost\} = min\{J\} = min\{\frac{1}{2}\int_0^T \dot{a}^2(t)dt\} \tag{2.23}$$

where $\dot{a}$ is the rate of change of acceleration (jerk) of the movement. Note that the minimum-jerk cost solution will result in $x(t)$ trajectories which are equivalent to the cubic spline interpolation results. Furthermore, the $x(t)$ trajectory of the minimum-jerk cost principle are similar to the trajectories generated by the task-dynamic model described in previous section.

CHAPTER 3

ESTIMATION OF DYNAMIC ARTICULATORY PARAMETERS

3.1 Cubic Spline Method

A major difficulty in generating natural-sounding speech using articulatory synthesis is the lack of sufficient data on the motion of the articulators that control the parameters of the synthesizer. Suppose that we have N discrete phonemes with their static articulatory parameters denoted by $f_{i,j}, i = 1, 2, \cdots, N, \quad j = 1, 2, \cdots, M$, here M denotes the number of articulatory parameters per phoneme ($M = 8$ in our case). Our objective is to estimate the articulatory parameters of the K frames between consecutive phonemes i and $i + 1$. We can solve this problem by obtaining a cubic spline function which interpolates the function f at $x_0, x_1, \cdots, x_N$. In each of the subintervals $I_i = [x_i, x_{i+1}]$ of the interpolation range, s is a polynomial of degree at most three. Denote this polynomial by s_i, then we have:

$$s(x) = s_i(x) \qquad x \in I_i, i = 0, 1, \cdots, N - 1 \tag{3.1}$$

A convenient formulation of s_i will be in terms of the distance of x from the two ends of the interval I_i, and so we can define the new variable u_i by:

$$u_i = x - x_i \qquad i = 0, 1, \cdots, N \tag{3.2}$$

Observe that $\frac{du_i}{dx} = 1$ for every i, and so differentiation or integration with respect to x and with respect to u_i will be equivalent. We denote the step lengths between the knots by $h_i = x_{i+1} - x_i = u_i - u_{i+1}$. The conditions which must be satisfied are that s must interpolate f at $x_0, x_1, \cdots, x_N$ and $\dot{s}$, $\ddot{s}$ must be continuous at the interior knots $x_1, x_2, \cdots, x_{N-1}$. We will begin with the continuity of $\ddot{s}$. On each of the intervals I_i, s is a cubic, and so $\ddot{s}$ is the first-degree polynomial $\ddot{s}_i$. Let's denote its values at the knots by:

$$\ddot{s}(x_i) = A_i, \qquad i = 0, 1, \cdots, N \tag{3.3}$$

Since $\ddot{s}_i$ is a linear function, we have, for each i,

$$\ddot{s}_i(x) = \frac{A_{i+1}(x - x_i) - A_i(x - x_{i+1})}{h_i} = \frac{A_{i+1}u_i - A_i u_{i+1}}{h_i} \tag{3.4}$$

We may integrate Eq. (3.4) twice to get:

$$s_i(x) = \frac{A_{i+1}u_i^3 - A_i u_{i+1}^3}{6h_i} + cx + d \tag{3.5}$$

where c and d are constants of integration. This can be conveniently written in the form:

$$s_i(x) = \frac{A_{i+1}u_i^3 - A_i u_{i+1}^3}{6h_i} - B_i u_{i+1} + C_i u_i \tag{3.6}$$

Now we should choose the coefficients of Eq. (3.6) for $i = 0, 1, \cdots, N-1$ so that both the interpolation conditions and the first-derivative continuity are satisfied. Consider

first the interpolation conditions. At the point x_i, we have $u_i = 0$ and $u_{i+1} = -h_i$. Denoting $f(x_i)$ by f_i and substituting these values into Eq. (3.6), we get:

$$f_i = \frac{A_i h_i^2}{6} + B_i h_i, \qquad i = 0, 1, \cdots, N-1 \tag{3.7}$$

$$f_{i+1} = \frac{A_{i+1} h_i^2}{6} + C_i h_i, \qquad i = 0, 1, \cdots, N-1 \tag{3.8}$$

Solving Eqs. (3.7) - (3.8) for B_i and C_i yields:

$$B_i = \frac{f_i}{h_i} - \frac{A_i h_i}{6} \tag{3.9}$$

$$C_i = \frac{f_{i+1}}{h_i} - \frac{A_{i+1} h_i}{6} \tag{3.10}$$

The final system of equations is derived from the first-derivative continuity condition. These equations are obtained by differentiating Eq. (3.6) with respect to x. We obtain:

$$\dot{s}_i(x) = \frac{A_{i+1} u_i^2 - A_i u_{i+1}^2}{2h_i} - B_i + C_i \tag{3.11}$$

from which we may deduce that:

$$\dot{s}_i(x_i) = C_i - B_i - \frac{A_i h_i}{2} \tag{3.12}$$

The continuity of $\dot{s}$ will be guaranteed if for every interior knot x_i, we have $\dot{s}_i(x_i) = \dot{s}_{i-1}(x_i)$ which yields the equation:

$$\frac{(h_{i-1} + h_i) A_i}{2} + B_i - C_i - (B_{i-1} - C_{i-1}) = 0 \qquad i = 1, 2, \cdots, N_1 \tag{3.13}$$

From Eqs. (3.9) - (3.10), we can get:

$$B_i - C_i = \frac{(A_{i+1} - A_i) h_i}{6} - \frac{f_{i+1} - f_i}{h_i} \qquad i = 0, 1, \cdots, N_1 \tag{3.14}$$

Denote $\frac{f_{i+1} - f_i}{h_i}$ by d_i and substituting Eq. (3.14) into Eq. (3.13) for both i and $i-1$, we get:

$$\frac{h_{i-1} A_{i-1}}{6} + \frac{(h_{i-1} + h_i) A_i}{3} + \frac{h_i A_{i+1}}{6} = d_i - d_{i-1} \qquad i = 1, 2, \cdots, N-1 \tag{3.15}$$

Multiplying Eq. (3.15) by 6 and denoting $d_i - d_{i-1}$ by $\triangle d_{i-1}$, the natural cubic spline interpolating f at $x_0, x_1, \cdots, x_N$ is obtained with the coefficients A_i satisfying the *tridiagonal* system:

$$\begin{bmatrix} 2(h_0+h_1) & h_1 & 0 & \cdots & 0 \\ h_1 & 2(h_1+h_2) & h_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & h_{N-3} & 2(h_{N-3}+h_{N-2}) & h_{N-2} \\ 0 & \cdots & 0 & h_{N-2} & 2(h_{N-2}+h_{N-1}) \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_{N-2} \\ A_{N-1} \end{bmatrix} = \begin{bmatrix} 6\triangle d_0 \\ 6\triangle d_1 \\ \vdots \\ 6\triangle d_{N-3} \\ 6\triangle d_{N-2} \end{bmatrix} \tag{3.16}$$

The system described by Eq. (3.16) is an interpolation system. The shape it assumes is the shape with minimum stored energy, and it turns out to be a natural spline.

3.2 Signal Approximation Method

3.3 Review of the Signal Representation Techniques

Interpolation is one of the basic operations in signal processing. It is used extensively in image reconstruction, magnetic resonance imaging and other signal processing applications. Suppose that we are given an vector space of functions X and a set of samples from a function $x(t) \in X$, the objective of interpolation is to find an element $\hat{x} \in X$ that is the optimal approximation to $x(t)$ according to certain minimum error criterion. The traditional signal processing theory uses the sinc-interpolation for bandlimited functions based on the Shannon's sampling theory which is stated as following. If the highest frequency contained in an analog signal $x_a(t)$ is $F_{max} = B$ and the signal is sampled at a rate $F_s > 2F_{max} = 2B$, then $x - a(t)$ can be exactly recovered from its sample values

using the interpolation function:

$$g(t) = \frac{sin(2\pi Bt)}{2\pi Bt} \tag{3.17}$$

$$x_a(t) = \sum_{n=-\infty}^{+\infty} x(n)g(t - \frac{n}{F_s}) \tag{3.18}$$

where $x(n) = x_a(\frac{n}{F_s})$ are the samples of $x_a(t)$.

In practice, most of the researchers use short kernel methods such as bilinear interpolation, cubic convolution or polynomial spline interpolation, which are much more efficient to implement, especially in higher dimensions. These methods are all convolution-based in the sense that they use an interpolation model of the form:

$$s_h(x) = \sum_{k \in Z} c_h(k)\varphi(\frac{x}{h} - k) \tag{3.19}$$

where h is the sampling step and $\varphi(x)$ is the basic interpolation kernel. The expansion coefficients in Eq. (3.19) typically correspond to the samples of the input function $s(x)$ taken on a uniform grid: $c_h(k) = s(hk)$ Fig. 3.1 shows the block diagram of the convolutional-based interpolator. Where the impulse response of the reconstruction filter is $\varphi_h(x) = \varphi(\frac{x}{h})$.

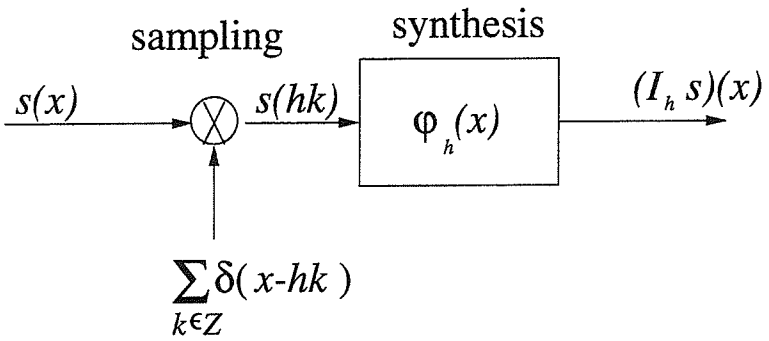


Figure 3.1 Block diagram of the convolutional-based interpolator

More recently, researchers have proposed a systematic formulation of this class of representations using the Hilbert space framework [130]. The corresponding least square (LS) solution can be obtained through a simple modification of the basic interpolation

procedure, which consists of applying an appropriate prefilter to $s(x)$ prior to sampling as shown in Fig. 3.2. Where the signal approximation $P_h s$ corresponds to the orthogonal projection of s onto the signal subspace $V_h = span\{\varphi(\frac{x}{h} - k)\}_{k \in Z}$. The impulse response of the optimal prefilter is $h^{-1} \cdot \tilde{\varphi}(-\frac{x}{h})$.

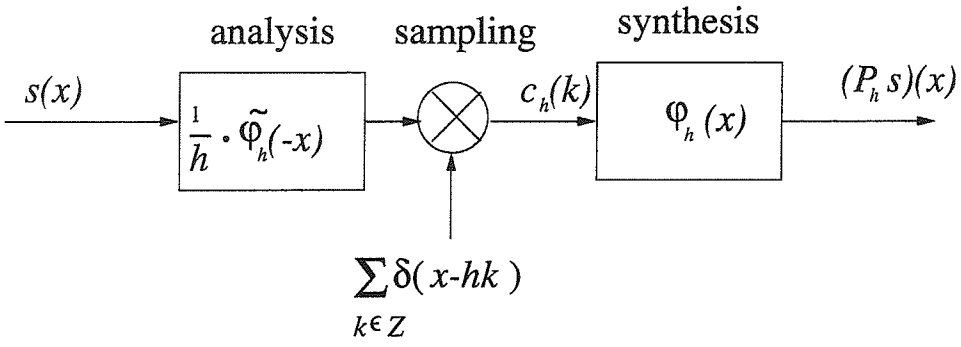


Figure 3.2 Block diagram of the convolutional-based least squares signal approximation.

The non-bandlimited convolution-based signal representations have a lot of advantages for signal processing such as computational efficiency and simplicity of implementation. In [130], the authors answered two fundamental question in this signal representation. First, they provided quantitative error estimates that can be used for the appropriate selection of the sampling step h. Secondly, they compared three signal approximation methods namely least squares approximation, interpolation and quasi-interpolation and derived some quantitative error bounds on these methods. Before the error bounds analysis of these methods, we will first introduce some notations.

3.4 Notations

3.4.1 L_2 Space

L_2 is the space of measurable, square-integrable, real-valued functions or signals $s(x), x \in R$. It is a Hilbert space whose L_2-norm is induced from the inner product:

$$< s(x), r(x) >= \int_{-\infty}^{+\infty} s(x)r(x)dx = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \overline{S(\omega)}R(\omega)d\omega \tag{3.20}$$

where $S(\omega)$ and $R(\omega)$ denote the Fourier transforms of $s(x)$ and $r(x)$, respectively. The L_∞ is defined as:

$$\|s\|_\infty = \lim_{p\to+\infty}[\int_{-\infty}^{+\infty} |s(x)|^p]^{\frac{1}{p}} = \sup_{x\in R}|s(x)| \tag{3.21}$$

The class of smoothness of a signal will be specified by its distance to the Sobolev space W_2^L (or W_∞^L), which is the space of functions whose L first derivatives are defined in the L_2 (L_∞) sense.

3.4.2 Convolution-based Signal Representations

A general approach to specify continuous signal representation is to consider the class of functions generated from the integer translates of a single function $\varphi(x) \in L_2$. We can adjust the resolution by varying the sampling step h and rescaling φ accordingly. The corresponding function space $V_h(\varphi) \subset L_2$ is defined as:

$$V_h(\varphi) = \{s_h(x) = \sum_{k\in Z} c_h(k)\varphi(\frac{x}{h} - k)|c_h \in l_2\} \tag{3.22}$$

where l_2 is the vector space of square-summable sequences. The only restriction on the choice of the *generation function* φ is that the set $\varphi(x/h - k)_{k\in Z}$ is a Riesz basis of $V_h(\varphi)$. This is equivalent to the condition:

$$0 < A \leq a_\varphi(\omega) = \sum_{k\in Z} |\Phi(\omega + 2\pi k)|^2 \leq B < +\infty \quad a.e. \tag{3.23}$$

where $\Phi(\omega)$ is the Fourier transform of $\varphi(x)$, and the constants A and B are the Riesz bounds. This constraint ensures that each function $s_h(x)$ in $V_h(\varphi)$ is uniquely characterized by the sequence of its coefficients $c_h(k)$.

3.4.3 Interpolation and Quasi-Interpolation

The simplest way to represent a continuous signal $s(x) \in L_2$ in $V_h(\varphi)$ is to use its samples as the coefficients of the representation in Eq. (3.22). The corresponding *interpolation* operator, which is schematically represented by the block diagram in Fig. 3.1 is defined as:

$$(I_h s)(x) = \sum_{k \in Z} s(hk)\varphi(\frac{x}{h} - k) \tag{3.24}$$

The operator I_h is bounded, provided that the input signal is sufficiently smooth; for example, $s \in W_2^1$. Note that with this definition, the samples of the signal $s(x)$ and of its interpolation $(I_h s)(x)$ are not necessarily identical. To get a true interpolation (i.e., $\forall k \in Z, s(x)|_{x=hk} = (I_h s)(x)|_{x=hk}$), we need to select a generating function $\varphi_{int} \in V_1(\varphi)$ that satisfies the interpolation property:

$$\varphi_{int}(x)|_{x=k} = \delta(k) \tag{3.25}$$

where $\delta(k)$ denotes the discrete unit impulse at the origin. For a given subspace $V_h(\phi) \subset L_2$, the interpolation function $\varphi_{int}(\frac{x}{h}) \in V_h(\varphi)$ is generally unique.

Furthermore, we can relax the interpolation condition without any noticeable loss in performance. This leads to the concept of a $quasi - interpolation$, which is a standard notion in approximation theory, but has not yet been exploited in signal processing. By definition, a quasi-interpolant of order $L = n + 1$ is a function φ_{QI} that interpolates all polynomial $p_n(x)$ of degree n:

$$\forall p_n(x) \in P^n, \quad \sum_{k \in Z} p_n(k)\varphi_{QI}(x - k) = p_n(x) \tag{3.26}$$

where P^n denotes the space of polynomials of degree n. The equivalent formulation of this condition in the frequency domain is:

$$\Phi_{QI}(\omega)|_{\omega=2\pi k} = \delta(k) \tag{3.27}$$

$$\Phi_{QI}^{(m)}(\omega)|_{\omega=2\pi k} = 0, \quad (m = 1, \cdots, L-1) \tag{3.28}$$

where $\Phi_{QI}(\omega)$ denotes the Fourier transform of φ_{QI} and $\Phi_{QI}^{(m)}(\omega)$ denotes the m^{th} derivative of $\Phi_{QI}(\omega)$. In other words, $\Phi_{QI}(\xi) = 1 + O(\xi^L)$ as $\xi \to 0$. Whether or not it is possible to construct quasi-interpolants within a certain subspace $V_h(\varphi)$ depends on its order of approximation, which will be described in the Strang-Fix conditions.

3.4.4 Convolution-Based Least Squares

A more sophisticated approach for obtaining a representation of the signal $s(x) \in L_2$ in $V_h(\varphi)$ is to determine its minimum L_2-norm approximation (orthogonal projection). This least squares approximation is given by:

$$(P_h s)(x) = \sum_{k \in Z} c(k)\varphi(\frac{x}{h} - k) \tag{3.29}$$

$$c(k) = \frac{1}{h} < s(x), \tilde{\varphi}(\frac{x}{h} - k) > \tag{3.30}$$

where $\tilde{\varphi}$ is the dual of φ and is defined by:

$$\tilde{\Phi}(\omega) = \frac{\Phi(\omega)}{a_\varphi(\omega)} \tag{3.31}$$

$$a_\varphi(\omega) = \sum_{n \in Z} |\Phi(\omega + 2n * \pi)|^2 \tag{3.32}$$

where $\Phi(\omega)$ and $\tilde{\Phi}(\omega)$ denotes the Fourier transform of $\varphi(x)$ and $\tilde{\varphi}(x)$, respectively.

The orthogonal projection operator on $V_h(\varphi)$ can also be expressed in the more compact form:

$$(P_h s)(x) = \int_{-\infty}^{+\infty} s(y)\frac{1}{h}K(\frac{x}{h}, \frac{y}{h})dy \tag{3.33}$$

where $K(x,y)$ is the reproducing kernel associated with the basic approximation space $V_1(\varphi)$:

$$K(x,y) = \sum_{k \in Z} \varphi(x-k)\tilde{\varphi}(y-k) \tag{3.34}$$

The only difference between the (quasi-)interpolation procedure and the least squares approach is the presence of the prefiltering module, which has a role similar to the anti-aliasing filter required in conventional sampling theory. In fact, if $\varphi(x) = sinc(x)$, then the optimal filter is precisely shannon's ideal lowpass filter with the appropriate cutoff at the Nyquist frequency.

3.4.5 Strang-Fix Conditions

The Strang-Fix conditions relate the approximation power of the representation to the spectral characteristics of the generating function and are described as following.

Let φ be a valid generating function with appropriate decay. The following statements are equivalent [116]:

(1). The function spaces $V_h(\varphi)$ reproduce polynomials of degree $n = L-1$, which is equivalent to say that there exists a function $\varphi_{QI} \in V_1(\varphi)$ (not necessarily unique) that is a quasi-interpolant of order L.

(2). There exists a function $\varphi_{QI} \in V_1(\varphi)$ such that:

$$\forall x \in R, \quad \sum_{k \in Z} \varphi_{QI}(x-k) = 1 \tag{3.35}$$

$$\forall x \in R, \quad \sum_{k \in Z} (x-k)^m \varphi_{QI}(x-k) = 0, \quad (m = 1, \cdots, L-1) \tag{3.36}$$

(3). $\Phi(\omega)$, which is the Fourier transform of φ, is nonvanishing at the origin and has zeros of at least multiplicity L at all nonzero frequencies that are integer multiples of 2π.

(4). There exists a constant C such that approximation error at step size h is bounded as:

$$\forall s \in W_2^L, \quad inf_{s_h \in V_h(\varphi)} \|s - s_h\|_2 \leq C \cdot h^L \cdot \|s^{(L)}\|_2 \tag{3.37}$$

3.5 Pointwise Error Analysis

Although the Strang-Fix bound in Eq. (3.37) is of considerable theoretical interest, it needs to be made more specific and quantitative to be of direct use for signal processing. In this section, we will describe the pointwise error analysis of different signal approximation methods. The proposition 1 is proved in [130]. We proved an improved error bound for interpolation in proposition 2 and an error bound for LS approximation in proposition 3. The basic tool for this pointwise analysis is the Taylor series expansion. Specifically, if the signal s is $(n+1)$ times continuously differentiable (i.e. $s \in W_\infty^{n+1}$), we can write:

$$s(y) = s(x) + (y-x)s^{(1)}(x) + \frac{(y-x)^2}{2!}s^{(2)}(x) + \cdots + \frac{(y-x)^n}{n!}s^{(n)}(x) + R_{n+1}(y) \tag{3.38}$$

$$R_{n+1}(y) = \frac{(y-x)^{n+1}}{n!} \cdot \int_0^1 (1-\tau)^n s^{(n+1)}(\tau y + (1-\tau)x) d\tau \tag{3.39}$$

3.5.1 Interpolation Error

Let's consider the (quasi-)interpolation error defined by;

$$s(x) - (I_h s)(x) = s(x) - \sum_{k \in Z} s(hk)\varphi(\frac{x}{h} - k) \tag{3.40}$$

Replacing $s(hk)$ by its Taylor series in Eq. (3.38) with $y = hk$ and using the quasi-interpolation properties of φ, we get:

$$s(x) - (I_h s)(x) = \frac{-h^L}{(L-1)!} \sum_{k \in Z} (k - \frac{x}{h})^L \varphi(\frac{x}{h} - k) \cdot [\int_0^1 (1-\tau)^{L-1} s^{(L)}(\tau hk + (1-\tau)x) d\tau] \tag{3.41}$$

The only remaining terms are those associated with the remainders of the Taylor series because φ is designed to perfectly interpolate all expansion terms up to degree $n = L-1$. The uniform estimate of the error is given in the following proposition.

Proposition 1: If φ is a quasi-interpolant of order L with sufficient decay, then:

$$\forall s \in W_\infty^L, \quad \|s - (I_h s)\|_\infty \leq C_{\varphi,L} \cdot h^L \cdot \|s^{(L)}\|_\infty \tag{3.42}$$

$$C_{\varphi,L} = \frac{1}{L!} \sup_{x\in[0,1]} \sum_{k\in Z} |x-k|^L |\varphi(x-k)| \tag{3.43}$$

Proof: See [130].

If $\varphi(x)$ decays like $O(x^{-(L+1)})$, we can improve our estimate by considering one more term in the Taylor series expansion. In this section, we will give the detailed proof of the following error bound.

Proposition 2: If φ decays like $O(x^{-(L+1)})$, then the pointwise estimate for $s \in W_\infty^{L+1}$ is given by:

$$s(x) - (I_h s)(x) = -\frac{h^L}{L!} E_L(\frac{x}{h}) s^{(L)}(x) + O(h^{L+1}) \tag{3.44}$$

$$E_L(x) = (-1)^L \sum_{k\in Z} (x-k)^L \varphi(x-k) \tag{3.45}$$

Proof: Let $y = hk$ and substitute it into Eq. (3.38), we get:

$$s(hk) = s(x) + \sum_{i=1}^{L-1} \frac{(hk-x)^i}{i!} s^{(i)}(x) + \frac{(hk-x)^L}{L!} s^{(L)}(x) + R_{L+1}(hk) \tag{3.46}$$

$$\begin{aligned}
(I_h s)(x) &= \sum_{k\in Z} s(hk) \varphi(\frac{x}{h} - k) \\
&= \sum_{k\in Z} s(x) \varphi(\frac{x}{h} - k) + \sum_{i=1}^{L-1} \sum_{k\in Z} \frac{(hk-x)^i}{i!} s^{(i)}(x) \varphi(\frac{x}{h} - k) \\
&+ \frac{h^L}{L!} E_L(\frac{x}{h}) s^{(L)}(x) + \sum_{k\in Z} R_{L+1}(hk) \varphi(\frac{x}{h} - k)
\end{aligned} \tag{3.47}$$

where $E_L(x)$ is defined in Eq. (3.45). From the Strang-Fix condition in Eqs. (3.35) - (3.36), we get:

$$\sum_{k\in Z} s(x) \varphi(\frac{x}{h} - k) = s(x) \tag{3.48}$$

$$\sum_{k\in Z}\frac{(hk-x)^i}{i!}s^{(i)}(x)\varphi(\frac{x}{h}-k)=0,\quad (i=1,\cdots,L-1) \tag{3.49}$$

Then we can get:

$$s(x)-(I_hs)(x)+\frac{h^L}{L!}E_L(\frac{x}{h})s^{(L)}(x)=\sum_{k\in Z}R_{L+1}(hk)\varphi(\frac{x}{h}-k)$$
$$=-\frac{(h)^{L+1}}{L!}\sum_{k\in Z}(k-\frac{x}{h})^{L+1}\varphi(\frac{x}{h}-k)\cdot\int_0^1(1-\tau)^Ls^{(L+1)}(\tau hk+(1-\tau)x)d\tau \tag{3.50}$$

$$|s(x)-(I_hs)(x)+\frac{h^L}{L!}E_L(\frac{x}{h})s^{(L)}(x)|\le\frac{h^{L+1}}{L!}\sum_{k\in Z}|\frac{x}{h}-k|^{L+1}|\varphi(\frac{x}{h}-k)|\cdot\int_0^1(1-\tau)^Lsup_{x\in R}|s^{(L+1)}(x)|d\tau \tag{3.51}$$

$$|s(x)-(I_hs)(x)+\frac{h^L}{L!}E_L(\frac{x}{h})s^{(L)}(x)|\le h^{L+1}\|s^{(L+1)}\|_\infty\frac{1}{(L+1)!}\sum_{k\in Z}|\frac{x}{h}-k|^{L+1}|\varphi(\frac{x}{h}-k)|$$
$$=C_{\varphi,L+1}\cdot h^{L+1}\cdot\|s^{(L+1)}\|_\infty=O(h^{L+1}) \tag{3.52}$$

This lead to the pointwise estimate of s in Eq. (3.44).

3.5.2 Least Squares Error

To simplify the analysis of error in the least squares case, we will use the reproducing kernel formalism. For the generating function φ which satisfies the decay condition $|\varphi(x)|\le K\cdot(1+|x|)^{-M}\quad M\ge L$, an equivalent form of the conditions Eqs. (3.35) - (3.36) is as following [130]:

$$e_0(x)=\int_{-\infty}^{+\infty}K(x,y)dy=1 \tag{3.53}$$

$$e_m(x)=\int_{-\infty}^{+\infty}(y-x)^mK(x,y)dy=0,\quad m=1,\cdots,L-1 \tag{3.54}$$

where the reproducing kernel $K(x,y)$ is defined in Eq. (3.34). The corresponding pointwise error bound of the least squares is given in the following proposition.

Proposition 3: If φ is such that the conditions in Eqs. (3.53) - (3.54) are satisfied, then:

$$\forall s\in W_\infty^L,\quad \|s-(P_hs)\|_\infty\le C_{K,L}\cdot h^L\cdot\|s^{(L)}\|_\infty \tag{3.55}$$

$$C_{K,L} = \frac{1}{L!} \sup_{x \in R} [\int_{-\infty}^{+\infty} |x-y|^L |K(x,y)| dy] \tag{3.56}$$

Proof: Using Eq. (3.33) and Eq. (3.53), we can write the approximation error as:

$$s(x) - (P_h s)(x) = \int_{-\infty}^{+\infty} [s(x) - s(y)] \frac{1}{h} K(\frac{x}{h}, \frac{y}{h}) dy \tag{3.57}$$

Substituting the Taylor series Eq. (3.38) into Eq. (3.57) and noting from Eq. (3.56) that the integers of $K(x,y)$ times difference monomials $(y-x)^m$ are zero for $m = 1, \cdots, L-1$, we get:

$$s(x) - (P_h s)(x) = -\int_{-\infty}^{+\infty} R_L(y) \frac{1}{h} K(\frac{x}{h}, \frac{y}{h}) dy \tag{3.58}$$

The remainder can also be written in the standard form:

$$R_L(y) = \frac{(y-x)^L}{L!} s^{(L)}(\xi) \tag{3.59}$$

where ξ is some value between x and y. This lead to the estimate:

$$|s(x) - (P_h s)(x)| \leq \frac{h^L}{L!} \cdot \|s^{(L)}\|_\infty \cdot \int_{-\infty}^{+\infty} |\frac{y}{h} - \frac{x}{h}|^L \frac{1}{h} |K(\frac{x}{h}, \frac{y}{h})| dy \tag{3.60}$$

We then make the change of variable $y' = \frac{y}{h}$ and take the supremum at both sides of Eq. (3.60) and get the error bound given in Eq. (3.55).

The pointwise error analysis indicates that the local behavior of the error is qualitatively the same in the (quasi-)interpolation and least squares cases. Comparing Eqs. (3.42) and (3.55) we can see that the only difference lies in the constants involved. In [130], the authors claimed that the constant $C_{K,L}$ for the least square case is smaller than the constant $C_{\varphi,L}$ for the (quasi-)interpolator case, but they didn't prove this claim.

3.6 L_2 Error Analysis

It's usually more informative to investigate the behavior of the L_2 error for quantification purposes. The most appropriate tool for this type of analysis is the Fourier transform. Before the error bound analysis, we first prove a useful lemma [130].

Lemma 1: If F is M times continuously differentiable and $F^{(m)}(2\pi k) = 0$ for all $k \in Z,\ k \neq 0$ and $m = 0, \cdots, M-1$, and $F^{(M)}$ decays fast enough so that $\sum_k |F^{(M)}(\xi + 2\pi k)| \leq C < +\infty$, then,

$$|\sum_{k\neq 0} F(\omega + 2\pi k)| \leq \frac{|\omega|^M}{M!} \sup_{|\xi|\leq\pi} |\sum_{k\neq 0} F^{(M)}(\xi + 2\pi k)|, \quad \forall |\omega| \leq \pi \tag{3.61}$$

3.6.1 L_2 Error of Quasi-Interpolation

We can write down Eq. (3.24) in the frequency domain:

$$\hat{(I_h s)}(\omega) = \sum_{k\in Z} S(\omega + \frac{2\pi}{h}k)\Phi(h\omega) \tag{3.62}$$

where $S(\omega)$, $\Phi(\omega)$ and $\hat{(I_h s)}(\omega)$ denotes the Fourier transform of $s(x)$, $\varphi(x)$ and $(I_h s)(x)$, respectively. This lead to the following error decomposition in the frequency domain.

$$S(\omega) - \hat{(I_h s)}(\omega) = [S(\omega)(1 - \Phi(h\omega))] - [\sum_{k\neq 0} S(\omega + \frac{2\pi}{h}k)\Phi(h\omega)] = E_1(\omega) + E_2(\omega) \tag{3.63}$$

where the $E_1(\omega)$ and $E_2(\omega)$ denotes the in-band and out-of-band error components, respectively. The L_2 error bound of quasi-interpolation is given by the following proposition.

Proposition 4: If φ is a quasi-interpolant of order L with sufficient decay, then:

$$\forall s \in W_2^L, \|s - I_h s\|_2 \leq C_{\varphi,L} \cdot h^L \cdot \|s^{(L)}\|_2 \tag{3.64}$$

where $C_{\varphi,L}$ is defined in Eq. (3.43).

The key idea of the proof of *Proposition 4* is to use the error decomposition in Eq. (3.63) and use the Taylor series expansion of $1 - \Phi(h\omega)$ and the Schwartz inequality. The detailed proof was described in [130].

3.6.2 L_2 Error of the LS Approximation

The Fourier representation of the LS approximation of Eq. (3.29) is given by:

$$\begin{aligned}(\hat{P_h s})(\omega) &= \Phi(h\omega)\sum_{k\in Z}\overline{\tilde{\Phi}(h\omega+2\pi k)}S(\omega+\frac{2\pi}{h}k)\\ &= \Phi(h\omega)\sum_{k\in Z}\frac{\overline{\Phi(h\omega+2\pi k)}}{a_\varphi(h\omega)}\cdot S(\omega+\frac{2\pi}{h}k)\end{aligned} \tag{3.65}$$

where $a_\varphi(\omega)$ is defined in Eq. (3.32). The approximation error in the Fourier domain can be decomposed into two components.

$$S(\omega)-(\hat{P_h s})(\omega) = e_1(\omega)+e_2(\omega) \tag{3.66}$$

$$e_1(\omega) = [1-\frac{|\Phi(h\omega)|^2}{a_\varphi(h\omega)}]S(\omega) \tag{3.67}$$

$$e_2(\omega) = \sum_{k\neq 0}\frac{\Phi(h\omega)}{a_\varphi(h\omega)}\cdot\overline{\Phi(h\omega+2\pi k)}\cdot S(\omega+\frac{2\pi}{h}k) \tag{3.68}$$

where $e_1(\omega)$ and $e_2(\omega)$ denotes the in-band error and out-of-band error, respectively. The L_2 error bound of the LS approximation is given by the following proposition.

Proposition 5: If $\Phi^{(m)}(2\pi k)=0,\ \ k\in Z,\ \ k\neq 0$ and $m=0,\cdots,L-1$, then:

$$\forall s\in W_2^L, \|s-P_h s\|_2 \leq K_{\varphi,2L}\cdot h^{2L}\cdot\|s^{(2L)}\|_2 + K_{\varphi,2L}^{1/2}\cdot h^L\cdot\|s^{(L)}\|_2 \tag{3.69}$$

$$K_{\varphi,2L} = \frac{1}{(2L)!}\cdot\frac{1}{A}\cdot\sup_\xi|\sum_{k\neq 0}(|\Phi|^2)^{(2L)}(\xi+2\pi k)| \tag{3.70}$$

where $A=inf_\omega[a_\varphi(\omega)]$ is the lower Riesz bound.

The key idea of the proof of *Proposition 5* is to use the error decomposition in Eq. (3.66) and the Taylor expansion in *Lemma 1* and the Schwartz inequality. The detailed proof of this proposition is given in [130].

3.6.3 Comparison

The L_2 bounds in Eqs. (3.64) and (3.69) are both consistent with the Strang-Fix conditions in Eq. (3.37). However, the error bound for the LS case provides a finer

characterization of the error. It consists of two distinct terms that represent the in-band (e_1) and out-of-band (e_2) contribution of the error, respectively. For small value of h, the first part of the error becomes neglible and the bound is dominated by the second $O(h^L)$ term. The corresponding constant $K_{\varphi,2L}^{1/2}$ turns out to be smaller than the constant C in the Strang-Fix bound in Eq. (3.37)., or $C_{\varphi,L}$ in Eq. (3.64). This is a first indication that there is an advantage in using least squares over interpolation. In addition, we note that in the case of larger values of h, the first term in the LS error bounds becomes dominant and it has the characteristic form of the error of an interpolator of order $2L$. We further observe that the function $\Phi_{2L}(\omega) = \frac{|\Phi(\omega)|^2}{a_\varphi(\omega)}$ represents the frequency response of an interpolator of order $2L$, Therefore under the condition that $\|e_1\| \gg \|e_2\|$, the LS solution of order L should perform as well as the corresponding interpolator with twice the order. This condition typically arises for larger h when the signal is somewhat undersampled. In the following section, we will apply this observation to a specific task where the observed discrete signal is undersampled.

3.7 Experimental Results

Suppose we are given N discrete phonemes with their static articulatory parameters denoted by $s_m(n), \quad m = 1, 2, \cdots M, \quad n = 1, 2, \cdots, N$, where M denotes the number of articulatory parameters per phoneme ($M = 8$ in our case). Our objective is to estimate the dynamic articulatory parameters between consecutive phonemes. This problem can be restated as given N discrete observations $s_m(n)$ of M continuous function $s_m(x)$, we want to construct the functions $\hat{s}_m(x)$ which is an "optimal" approximation to $s_m(x)$ according to certain criterion. Due to the physical constraints of human vocal apparatus, all of its parameters such as tongue body height have finite values and its movements are continuous over time and their L^{th} derivatives also have finite value ($L \geq 2$). Thus

the functions $s_m(x) \in W_2^L(W_\infty^L)$ and we can use the interpolation or LS approximation discussed above to slove the dynamic parameter estimation problem. In this section, we investigated the cubic spline interpolation and the LS approximation which is realized by an "optimal" prefiltering followed by cubic spline interpolation.

The LS approximation in Eq. (3.29) can be implemented by first prefiltering the discrete data $s_m(n) = s_m(x)|_{x=nh}$ by an "optimal" filter $P(z)$, which yields the sequence $c(k)$ in Eq. (3.29), and then by applying the interpolation in Eq. (3.24). In our experiment, the Z-transform of the "optimal" filter for cubic spline interpolation φ is given by:

$$P(z) = \frac{3}{16} \frac{(z^4 + z^{-4}) + a_3(z^3 + z^{-3}) + a_2(z^2 + z^{-2}) + a_1(z + z^{-1}) + a_0}{(z^5 + z^{-5}) + b_4(z^4 + z^{-4}) + b_3(z^3 + z^{-3}) + b_2(z^2 + z^{-2}) + b_1(z + z^{-1}) + b_0} \tag{3.71}$$

Where $a_3 = 6552$, $a_2 = 331612$, $a_1 = 2485288$, $a_0 = 4675014$, $b_4 = 196$, $b_3 = 10541$, $b_2 = 120608$, $b_1 = 467858$, and $b_0 = 736952$, respectively.

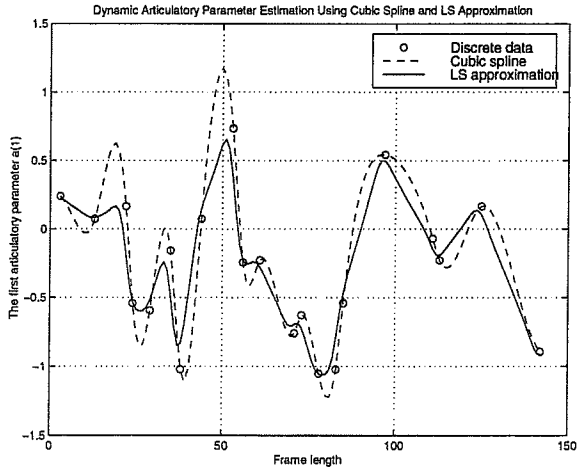


Figure 3.3 Estimated first articulatory parameter using cubic spline interpolation and LS approximation

Figures (3.3) - (3.6) shows the estimated dynamic articulatory parameters (first articulatory parameter to the fourth articulatory parameters) of English sentence "Where were you while we were away" using the approximation methods discussed above. For

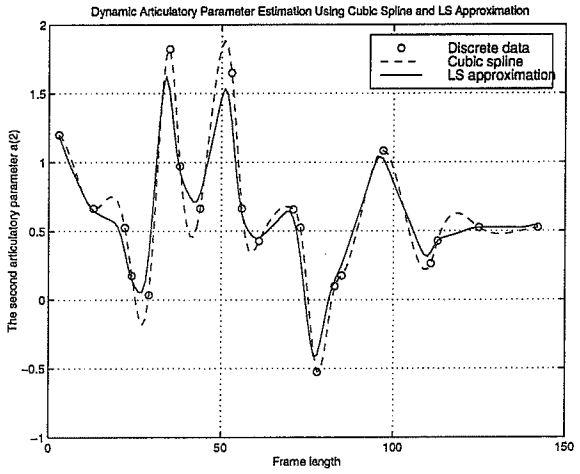


Figure 3.4 Estimated second articulatory parameter using cubic spline interpolation and LS approximation

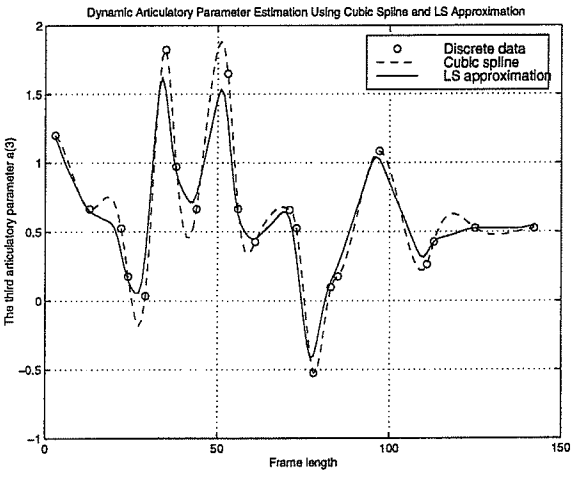


Figure 3.5 Estimated third articulatory parameter using cubic spline interpolation and LS approximation

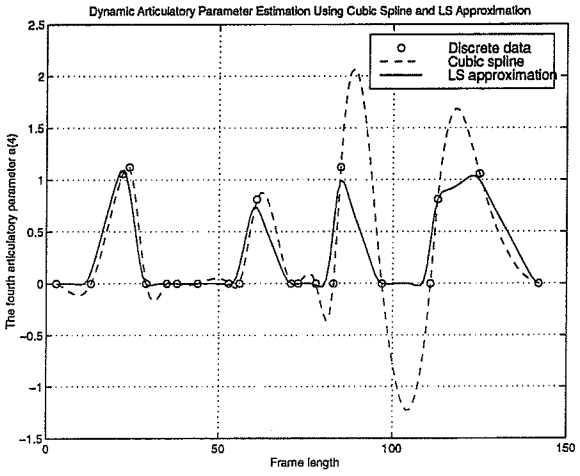


Figure 3.6 Estimated fourth articulatory parameter using cubic spline interpolation and LS approximation

each figure, the circle points denotes the static articulatory parameters computed from the method in [64], for example, the first circle in Fig. (3.3) denotes the first articulatory parameter (anterior/posterior position of the tongue body) of the first phoneme /W/; the dashed line denotes the approximated continuous articulatory parameters using the cubic spline interpolation method; and the solid line denotes the estimated continuous articulatory parameters using the LS approximation approach. We can see from these figures the LS approach achieves smoother approximation of the articulatory parameter trajectory over time ($s_m(t)$ and doesn't have the "overshooting" problems of the cubic spline methods. One hypothesis in human speech production is the "minimum jerk" hypothesis which basically says that articulation is planned in order to minimize the abrupt movements of neuro-muscular actions. We can see from these figures that the LS solution tends to be consistent with this hypothesis because its trajectory doesn't have the abrupt changes which exist in some regions of the cubic spline solutions.

Fig. 3.7 shows the snapshots of the estimated vocal tract movements of phoneme /EH/ in /W/-/EH/-/R/ ("were") using the LS approximation technique described above. We can see from this figure that the LS approximation method can successfully estimate the articulatory trajectory and thus the dynamic vocal tract movements. Although the vocal tract shape of the phoneme /EH/ is unchanged with time when produced in isolation, we can see from this figure that the vocal tract geometry of /EH/ will change dynamically with respect to its neighbor phonemes and thus result in the change of its acoustic properties at different time frames, which is called *coarticulation* in phonetics. Once we can get a good estimate of articulatory parameter trajectories thus the dynamic vocal tract movements, we can apply these dynamic geometries as the boundary conditions of solutions of the Navier-Stokes equations and then synthesize high intelligibility, naturally sounded continuous speech.

3.8 Discussion

Signal approximation in the framework of Hilbert space [130] provides a powerful tool in signal processing because it is unbandlimited, more efficient to implement in high dimensions and has more quantitative error bounds than the conventional interpolation technique based on Shannon's sampling theory. The pointwise error analysis indicates that the (quasi-)interpolation and LS approximation are qualitatively the same. The only difference is the constant terms which tend to be smaller for the LS case. For small values of the sampling step h, the in-band component of the L_2 error of LS approach becomes negligible and the bound is dominated by the out-of-band error term ($O(h^L)$) and the LS error still tends to be smaller than the interpolation error in this case. For larger values of h (undersampling), the first term in the LS error bounds becomes dominant and it has the characteristic form of the error of an interpolator of order $2L$. In

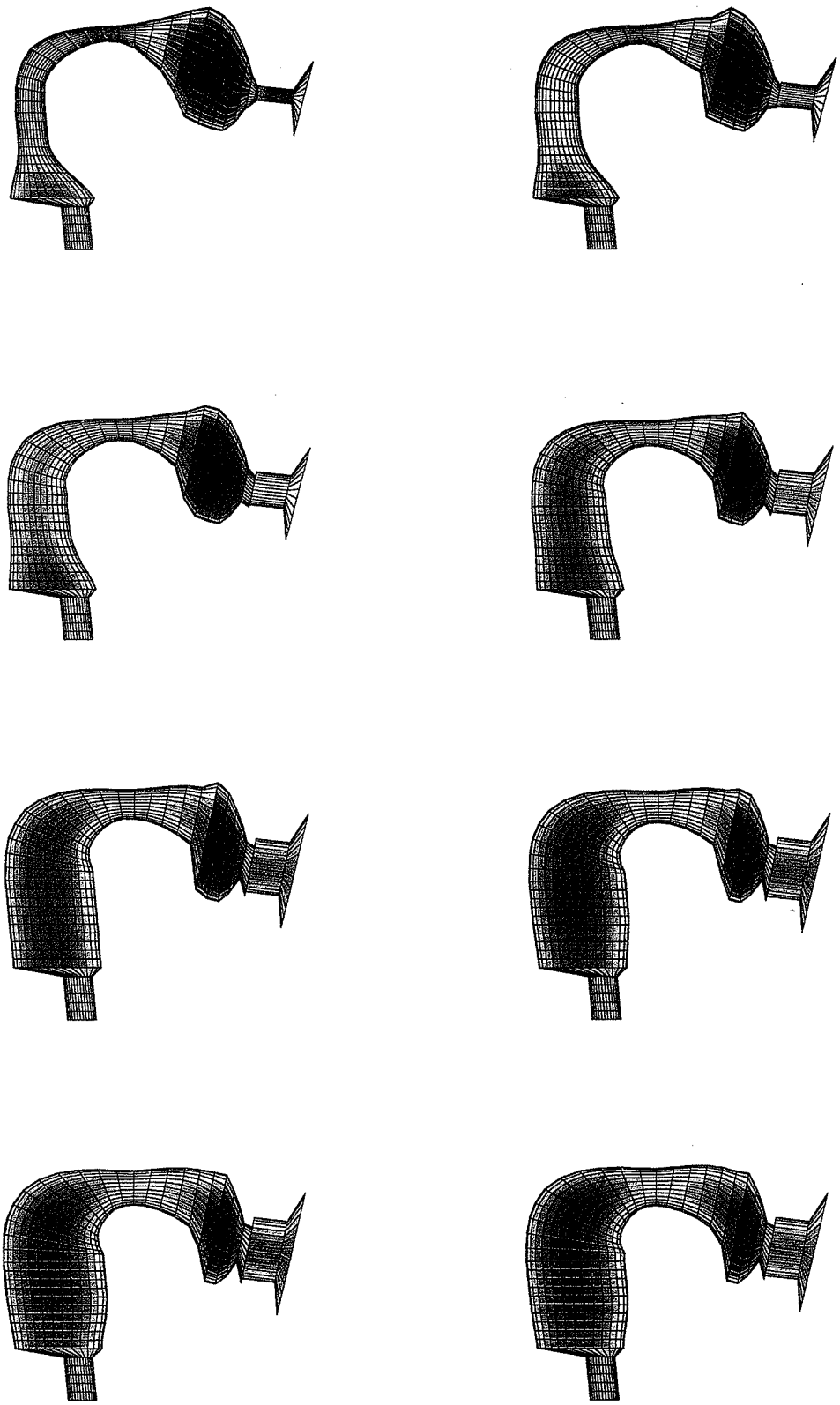


Figure 3.7 Snapshots of the vocal tract movements of phoneme /EH/ in /W/-/EH/-/R/.

both cases, the LS approach has a smaller error bound than the interpolation approach although the difference becomes smaller as h goes smaller . Both the interpolation and LS approximation approach can be successfully applied to estimate the continuous articulatory parameter trajectory given the static parameters of the discrete phonemes. Experimental results indicate that the LS approach achieves smoother estimation and doesn't have the "over-shooting" problem in the spline interpolation methods.

CHAPTER 4

CONSTRUCTION OF ARTICULATORY MODEL BASED ON MRI DATA

4.1 Problem Formulation

A qualitative physical picture of speech sound production has developed over a long time. Fant and Flanagan [26], [31] first noted the necessity of turbulent airflow for producing unvoiced and voiced sounds. Stevens [113] incorporated many ideas from the aeroacoustics literature and concluded that the sound is produced not where the turbulent flow is formed, but instead where that turbulent flow interacts with an obstacle such as teeth in the vocal apparatus. Shadle [102] strongly suggested that the sound produced by airflow in the vocal tract is sensitive to the three-dimensional details of vocal tract geometry, so that a simple axial area distribution may not be enough to characterize the vocal tract for the purpose of producing speech sounds.

Traditional articulatory models use $7 - 15$ physical parameters to model the mid-sagittal section of the human vocal tract and use certain error minimization criterion to estimate the articulatory parameters from the measured acoustical parameters (usu-

ally called the inverse problem) [14], [67], [71], [103], [95], [112]. The disadvantage of this type of articulatory models is that they are basically 2-D models and it's difficult to overcome the ambiguity problem because there is a non-denumerable infinity of area functions corresponding to a given set of acoustical parameters (such as formant frequencies). Furthermore, these articulatory models can't get fine details of the sharp constriction geometry of the 3-D vocal tract which is important for the analysis of airflow inside the vocal apparatus. An alternative is to directly measure the vocal tract shape using the X-ray or magnetic resonance imaging (MRI) techniques. The disadvantage of this approach is that it will produce a very high dimensional signal space. For example, a typical MRI image of an English phoneme contains thousand of data points and thus make it impossible for direct analysis. However, a number of studies suggest that articulatory movements in speech can be approximately modeled using a few elementary articulatory parameters [75], [114], [38]. In this chapter, we propose a latent variable model-based method to reduce the dimensionality of the MRI vocal tract shape data. Our objective is to construct an articulatory speech model with a few parameters which still can represent the fine details of the 3-D vocal tract shape for different phonemes.

4.2 Review of dimensionality reduction methods

Several techniques to extract the model parameters or other kind of condensed information from high dimensional data have been developed. One type of these techniques is based on linear combination of the data vector components. We refer to such linear combination as *linear indices* and represent them using a D-dimensional basis vector $\vec{v}$. The score of a D-dimensional observation data $\vec{t}$ with respect to index $\vec{v}$ is thus the projection of $\vec{t}$ on $\vec{v}$, i.e., $\vec{v}^T\vec{t} = \sum_{i=1}^{D} v_i t_i$. For example, Nguyen [77] applied the discrete cosine transform (DCT) to the electropalatographic (EPG) frames considered as images

to obtain several linear indices, corresponding to low spatial frequencies. The disadvantage of this method is that the set of indices is fixed across data sets (either using DCT or other linear mapping) and is usually insufficient to capture the inter-speaker and intra-speaker differences. Another type of methods use some linear approach to extract structure from a data set without requiring any *a priori* knowledge about it, such as principal component analysis (PCA), rotated principle component analysis (RPCA) and factor analysis (FA) [38], [125]. Other researchers use some nonlinear methods such as neural network (NN) to find the relationship between the articulatory model parameters and the observed data sequence [76], [47]. Usually the NN approach will infer a set of basis vectors that spans a subspace similar to that defined by the PCA [4], [10]. However, these methods lack a natural interpretation as probability model. In this work, we are concerned with methods that explicitly define a full probability model.

4.3 Latent Variable Models

The assumption of a latent variable model is that the observed high-dimensional data is generated from an underlying low-dimensional process. The high dimensionality arises from several reasons, including stochastic variation and the measurement process [111]. The objective is to learn the low dimensional generating process (defined by a small number of *latent variables* or *hidden causes*) along with a noise model, rather than directly learning a dimensionality reduction mapping.

4.3.1 Definition of general latent variable model

Consider an observed sample in data space $\{\vec{t}_n\}_{n=1}^N \subset \Re^D$ of N D-dimensional real vectors that has been generated by an unknown distribution. In latent variable modeling

we assume that the distribution in data space $\Re^D$ is actually due to a small number $L < D$ of latent variables acting in combination. We refer to this L-dimensional space as the *latent space*. Thus, a point $\vec{x}$ in latent space $\Re^L$ is generated according to a prior distribution $p(\vec{x})$ and is mapped onto data space $\Re^D$ by a smooth mapping f. Because $f(\Re^L)$ is an L-dimensional manifold in $\Re^D$, in order to extend it to the whole D-dimensional data space we define a distribution $p(\vec{t}|\vec{x}) = p(\vec{t}|f(\vec{x}))$ on $\Re^D$, called the noise or error model. The prior in latent space $p(\vec{x})$, the smooth mapping f and the noise model $p(\vec{t}|\vec{x})$ are collectively called parameter set Θ. In latent variable modeling these parameters are typically maximized according to the maximum likelihood (ML) criterion of the observed data given the parameters, $p(\vec{t_n}|\Theta)$. This optimization is often implemented using an expectation-maximization (EM)algorithm [19].

In summary, a latent variable model is defined by:

1. the functional form of the prior $p(\vec{x})$ in latent space;
2. the smooth mapping $f : \Re^L \mapsto \Re^D$ from latent space to data space;
3. the noise model $p(\vec{t}|\vec{x})$ in data space.

Fig. 4.1 illustrates the idea of latent variable model with a three-dimensional data space and a two-dimensional latent space.

The joint probability density function (pdf) in the product space $\Re^D \times \Re^L$ is $p(\vec{t}, \vec{x})$ and integrating over latent space gives the marginal distribution in data space,

$$p(\vec{t}) = \int p(\vec{t}, \vec{x}) d\vec{x} = \int p(\vec{t}|\vec{x}) p(\vec{x}) d\vec{x} \tag{4.1}$$

The log-likelihood of the parameters given the sample $\{\vec{t_n}\}_{n=1}^N$ is:

$$l(\Theta) = log \prod_{n=1}^{N} p(\vec{t_n}|\Theta) = \sum_{n=1}^{N} log p(\vec{t_n}|\Theta) \tag{4.2}$$

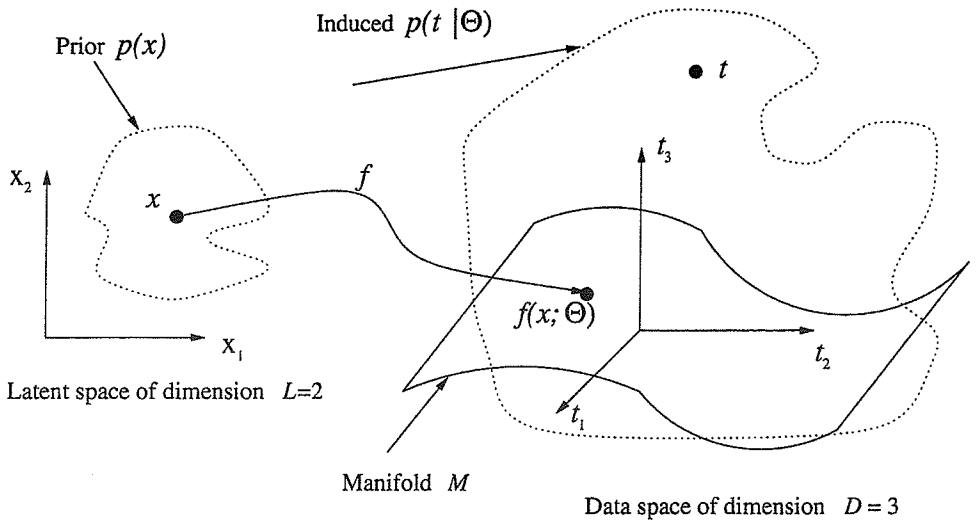


Figure 4.1 Schematic illustration of a latent variable model with a three-dimensional data space and a two-dimensional latent space.

where the parameter set Θ is estimated using the maximum likelihood (ML) criterion $\Theta^* = argmax_\Theta l(\Theta)$ corresponding to a local maximum of the log-likelihood. Once the parameters Θ are fixed, Bayes' theorem gives the posterior distribution in latent space given a data vector $\vec{t}$, i.e., the distribution of the probability that a point $\vec{x}$ in latent space was responsible for generating $\vec{t}$.

$$p(\vec{x}|\vec{t}) = \frac{p(\vec{t}|\vec{x})p(\vec{x})}{p(\vec{t})} = \frac{p(\vec{t}|\vec{x})p(\vec{x})}{\int p(\vec{t}|\vec{x})p(\vec{x})d\vec{x}} \tag{4.3}$$

Summarizing this distribution in a single latent space point $\vec{x}^*$ results in a reduced-dimension representative of $\vec{t}$. This defines a corresponding mapping F from data space onto latent space, so that every data point $\vec{t}$ is assigned a representative in latent space, $\vec{x}^* = F(\vec{t})$. Thus, it can be considered as an inverse mapping of f. This mapping F will be most successful when the posterior distribution $p(\vec{x}|\vec{t})$ is unimodal and sharply peaked. Typical choices for $F(\vec{t})$ are the minimum mean square error (MMSE) estimator $E\{\vec{x}|\vec{t}\}$ or the maximum *a posteriori* (MAP) estimator $argmax_{\vec{x}} p(\vec{x}|\vec{t})$. On the other hand, applying the mapping f to the reduced-dimension representative we obtain the reconstructed data vector $\vec{t}^* = f(\vec{x}^*)$. The reconstruction error for that point $\vec{t}$ is defined

in terms of some distance $\triangle(\vec{t}, \vec{t}^*)$ in data space and the average reconstruction error for the sample as:

$$E_\triangle = \frac{1}{N}\sum_{n=1}^{N} \triangle(\vec{t}_n, \vec{t}_n^*) \tag{4.4}$$

For example, the Euclidean distance provides with the mean squared error criterion $E_2 = \frac{1}{N}\sum_{n=1}^{N} \|\vec{t}_n - \vec{t}_n^*\|_2^2$.

In this section, we will investigate a specific latent variable model, namely the factor analysis. In the FA case, the parameters of the model are estimated using the EM algorithm [19]. The E-step involves estimating the posterior distribution of the latent variables using the current parameter values. The M-step is a maximization of the log-likelihood function in which the posterior distribution estimated in the E-step is used to fill in the latent variables.

4.3.2 Factor Analysis

Factor analysis uses a Gaussian distribution prior and noise model, and a linear mapping from the data space to the latent space. Specifically, a FA model is defined by:

1. The latent space prior $p(\vec{x})$ has unit normal distribution:

$$p(\vec{x}) \sim N(\vec{0}, I) \tag{4.5}$$

The latent variable $\vec{x}$ is often referred to as the factors.

2. The smooth mapping f is linear:

$$f(\vec{x}) = \Lambda\vec{x} + \vec{\mu} \tag{4.6}$$

The columns of the $D \times L$ matrix A are referred to as the factor loadings.

3. The data space noise model is normal centered in $f(\vec{x})$ with diagonal covariance matrix Φ:

$$p(\vec{t}|\vec{x}) \sim N(f(\vec{x}), \Phi) \tag{4.7}$$

The D diagonal elements of Φ are referred to as the uniqueness.

The marginal distribution in data space is normal with a constrained covariance matrix. Note that Eqs. (4.5) - (4.6) define a conventional two-way factor analysis. In three-way factor analysis such as PARAFAC, the smooth mapping function $f(\vec{x})$ in Eq. (4.6) is nonlinear.

CHAPTER 5

VOCAL FOLD EXCITATION MODELS

In articulatory synthesis, the vocal tract must be excited by some appropriate signal. The choice and method employed to excite the synthesizer play a vital part in determining the spectral characteristics and the degree of naturalness of the synthesized speech. Most speech synthesizers based on LPC techniques assume source-tract separability. For voiced excitations, a sequence of impulses is generated with the spacing determined by the desired fundamental frequency. The unvoiced excitation is generated by a band-limited white nosie signal. In order to incorporate source-tract interaction into an articulatory synthesizer, a more realistic model of the glottal excitation must be considered. The following sections present two methods for voiced excitation : parametric models of the glottal excitation and the multi-mass model of the vocal fold.

5.1 Parametric Models

Holmes [42], Rothenberg [92] and several other researchers investigated an alternative approach to inverse filtering of the speech waveform to generate the excitation signal. There findings show that this approach is capable of giving a good estimate of the glottal

volume velocity. In the following parts, we will describe two representative parametric models developed by Rosenberg [91] and Titze [126], respectively.

5.1.1 Rosenberg's Model

In 1971, Rosenberg applied a pitch-synchronous re-synthesis method to produce speech utterances with various source waveforms [91]. In his perceptual tests, the most natural excitation signal involves specification of several parameters. In order to better explain Rosenberg's model, we need to first introduce the general waveform of the glottal area during the vocal fold excitation as shown in Fig. 5.1

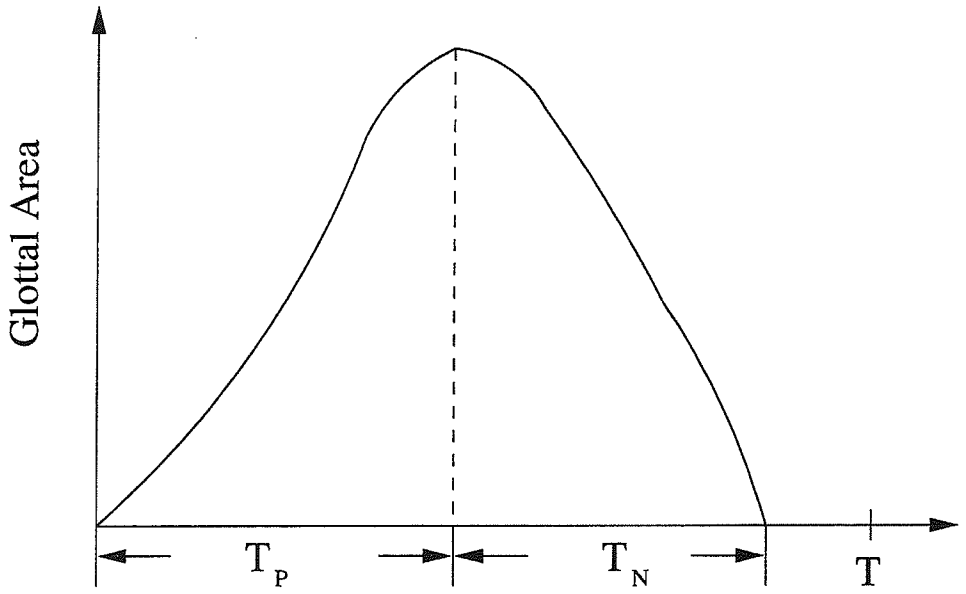


Figure 5.1 General waveform of the glottal area during excitation.

As shown in Fig. 5.1, T denotes the pitch period, T_P denotes the opening time during the glottal excitation and T_N denotes the closing time during the glottal excitation, respectively. In Rosenberg's model, the glottal waveform is specified by four parameters, namely, amplitude factor α, pitch period T, *open quotient* $\frac{T_P+T_N}{T}$ which denotes the ratio

of pulse duration to pitch period, and *speed quotient* $\frac{T_P}{T_N}$ which denotes the ratio of the rising to falling pulse durations. The glottal area function $A_g(t)$ is given by:

$$A_g(t) = \begin{cases} \alpha[3(\frac{t}{T_P})^2 - 2(\frac{t}{T_P})^3] & 0 \leq t \leq T_P \\ \alpha[1 - (\frac{t-T_P}{T_N})^2] & T_P < t \leq T_P + T_N \\ 0 & T_P + T_N < t \end{cases} \tag{5.1}$$

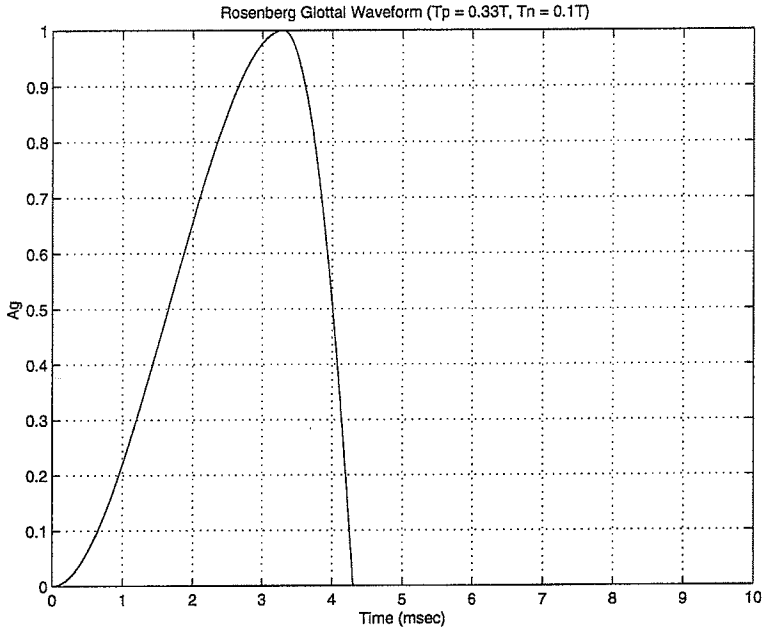


Figure 5.2 Glottal waveform computed from Rosenberg's model .

Fig. 5.2 shows a glottal waveform computed from Rosenberg's model, where the parameters are set as: $\alpha = 1.0$, $T = 10ms$, $T_P = 0.33T$ and $T_N = 0.1T$, respectively.

5.1.2 Titze's Model

In 1982, Titze proposed a parametric model to represent the glottal area [126]. The glottal waveform in Titze's model is essentially similar to the Rosenberg pulse described in Eq. (5.1), with an extra parameter β to determine the residual decay of the falling slope. The glottal area function in this model is given by:

$$A_g(t) = \begin{cases} \alpha[(\frac{\theta}{\theta_m})^{-\theta_m cot\theta_m}(\frac{sin\theta}{sin\theta_m})]^{\beta} & \theta \leq \pi \\ 0 & \theta \end{cases} \tag{5.2}$$

where $\theta \stackrel{\Delta}{=} \frac{\pi t}{T_P+T_N}$ and $\theta_m \stackrel{\Delta}{=} \frac{\pi T_P}{T_P+T_N}$, respectively.

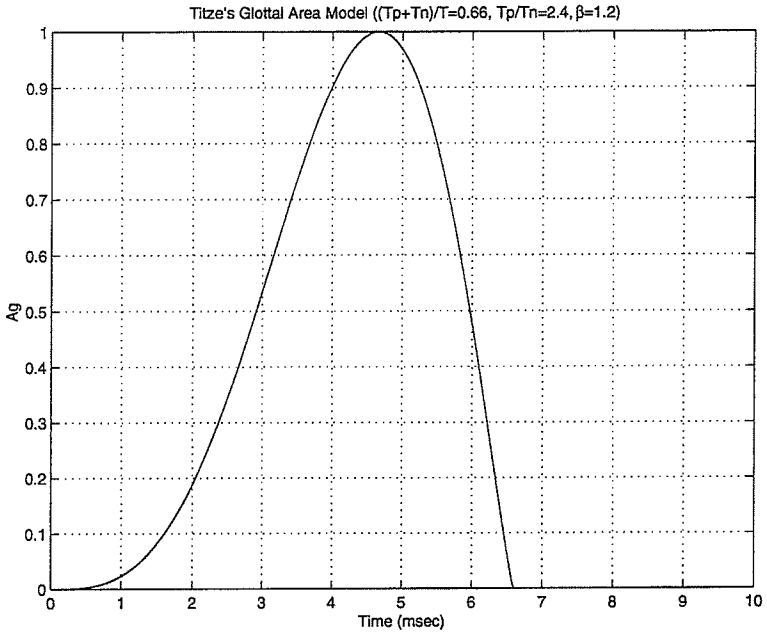


Figure 5.3 Glottal waveform computed from Titze's model.

Fig. 5.3 shows a glottal waveform computed from Titze's model, where the parameters are set as: $\alpha = 1.0$, $T = 10ms$, $T_P = 2.4T_N$, $(T_P+T_N) = 0.66T$, and $\beta = 1.2$, respectively.

5.2 Mechanical Model

Over the years, several researchers have developed a number of different methods for the realistic modeling of the vocal fold excitation during speech production. In 1972, Ishizaka and Flanagan [54] developed a two-mass model of the vocal fold by chosing the parameters from mechanical considerations and including aerodynamics with estimations of glottal waves produced in normal phonation. Other researchers have proposed similar

vocal fold models which differ from each other by the parameters and the way of aerodynamic influences [61], [115]. Recently, researchers have introduced finite-element method (FEM) to model the detailed geometric and material information in vocal fold [6], [7]. In this thesis, we focus on the lumped parameter model of the vocal fold and its application to articulatory synthesis and the experimental evidence of source-tract interaction.

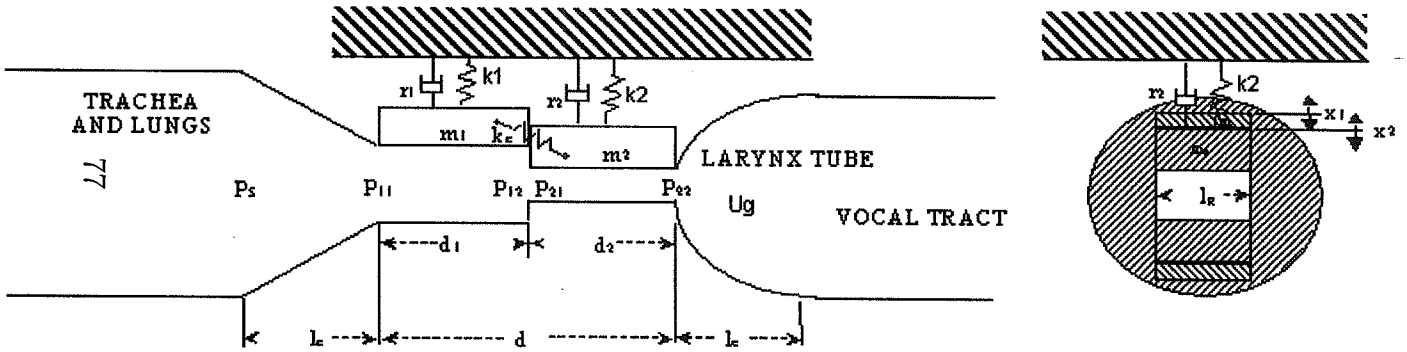
5.2.1 Two-Mass Model

Fig. 5.4 shows that the two-mass model describes one vocal fold by two coupled oscillators, where P_s, A_g and U_g denote the sub-glottal pressure, glottal area, and glottal volume velocity, respectively. Each oscillator consists of a mass, a spring stiffness, and a damper. Mass m_1, linear spring stiffness k_1, and damper r_1 represent the lower part of the vocal fold with thickness d_1. Mass m_2, linear spring stiffness k_2, and dumper r_2 represent the upper part with thickness d_2. The two masses are coupled by a spring stiffness k_c. The two masses, m_1 and m_2 are permitted to move in a lateral direction. The deflections of m_1 and m_2 are x_1 and x_2, respectively. In the two-mass model, symmetry along the length of the glottis is assumed, therefore only one vocal fold is considered. When the vocal fold approaches the symmetry line within a very short distance, collision springs will be activated and have an influence on m_1 and m_2, respectively, in the contralateral direction.

The dynamic response of the mechanical part of the two-mass model can be described by the equations of motion of the two masses as follows.

$$m_1\ddot{x}_1 + r_1\dot{x}_1 + (k_1 + k_c)x_1 - k_c x_2 = F_1 \tag{5.3}$$

$$m_2\ddot{x}_2 + r_2\dot{x}_2 + (k_2 + k_c)x_2 - k_c x_1 = F_2 \tag{5.4}$$


TRACHEA
AND LUNGS
77
Ps
P11
P12
P21
P22
Ug
VOCAL TRACT
LARYNX TUBE
r1
k1
r2
k2
m1
kc
m2
d1
d2
lc
d
lc

r2
k2
x1
x2
lR

where F_1 and F_2 are the aerodynamic forces exerted on the masses. $\dot{x}$ and $\ddot{x}$ denotes the first and second derivatives of the variable x, respectively. In our system, an improved two-mass model proposed by Don Davis and Scott Slimon in Electronic Boat corporation is used. The improved equations of motion are given as follows.

$$m_1\ddot{x}_1 + r_1\dot{x}_1 + s_1(x_1) + k_c(x_1 - x_2) = F_1 \tag{5.5}$$

$$m_2\ddot{x}_2 + r_2\dot{x}_2 + s_2(x_2) + k_c(x_2 - x_1) = F_2 \tag{5.6}$$

$$s_1(x_1) = \begin{cases} k_1(x_1 + \eta_{k1}x_1^3) + h_1[(x_1 + \frac{A_{g01}}{2l_g}) + \eta_{h1}(x_1 + \frac{A_{g01}}{2l_g})^3], & x_1 \leq -\frac{A_{g01}}{2l_g} \\ k_1(x_1 + \eta_{k1}x_1^3), & x_1 > -\frac{A_{g01}}{2l_g} \end{cases} \tag{5.7}$$

$$s_2(x_2) = \begin{cases} k_2(x_2 + \eta_{k2}x_2^3) + h_2[(x_2 + \frac{A_{g02}}{2l_g}) + \eta_{h2}(x_2 + \frac{A_{g02}}{2l_g})^3], & x_2 \leq -\frac{A_{g02}}{2l_g} \\ k_1(x_2 + \eta_{k2}x_2^3), & x_2 > -\frac{A_{g02}}{2l_g} \end{cases} \tag{5.8}$$

where A_{g01} and A_{g02} denote the area through mass 1 and mass 2 at the phonation neutral position, l_g denotes the depth of glottis slit, k_1 and k_2 denote the open linear spring constant of mass 1 and mass 2, h_1 and h_2 denote the closed linear spring constant of mass 1 and mass 2, η_{k1} and η_{k2} denote the open non-linear spring constant of mass 1 and mass 2, η_{h1} and η_{h2} denote the closed non-linear spring constant of mass 1 and mass 2.

The vocal fold model is coupled with the aerodynamic equations by relating the aerodynamic forces F_1 and F_2 to the glottal volume velocity U_g in the following way.

$$P_s - P_{11} = 1.37\frac{\rho}{2}(\frac{U_g}{A_{g1}})^2 + \int_0^{l_c} \frac{\rho}{A_c(x)}dx\frac{dU_g}{dt} \tag{5.9}$$

$$P_{11} - P_{12} = 12\frac{\mu l_g^2 d_1}{A_{g1}^3}U_g + \frac{\rho d_1}{A_{g1}}\frac{dU_g}{dt} \tag{5.10}$$

$$P_{12} - P_{21} = \frac{\rho}{2}U_g^2(\frac{1}{A_{g2}^2} - \frac{1}{A_{g1}^2} \tag{5.11}$$

$$P_{21} - P_{22} = \frac{\mu l_g^2 d_2}{A_{g2}^3}U_g + \frac{\rho d_2}{A_{g2}}\frac{dU_g}{dt} \tag{5.12}$$

$$P_{22} - P_1 = -\frac{\rho}{2}(\frac{U_g}{A_{g2}})^2 \frac{A_{g2}}{A_1}(1 - \frac{A_{g2}}{A_1}) \tag{5.13}$$

$$F_1 = \frac{1}{2}(P_{11} + P_{12}) l_g d_1 \tag{5.14}$$

$$F_2 = \frac{1}{2}(P_{21} + P_{22}) l_g d_2 \tag{5.15}$$

where $A_c(x)$ denotes the contraction area at x, A_1 denotes the area after expansion, A_{g1} and A_{g2} denotes the area through mass 1 and mass 2, respectively.

The numerical solution of the vocal fold excitation signals are computed in the folliwng iterative way. In the n^{th} iteration, the volume velocity computed from the previous iteration $U_g^{(n-1)}$ is used to compute the aerodynamic forces $F_1^{(n)}$ and $F_2^{(n)}$ using Eqs. (5.9) - (5.15). Then the computed forces $F_1^{(n)}$ and $F_2^{(n)}$ in the n^{th} iteration are subtituted into Eqs. (5.5) - (5.8) to compute the mass displacements $x_1^{(n)}$ and $x_2^{(n)}$. From $x_1^{(n)}$ and $x_2^{(n)}$, we further compute the areas $A_{g1}^{(n)}$ and $A_{g2}^{(n)}$. Finally, the glottal area in the n^{th} iteration is determined by $A_g^{(n)} = min\{A_{g1}^{(n)}, A_{g2}^{(n)}\}$. This glottal area is used to compute the volume velocity in next iteration, which is being used as one of the boundary conditions of the governing equations (Navier-Stokes equations) of fluid flow in next iteration.

5.2.2 M-Mass Model

In the current synthesis system, the two-mass vocal fold model described in previous section is used to compute the glottal excitation signals for articulatory speech synthesis. However, an M-mass model is needed to simulate the velocity field and vorticity field of the vocal apparatus because a two-mass model is insufficient to describe the flow inside vocal fold in this case. Furthermore, an M-mass model has the potential to comute more accurate glottal excitation siganls for articulatory speech synthesis.

Similarly, we can get the equations of motion of the M-mass model as following:

$$
\begin{aligned}
&m_1\ddot{x}_1 + r_1\dot{x}_1 + k_1x_1 + K_{1,2}(x_1 - x_2) = F_1 \\
&m_i\ddot{x}_i + r_i\dot{x}_i + k_ix_i + k_{i,i-1}(x_i - x_{i-1}) + k_{i,i+1}(x_i - x_{i+1}) = F_i, \quad i = 2, \cdots, M-1 \\
&m_M\ddot{x}_M + r_M\dot{x}_M + k_Mx_M + K_{M,M-1}(x_M - x_{M-1}) = F_M
\end{aligned} \tag{5.16}
$$

where k_i denotes the spring stiffness of mass m_i, r_i denotes the damping coefficient of mass m_i, k_{ij} denotes the coupling spring stiffness between mass m_i and mass m_j. Alternatively, we can write the equations of motion of the M-mass model in a matrix form as follows:

$$
M_{mass}\ddot{\vec{x}} + R\dot{\vec{x}} + K\vec{x} = \vec{F} \tag{5.17}
$$

where M_{mass}, R and K denotes the mass matrix, the damping matrix and the stiffness matrix K, respectively. In a two-mass model, these matrices are given by $M_{mass} = \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix}$, $R = \begin{bmatrix} r_1 & 0 \\ 0 & r_2 \end{bmatrix}$, $K = \begin{bmatrix} k_1 + k_{12} & -k_{12} \\ -k_{12} & k_2 + k_{12} \end{bmatrix}$. However, if we want to consider the non-linear effects of mass spring and both the open spring constant and closed spring constant like the improved two-mass model described in previous section, the simple matrix form in Eq. (5.17) can not be used and terms $k_ix_i, \quad i = 1, \cdots, M$ in Eq. (5.16) should be accordingly replaced by $s_i(x_i)$. The $s_i(x_i)$ is given by:

$$
s_i(x_i) = \begin{cases} k_i(x_i + \eta_{ki}x_i^3) + h_i[(x_i + \frac{A_{g0i}}{2l_g}) + \eta_{hi}(x_i + \frac{A_{g0i}}{2l_g})^3], & x_i \leq -\frac{A_{g0i}}{2l_g} \\ k_i(x_i + \eta_{ki}x_i^3), & x_i > -\frac{A_{g0i}}{2l_g} \end{cases} \tag{5.18}
$$

where A_{g0i} denotes the area through mass i at the phonation neutral position, k_i denotes the open linear spring constant of mass i, h_1 denotes the closed linear spring constant of mass i, η_{ki} denotea the open non-linear spring constant of mass i, η_{hi} denotes the closed non-linear spring constant of mass i, for $i = 1, \cdots, M$.

5.3 Experimental Results

In this parts, we present the simulation results on vocal excitation based on the M-mass model described above. Our experimental results include the glottal area A_g, glottal particle velocity V_g and glottal volume velocity U_g of an English sentence *"Where were you while we were away"*. The glottal particle velocity V_g and glottal volume velocity U_g will be used as initial conditions for the Navier-Stokes (N-S) equations for articulatory synthesis which will be discussed in the later chapter. The different experimental conditions are listed in Table 5.1.

Table 5.1 Different experimental conditions for the simulation of glottal excitation.

Filename	VTLN	Time-varying pitch	Duration of sentence
runjj	Yes (factor = 1.4)	No	2.4 sec
runkk	No	No	2.4 sec
runll	No	Yes	2.4 sec
runmm	No	Yes	1.2 sec

In Table 5.1, VTLN denotes vocal tract length normalization. The reason of using VTLN is to compensate the unrealistic vocal tract length (greater than 23 cm in most cases) computed from Coker's model. In the case of time-invarying pitch, a single pitch ($F_0 = 100$ Hz is used throughout the simulation. In the case of time-varying pitch, the pitch contour of a waveform of the sentence *"Where were you while we were away"* from the TIMIT database is extracted using the ESPS/Xwaves+ tool and applied to the glottal excitation simulation. The duration of the original sentence from TIMIT database is 1.2 sec. In our experiments, we found out that the speaking rate of this sentence is too fast to synthesize natural sounding speech signal so that we artificially double the synthesis length in order to compensate the flaw from the database. We have also segmented the phonetic boundaries of several recorded normal speaking rate utterances and will use this

prosodic information to re-synthesize continuous speech sentences later. The waveforms of the simulated glottal area A_g, glottal particle velocity V_g and glottal volume velocity U_g are shown in Figs. 5.5 - 5.8, Figs. 5.9 - 5.12 and Figs. 5.13 - 5.16, respectively. In Figs. 5.5 - 5.16, one time step corresponds to 10^{-5} second.

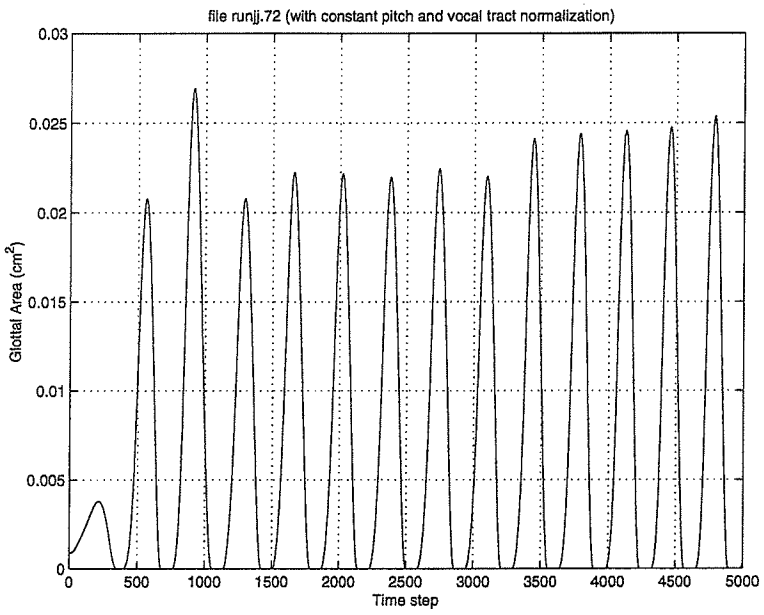


Figure 5.5 Simulated glottal area under the experimental condition runjj.

5.4 Discussion

From Figs. 5.5 - 5.16, we can see that the M-mass model of the vocal folds is able to generate more realistic results of glottal area and glottal volume velocity. This method is better than the parametric models described in Section 5.1 because we can clearly observe ripples especially in the positive glottal opening interval of the simulated U_g waveforms, which is a strong evidence of the source-tract interaction. Another evidence of source-tract interaction is the observation of the skewing effect of the glottal waveform due to the existence of vocal tract constriction. Our results are similar to those reported

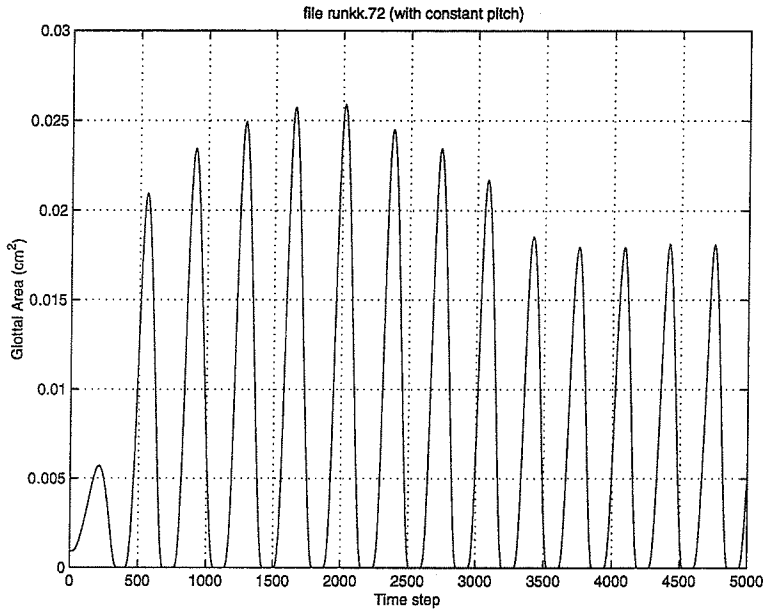


Figure 5.6 Simulated glottal area under the experimental condition runkk.

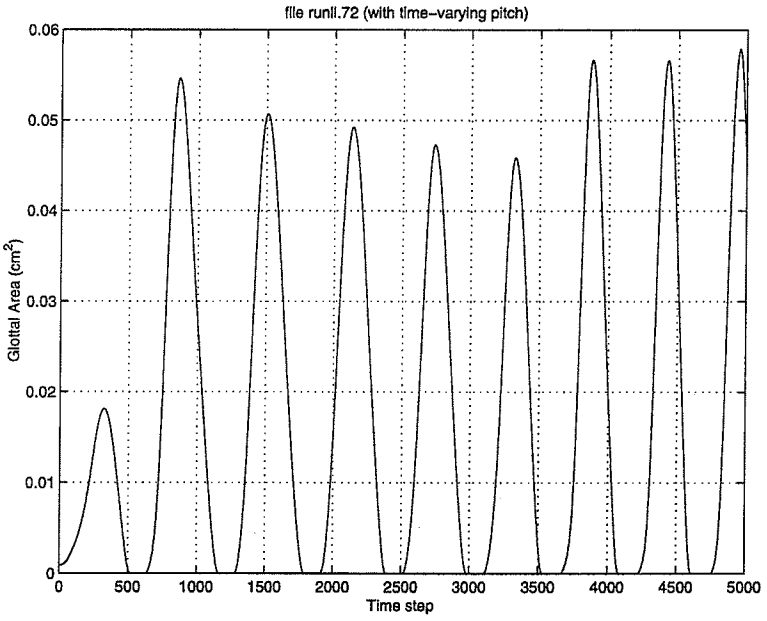


Figure 5.7 Simulated glottal area under the experimental condition runll.

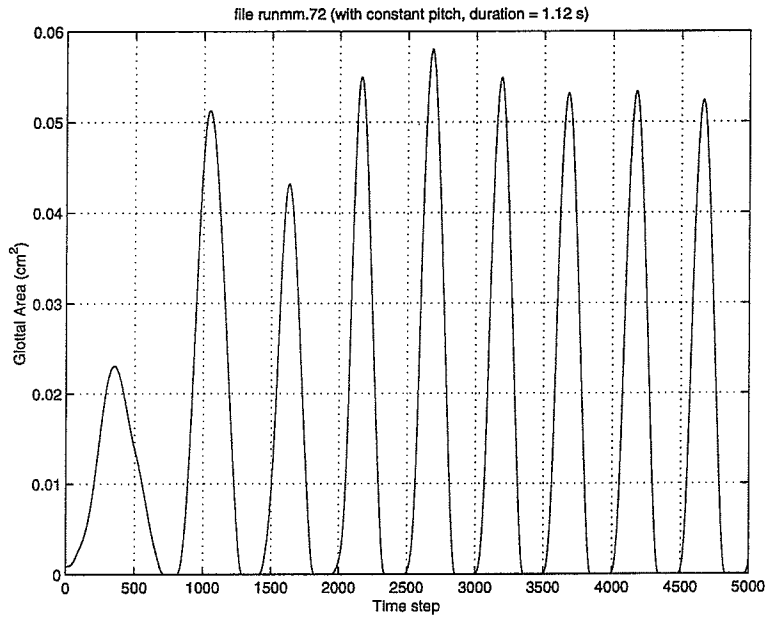


Figure 5.8 Simulated glottal area under the experimental condition runmm.

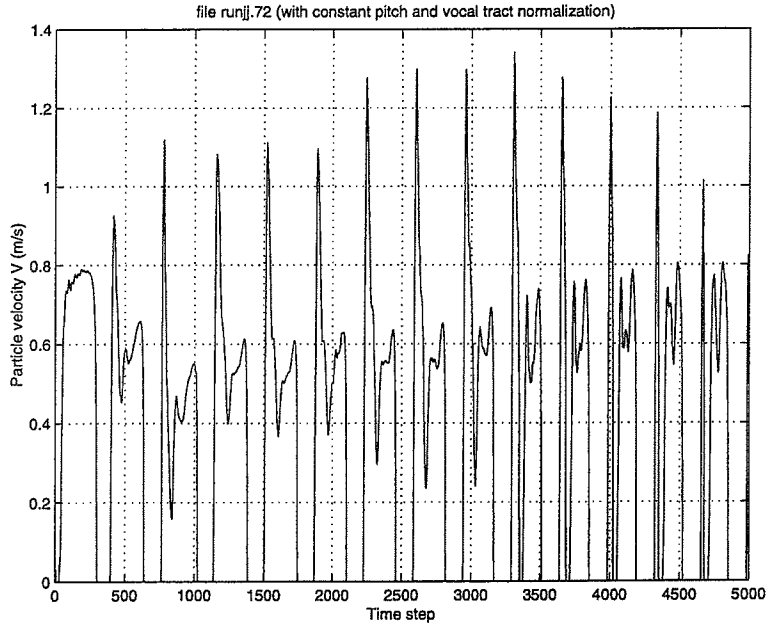


Figure 5.9 Simulated glottal particle velocity under the experimental condition runjj.

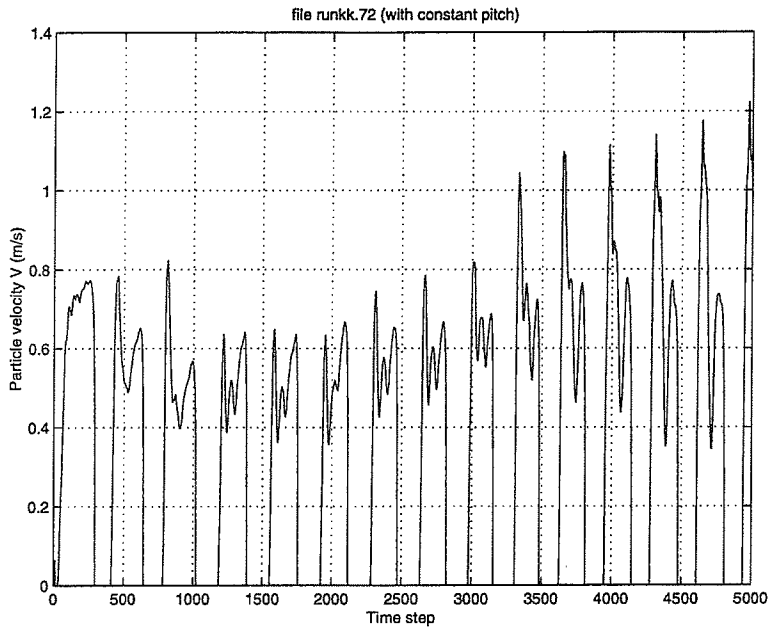


Figure 5.10 Simulated glottal particle velocity under the experimental condition runjkk.

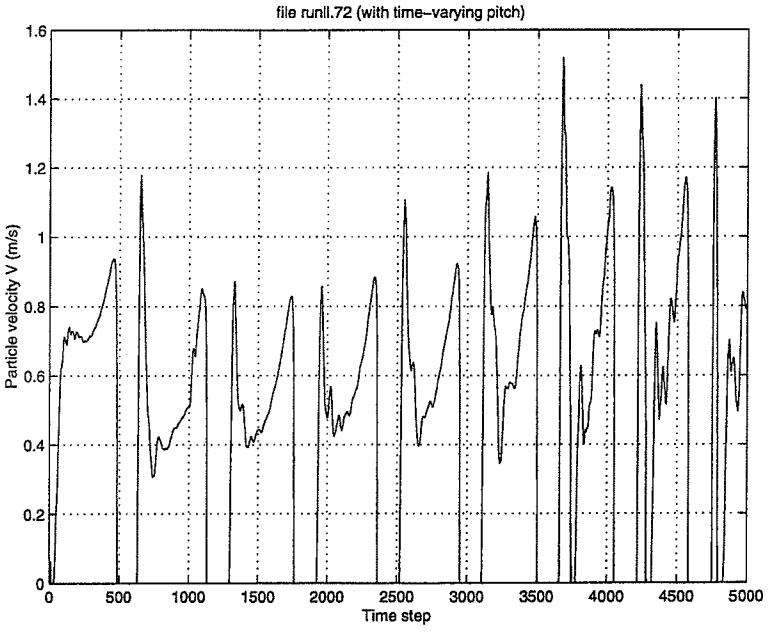


Figure 5.11 Simulated glottal particle velocity under the experimental condition runll.

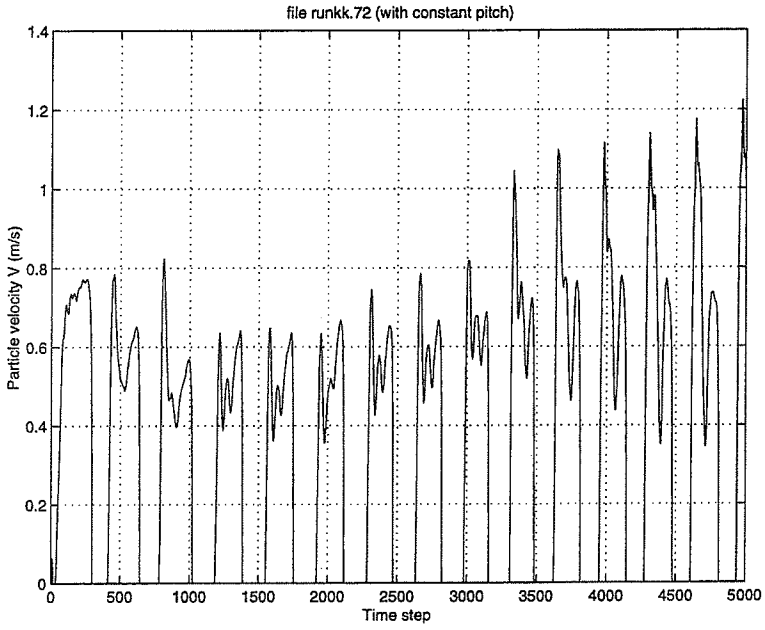


Figure 5.12 Simulated glottal particle velocity under the experimental condition runjmm.

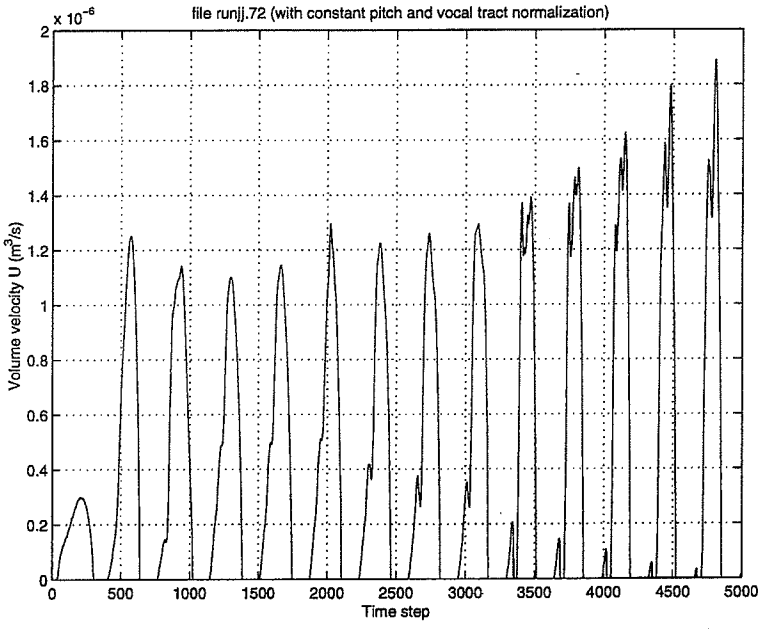


Figure 5.13 Simulated glottal volume velocity under the experimental condition runjj.

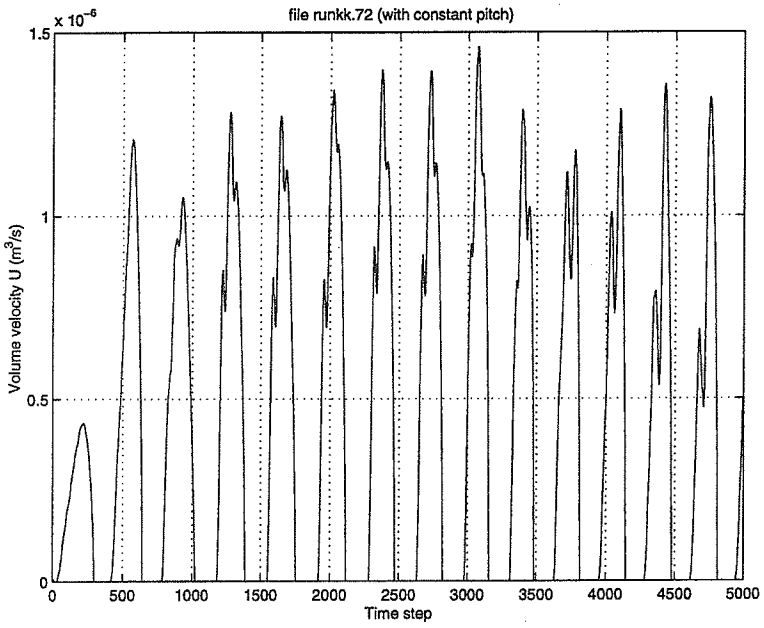


Figure 5.14 Simulated glottal volume velocity under the experimental condition runkk.

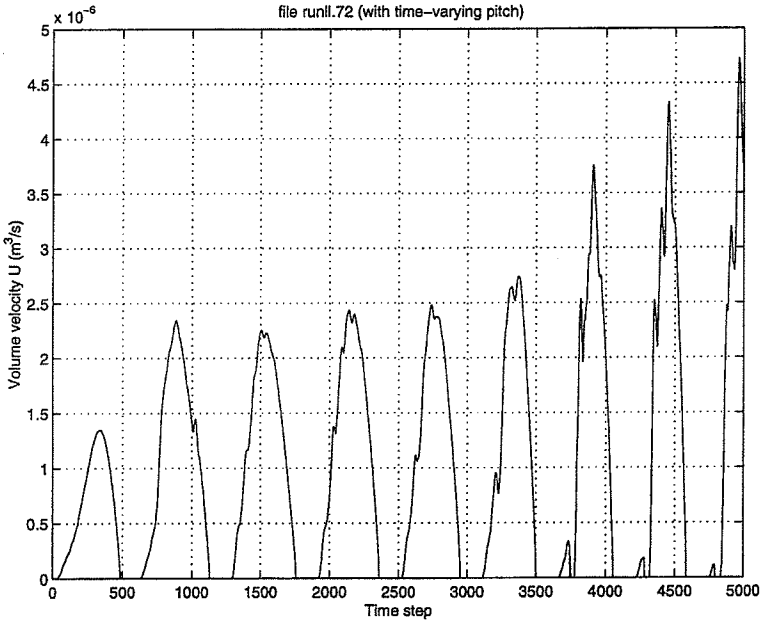


Figure 5.15 Simulated glottal volume velocity under the experimental condition runll.

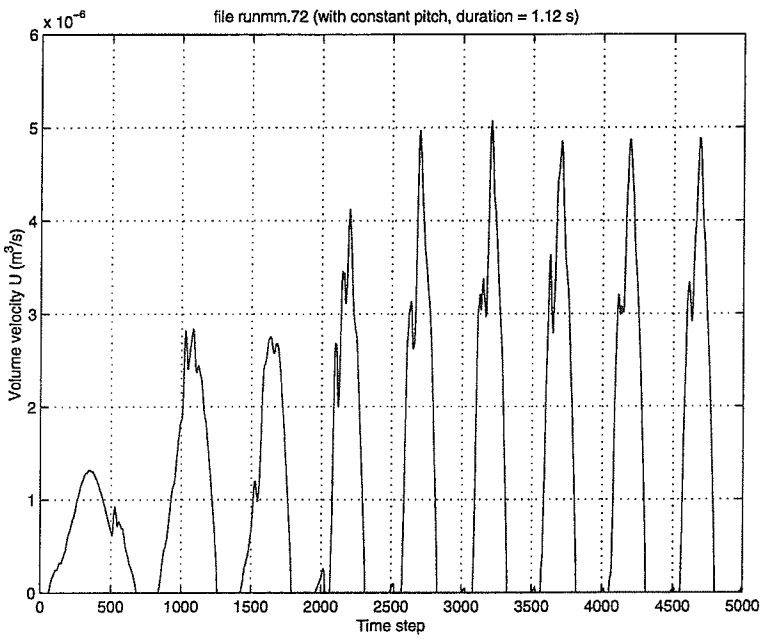


Figure 5.16 Simulated glottal volume velocity under the experimental condition runmm.

by Fant *et al.* [28], [29], [30]. However, our experimental results also simulate the glottal particle velocity which is not done in Fant's work. Furthermore, multiples ripples can be observed in our simulation results of glottal particle velocity and glottal volume velocity. Our assumption is that the source-tract interaction is due the influences of both the first and the second formants, although some Fourier analysis need to be done to verify this assumption. The source-tract separability is unsuitable during the production of unvoiced sounds due to the additional excitation source required at the vocal tract constriction. Even during the production of voiced sounds this assumption only holds as a first order approximation since the glottal excitation signal depends on the transglottal pressure which, in turn, is controlled by the subglottal pressure and the vocal tract shape. A more realistic model of the vocal fold excitation which considers the non-linear effects due to source-tract interaction is not only important for speech synthesis, but also beneficial for low-bit rate perceptual speech coding.

CHAPTER 6

ANALYSIS OF THE VELOCITY AND VORTICITY FIELDS

6.1 Simulation Results

Simulation results of the velocity and vorticity fields of speech production. The disk space problem is still not solved...

6.2 Analysis

Analysis of the simulation results.

6.3 Relation to Speech Production Model

Can we use this analysis result to verify (part) of Krane's speech production model? But we need to synthesize both voiced and unvoiced sounds in this case.

CHAPTER 7

SPEECH SYNTHESIS AND ANALYSIS RESULTS

7.1 Brief Overview of Fluid Dynamics

From the prospect of fluid dynamics, the continuum hypothesis asserts that fluids exist and can be measured as a collection of infinitesimally small particles characterized by mechanically based continuous variables such as particle velocity $\vec{v}$, pressure P, density ρ, *et al.* The *Eulerian* description of particles uses a frame which is fixed in space for all time. The *Eulerian* velocity at a fixed point $\vec{x}$ and time t is the instantaneous velocity of the particle passing that point, $\vec{v} = \vec{v}(\vec{x}, t)$. Consider a material control volume (MCV) at two times, t and $t + dt$, enclosing some material, where α is the density of some quantity A per unit volume. The total amount of A in the MCV at any given time t is given by:

$$A = \int_{V(t)} \alpha(t) dV \tag{7.1}$$

Reynold's transport theorem is one of the basic principles for all kinds of fluid dynamics. It describes the conservation for the total amount of A within the MCV and can be stated as following. The total rate of change of A = The instantaneous rate of change

of A within V + the flux of A through the boundary of S [17]. This statement can be written in the following mathematical form:

$$\frac{D}{Dt}\int_{V(t)} \alpha dV = \int_{V(t)} \frac{\partial \alpha}{\partial t} dV + \int_{S(t)} (\alpha \vec{v}) \cdot \vec{n} dS \tag{7.2}$$

where $\vec{v} \cdot \vec{n}$ is the velocity normal to the surface of $S(t)$ and $(\alpha \vec{v}) \cdot \vec{n} dS$ is the rate of volume changes across dS. Note that the operation $\frac{D}{Dt}$ is called the *material derivative* and is different from the conventional derivative defined in Calculus. For example, if α is a scalar, then its material derivative is given by:

$$\frac{D\alpha}{Dt} = \left.\frac{\partial \alpha}{\partial t}\right|_{\vec{x}} + \vec{v} \cdot \nabla \alpha \tag{7.3}$$

Another fundamental theorem in fluid dynamics is the *divergence theorem.* It relates surface and volume integrals for suitably differentiable densities α, velocity field $\vec{v}$ and bounding surface normals $\vec{n}$. The divergence theorem can be written as following [17]:

$$\int_{S(t)} \alpha \vec{v} \cdot \vec{n} dS = \int_{V(t)} \nabla \cdot (\alpha \vec{v}) dV \tag{7.4}$$

Combining Eq. (7.2) and (7.4) we can get another form of Reynold's transportation theorem:

$$\frac{DA}{Dt} = \int_V \frac{\partial \alpha}{\partial t} + \nabla \cdot (\alpha \vec{v}) dV \tag{7.5}$$

Consider a special case where the quantity A denotes the fluid mass and the density α denotes the mass density of the fluid accordingly, we can get the following equation of the conservation of mass:

$$\frac{DM}{Dt} = \int_V [\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \vec{v})] dV \tag{7.6}$$

Another conservation law in fluid dynamics is the conservation of momentum. It states that the time rate of change of the momentum of the MCV = the sum of all the external forces acting on the MCV. This *momentum equation* is written in the following tensor form:

$$\rho\frac{\partial v_i}{\partial t} + \rho\frac{\partial v_i v_j}{\partial x_j} = \rho f_i + \frac{\partial \sigma_{ji}}{\partial x_j} \tag{7.7}$$

where v_i denote the three components of velocity, σ_{ji} denote the six components of stress tensor, ρ denotes the density of the medium and f_i denotes the extrinsic applied force. When we look at Eq. (7.6) and Eq. (7.7), we can find that wave equation based on linear acoustics omits the second term $v_j\frac{\partial \rho}{\partial x_j}$ in Eq. (7.6) and the second term $\rho\frac{\partial v_i v_j}{\partial x_j}$ and the fourth term $\frac{\partial \sigma_{ji}}{\partial x_j}$ in Eq. (7.7). The terms $v_j\frac{\partial \rho}{\partial x_j}$, $\rho\frac{\partial v_i v_j}{\partial x_j}$ and $\frac{\partial \sigma_{ji}}{\partial x_j}$ denote the convection of mass, convection of momentum and viscous force, respectively. This means that speech synthesis based on the fluid dynamic principles consider the convective mode and the viscous effects which are ignored in the conventional speech synthesis based on linear acoustics.

In this section, we briefly introduce the definition of several fluid dynamics notations which are directly related to our speech processing work. A *pathline* is a trajectory of a fluid particle, with a fixed particle name ξ, i.e. $\vec{x} = \vec{x}(\xi, t)$. A *streamline* is a line in space that is everywhere tangent to the Eulerian velocity field $\vec{v}(\vec{x}, t)$ at a given instant t. *Vorticity* is defined as the curl of the velocity field, i.e., $\vec{\omega} = \nabla \times \vec{v}$. A *vortex line* is a space curve that is everywhere tangent to the vorticity. The *circulation* about a given closed circuit C in a fluid is defined by the line integral:

$$\Gamma = \int_C \vec{v} \cdot \vec{t} ds = \int_S (\nabla \times \vec{v}) \cdot \vec{n} dS \tag{7.8}$$

where $\vec{t}$ is the tangent vector along the path C within the fluid and ds is the arclength along the path. The second equality comes from *Stoke's theorem.*

A *perfectly viscous fluid* is defined as a material which deforms irreversibly at a constant rate under the action of an applied shear force of constant magnitude. A *Newtonian fluid* is a perfectly viscous fluid in which the rate of deformation varies linearly with the magnitude of the local stress. Some properties of a Newtonian fluid and its constitutive relations are: (1). In the absence of shear forces there is no deformation of the fluid particles; (2) The stress tensor σ_{ij} is proportional to the rate of strain tensor E_{ij}. An *incompressible fluid* is the fluid with $\frac{D\rho}{Dt} = 0$ or $\nabla \cdot \vec{v} = 0$. An *irrotational flow* is the flow which satisfies $\nabla \times \vec{v} = 0$. The velocity of an irrotational flow can be represented as the gradient of a scalar potential Φ, i.e., $\vec{v} = \nabla\Phi$. An *unsteady flow* is the flow with non-zero time derivative of the potential Φ, i.e., $\frac{\partial\Phi}{\partial t} \neq 0$. An *ideal fluid* is defined as an *inviscid* ($\mu = 0$, $\lambda = 0$) and *incompressible* fluid. A *potential flow* is an ideal flow which is also irrotational. A *stagnation point* is the point in the velocity field where $\vec{v} = 0$. *Boundary layer* is the thin fluid layer adjacent to the surface of a body in which strong viscous effects exist.

7.2 Governing Equations

During the production of speech, both the irrotational and incompressible flow modes exist. In the irrotational flow, motion of air particles is characterized by the propagative transfer of energy and momentum. A fluid particle experience no net displacement from its initial position during the passage of a sound wave. The only forces active in a sound field are pressure forces, which act on the centroid of a fluid particle, producing no rotation upon it. The incompressible flow is by contrast characterized by transfer of energy and momentum through the actual displacement and rotation of air particles, which as a result end up far away from their initial location. At air speeds observed

in speech, which might approach $40m/s$ (Mach number = 0.12), these motions proceed without appreciable volume fluctuations.

The conservation laws of mass and momentum described in the previous section are Reynolds-averaged to get the Reynolds-Averaged Navier-Stokes (RANS) equations which are the governing equations describing both the irrotational and slightly compressible fluid motions during sound generation. The RANS equations can be written as the following tensor notation.

$$M^2 \frac{\partial p}{\partial t} + \frac{\partial v_i}{\partial x_i} = 0 \tag{7.9}$$

$$\frac{\partial v_i}{\partial t} + \frac{\partial v_i v_j}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j}\left[\frac{(\nu + \nu_t)}{Re}\frac{\partial v_i}{\partial x_j}\right] \tag{7.10}$$

where Eq. (7.9) is called the continuity equation, M denotes the Mach number and $M = \frac{convection\ \ speed}{speed\ \ of\ \ sound}$, v_i denote the mean velocity components, p denotes the static pressure, ν denotes the kinematic viscosity, ν_t denotes the turbulent eddy viscosity coefficient and Re denotes the Reynold number. In RANS equations, the flow is assumed to be slightly compressible and an isentropic assumption is used to relate pressure and density. These equations are applicable to time-dependent, low Mach number flows, such as those existing in the human vocal tract. Furthermore, a turbulence model is also used to simulate the turbulence effect in speech generation. Currently, we use an eddy viscosity mixing length model for articulatory synthesis. In this model, the eddy viscosity is expressed as a functionof local flow variables (velocity derivatives and wall distances) and a user-specified mixing length. Damping functions are included in the model for smooth transition of the eddy viscosity from its value in the log layer to the wall.

Until now, there is no closed-form solution to RANS equations which are highly non-linear partial derivative equations (PDE). In our work, a RANS-based software package developed at Electric Boat Corporation is used. This package is a multi-block, finite

difference based solver which is second order accurate in both space and time. This software package is run on a Origin 2000 supercomputer in the NCSA center of the University of Illinois. The key features of the RANS-based software package provided by Don Davis and Scott Slimon at Electric Boat Corporation can be summarized as follows.

1. It uses finite difference approximation , i.e., central difference.
2. It uses alternate direction implicit (ADI) solver.
3. It uses body-fitted coordinates.
4. It uses second and fourth order articicial dissipation to control oscillations.
5. It uses time-dependent metrices for moving grid problems.
6. The inlet boundary conditions are extrapolated pressure and specified axial particle velocities over a specified area. The velocities and area are determined from a multi-mass vocal fold model ($M = 2$ in our current synthesis results).

In our current system, a single grid system is used for RANS simulations. The resolution of this grid is designed such that it will capture both viscous and inviscous phenomena of sound generation.

7.3 Synthesized waveform

In this section, we present some experimental results of our articulatory speech synthesis system. The simulation time step of the RANS-solver is $10^(-5)$ second. The vocal fold model described in Chapter 5 is coupled with the RANS-solver and they are iteratively solved at the same time step. First, we present the synthesized speech waveform of a dipthong */AY/* ("uy" in buy). Forty vocal tract shapes were estimated using the dynamic parameter estimation method described in Chapter 3. The time step between each

consecutive vocal tract shapes is 10 *msec.* Moving grid , i.e., dynamic metrics terms were added to the RANS solver to accurately account for the actual movement. The boundary condition at the vocal tract inlet is specified by the particle velocity computed from the vocal fold excitation model. Fig. 7.1 shows the waveform of the synthesized dipthong */AY/*.

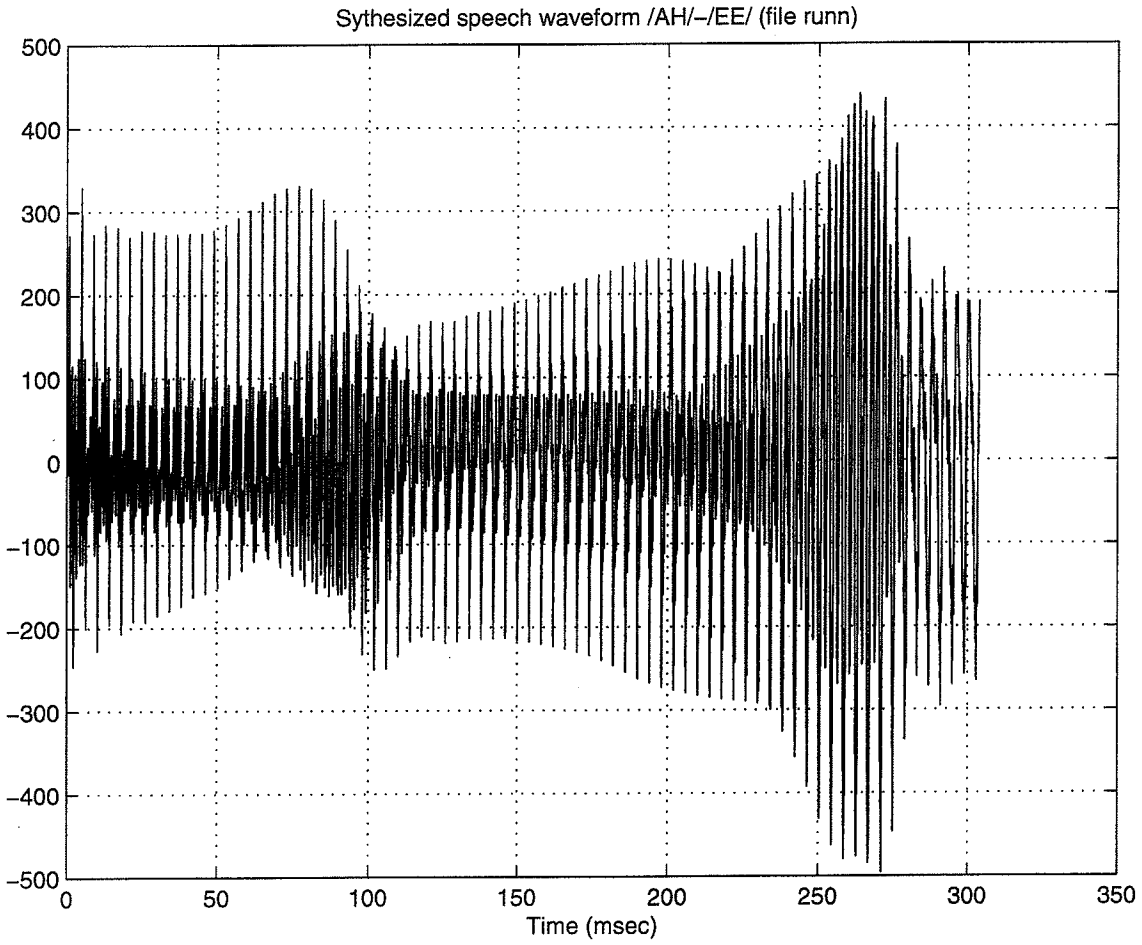


Figure 7.1 Waveform of the synthesized dipthong */AY/* ("uy" in buy).

In the next step, we synthesize the speech waveforms of an English sentence *"Where were you while we were away"*. The prosodic information (pitch contour, duration and energy) were extracted from a sentence in TIMIT database. Different simulation conditions were listed in Table 5.1. Figs. 7.2 - 7.5 show the waveforms of the synthesized

speech sentence under the experimental conditions of runjj, runkk, runll and runmm, respectively.

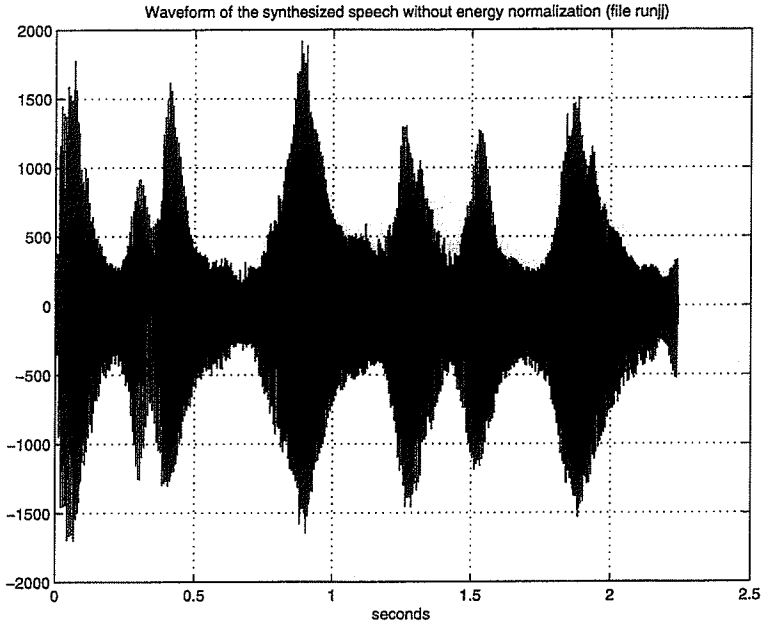


Figure 7.2 Waveform of the synthesized sentence *"Where were you while you were away"* under the experimental condition of runjj.

7.4 Speech Analysis Results

7.4.1 LPC Spectrum and the Short-Time Power Spectrum

In this section, LPC spectrum and the short-time power spectrum of the synthesized speech signal are computed. A preemphasis filter with the parameter $\alpha = 0.97$ is used to equalize the inherent spectral tilt in speech. A Hamming window with the length 24 *msec* is applied for preprocessing. The order of the LPC analysis is 12. The short-time power spectrum is computed through FFT technique. Fig. 7.6 shows the spectra of the second frame of the synthesized dipthong */AY/* (*uy* in *buy*, which corresponds to the spectra of vowel */AH/* (*u* in *but*). The upper figure denotes the LPC spectrum

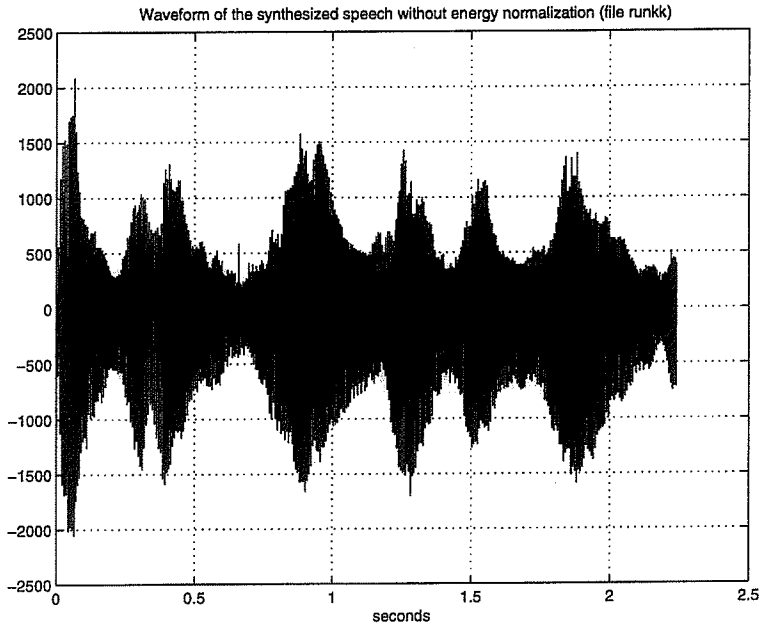


Figure 7.3 Waveform of the synthesized sentence *"Where were you while you were away"* under the experimental condition of runkk.

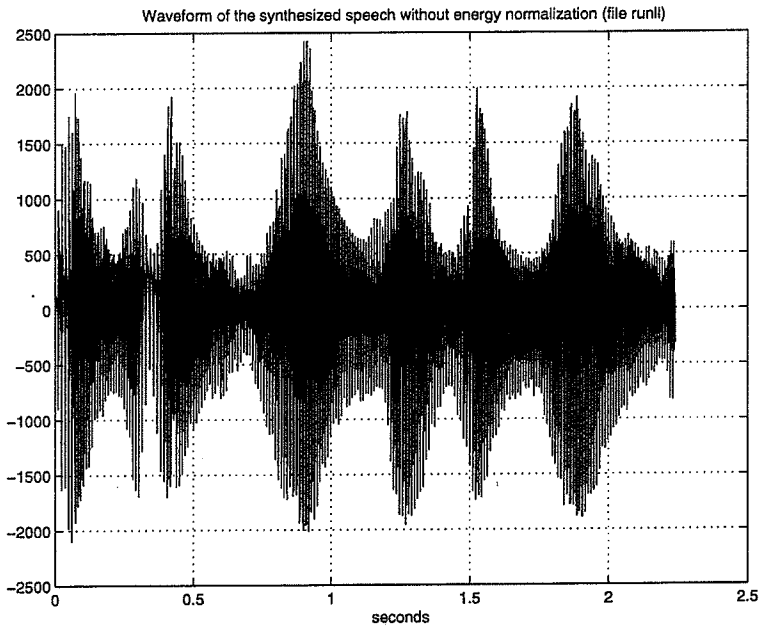


Figure 7.4 Waveform of the synthesized sentence *"Where were you while you were away"* under the experimental condition of runll.

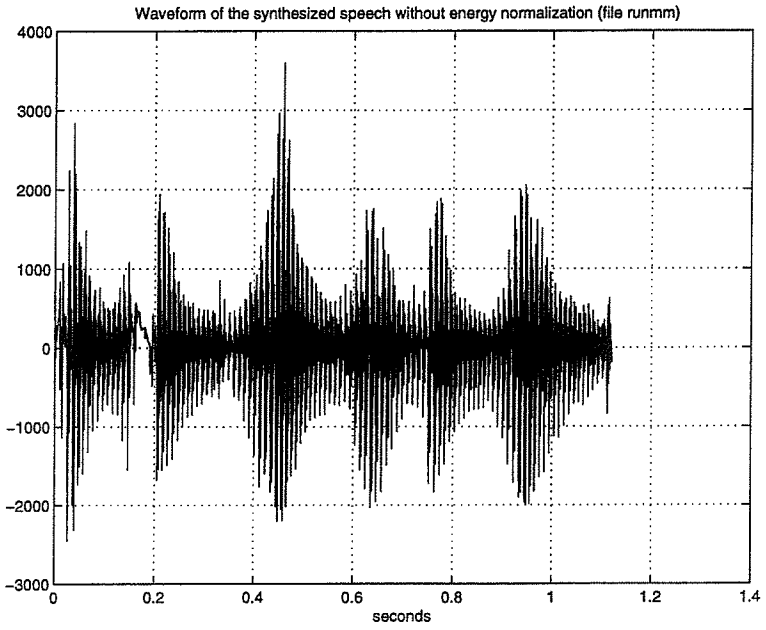


Figure 7.5 Waveform of the synthesized sentence *"Where were you while you were away"* under the experimental condition of runmm.

and the lower figure denotes the short-time power spectrum, respectively. The sampling frequency is 10 KHz.

7.4.2 Spectrogram

In this section, we present the spectrograms of the synthesized sentences *"Where were you while we were away"*. The same preemphasis and Hamming windowing approach as described in the previous section are applied. The window length is 25.6 *msec* and an 50% overlapping is applied before the time-frequency analysis. Figs. 7.7 - 7.10 show the spectrogram of the synthesized sentences under the simulation conditions of runjj, runkk, runll and runmm, respectively. The spectrogram of the original speech is shown in Fig. 7.11 for comparison.

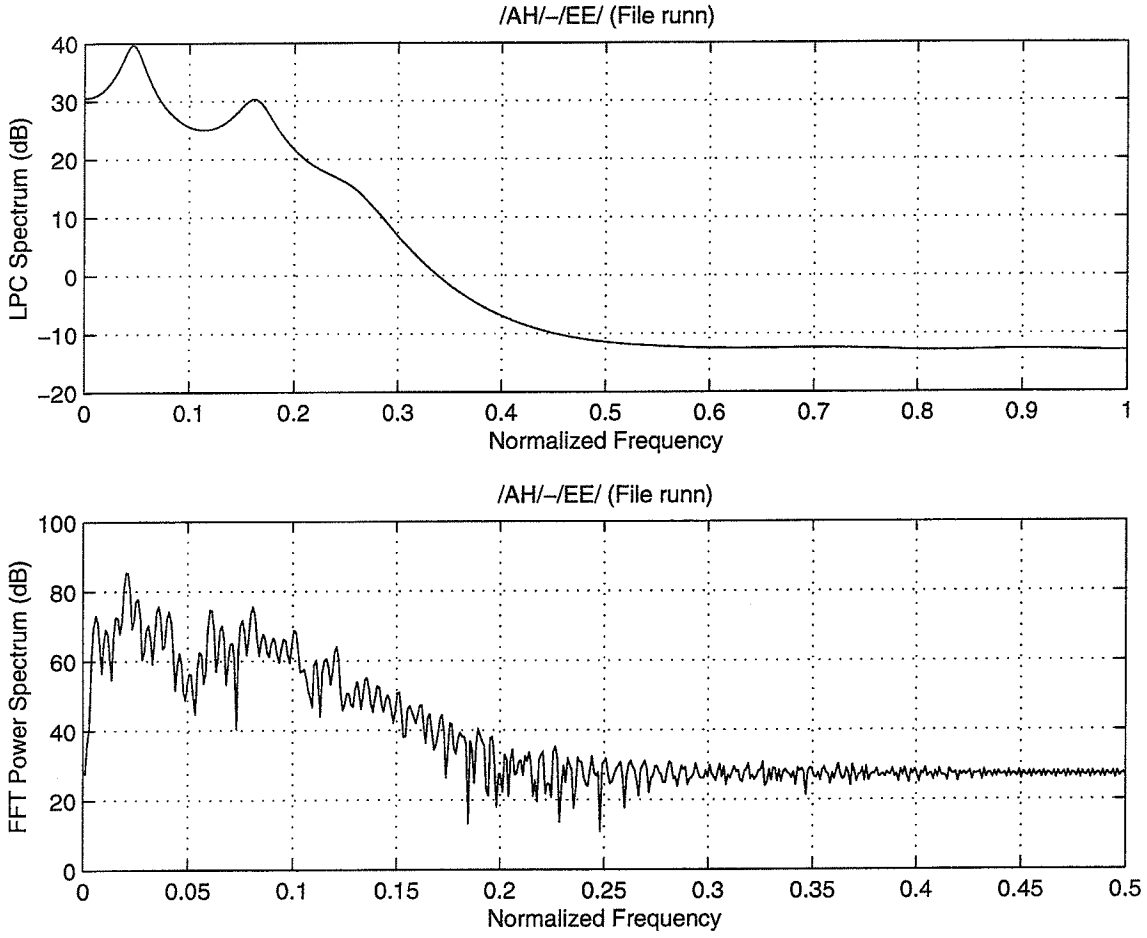


Figure 7.6 LPC spectrum and short-time power spectrum of the synthesized dipthong */AY/* ("uy" in buy).

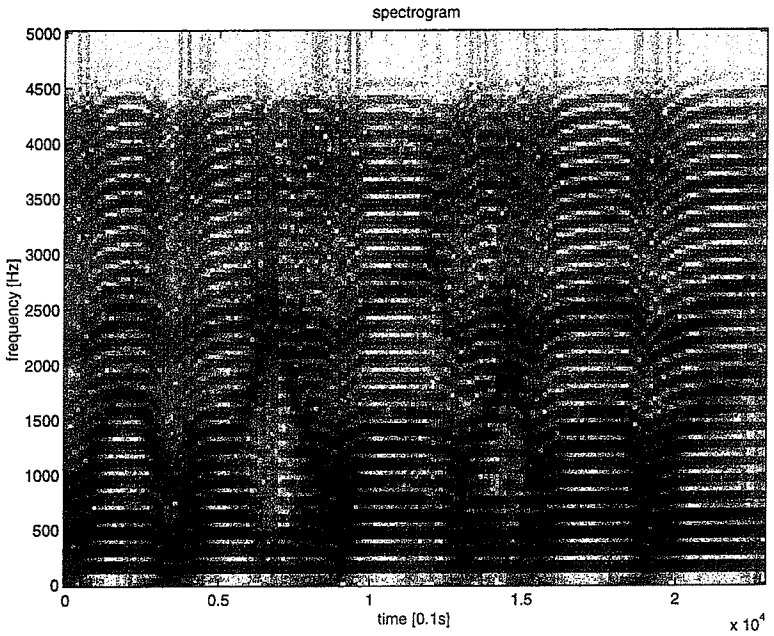


Figure 7.7 Spectrogram of the synthesized sentence *"Where were you while you were away"* under the experimental condition of runjj.

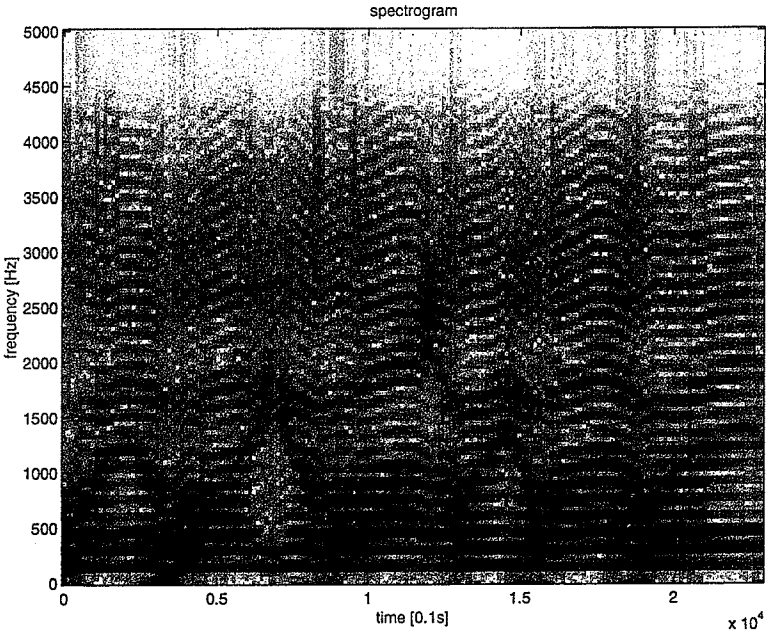


Figure 7.8 Spectrogram of the synthesized sentence *"Where were you while you were away"* under the experimental condition of runkk.

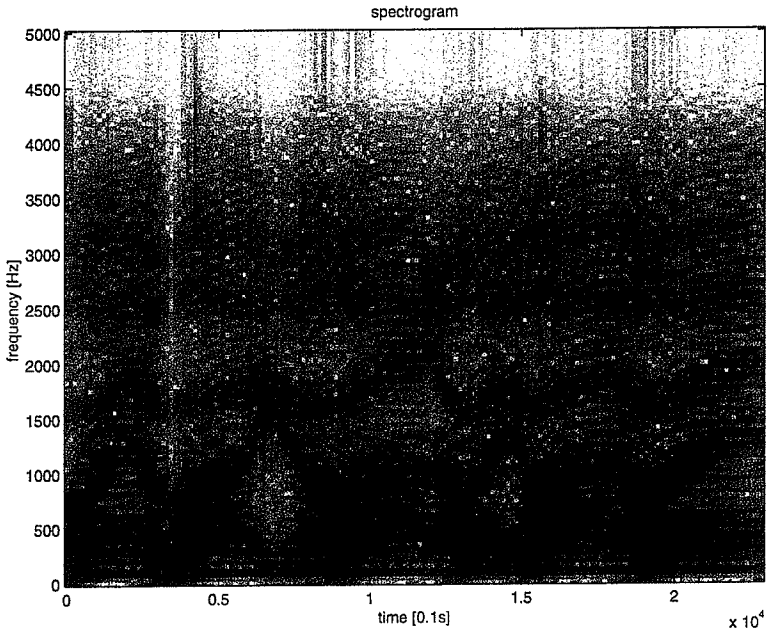


Figure 7.9 Spectrogram of the synthesized sentence *"Where were you while you were away"* under the experimental condition of runll.

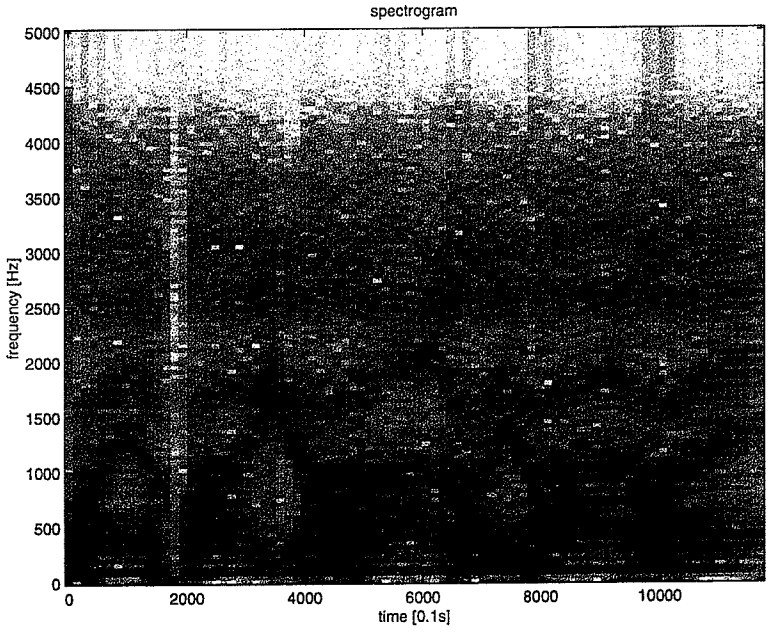


Figure 7.10 Spectrogram of the synthesized sentence *"Where were you while you were away"* under the experimental condition of runmm.

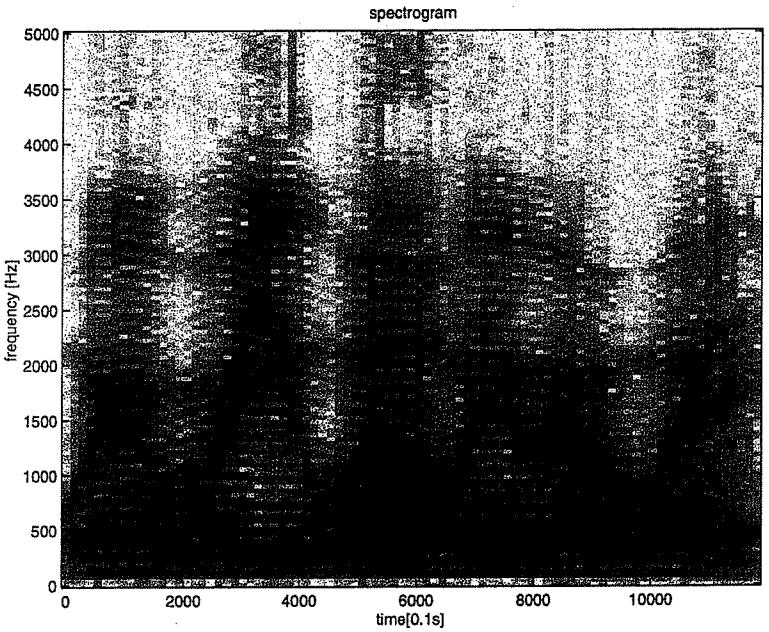


Figure 7.11 Spectrogram of the recorded sentence *"Where were you while you were away"*.

7.5 Discussion

In this chapter, we show the waveforms and the spectra analysis of the synthesized phonemes and continuous sentences using the articulatory synthesizer discussed in the previous chapters. Informal listening test shows that our articulatory synthesizer can generate good quality speech sounds. The synthesized speech signals with pitch contour information sounds more natural than the synthesized signals with monotone. In spectrogram analysis, we can see that when the pitch-contour information is included, a significant improvement can be seen from Fig. 7.9. We can also see that the articulatory synthesis has more flexibility than the conventional speech synthesis approaches. For example, our experimental results show that it's much more convenient for an articulatory synthesizer to adjust some physical parameters to achieve personalized speech synthesis. In our experiments, the synthesized speech under the conditions runll and runmm sound

muffled because in those cases vocal tract lengths (approximately 23 cm) simulating that of a very big male are used.

CHAPTER 8

CONCLUSION AND FUTURE WORK

In this thesis, a complete framework of an articulatory speech synthesizer and its relation to speech production are presented. An LS signal approximation technique was proposed to solve the motor control problem in speech production. This approach was proven to have lower error bounds than the interpolation-based methods and was successfully applied to estimate the moving vocal tract shape in articulatory synthesis. A vocal fold model based on fundamentals of mechanics and aeroacoustics was used to compute the excitation signal of vocal tract. Strong evidence of source-tract interaction was also found in the simulation results. Finally, an articulatory synthesis system was implemented to synthesize natural-sounding continuous sentences. Our finding is that articulatory synthesis has the potential to synthesize human-like speech sounds and has more flexibility than the conventional approaches. For example, articulatory synthesis uses interpolation methods naturally related to articulator movement planning to cope with the coarticulation and unit concatenation problem. The vocal fold excitation is naturally computed from a mechanical model and the aeroacoustics. On the other hand, smoothing between the unit boundaries is still a major problem in concatenative synthesis and an accurate model to represent the excitation signal is still not completely

found in formant synthesis. Furthermore, articulatory synthesizer is able to conveniently synthesize personalized speech sounds.

During the last decades, speech signal processing has experienced significant improvements along with the progress in the areas of digital signal processing, applied statistics, VLSI design, pattern recognition, artificial intelligence *et al.* However, the speech and language problem is far away from being completely solved. In [82], Rabiner and Levinson proposed an alternative view of the speech production/speech perception process. From their view, the discrete symbol information rate in the raw message text is about 50 bits per second (bps), the information rate after the language code conversation and inclusion of prosody is about 200 bps, the information rate at the neuromuscular level is about 2000 bps and finally the information rate of the acoustic signal is between 30,000 - 50,000 bps. On the other hand, we can clearly see that the current speech recognition, synthesis and coding techniques use much more information rate to process the speech signal while their performances are still much worse than those of the human being. Articulatory speech production model has the potential to lead to a more compact model of the speech signal and thus be beneficial to speech signal processing. Although limited research has been conducted in this area compared to other speech processing techniques, we can cautiously claim that investigation of speech production based on first physical principles and with the help from cross-discipline research including signal processing, statistics, fluid dynamics, aeroacoustics and aerodynamics might help us to solve some bottleneck problems encountered in the current speech signal processing.

Although articulatory synthesis and speech production modeling has the potential to benefit a wide range of problems in speech signal processing, we realize that a lot of research need to be investigated in this technique due to its unsolved problems, computational complexity and even performance. Future work in this thesis include the synthesis of higher quality sounds and the simulation and analysis of velocity and vorticity fields

(Chapter 7). Furthermore, it's found that vocal tract shape is the dominant factor in determining the spectral content of the turbulent jets in speech production [62]. If time permits, we will also investigate the construction of an articulatory model based on MRI data which will give us better detailed information about the vocal tract geometry. Moreover, a lot of important problems need to be solved after this thesis work. Some interesting topics are listed as following.

1. Investigation of the effects of non-symmetric, moving, compliant walls of the vocal tract on speech generation.
2. A much more simplified but still effective computational model which considers both the convective and the propagative modes during speech production.
3. Inclusion of the nasal tract in an articulatory model.
4. Investigation of relations between different speaking styles such as stress, accent and the articulatory parameters and its application to personalized speech synthesis.
5. Investigation of the solution of the RANS equations when there is a closure of vocal tract.
6. Investigation of the solution of the RANS equations when there is a closure of vocal cord.
7. Inclusion of the shear force due to fluid dynamics.

REFERENCES

[1] Anthony J. and Lawrence W., "A resonance analogue speech synthesizer", *Proc. 4th Int. Cong. Acoust.*, Copenhagen, Paper G43, pp. 1 - 2, 1962.

[2] Atal B. S. and Hanauer L., "Speech analysis and synthesis by linear prediction of the speech wave", *J. Acoust. Soc. Am.*, Vol. 50, pp. 637 - 655, 1971.

[3] Baer T., Gore J. C., Gracco L. C. and Nye P. W., "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels", *J. Acoust. Soc. Am.*, Vol. 90, No. 2, pp. 799 - 828, 1991.

[4] Baldi P. and Hornik K., "Neural networks and principal component analysis: Learning from examples without local minima," *Neural Networks*, Vol. 2, pp. 53 - 58, 1989.

[5] Bellegarda J. R., Silverman K. E. A., Lenzo K. and Anderson V., "Statistical prosodic modeling: from corpus design to parameter estimation", *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 1, pp. 52 - 66, 2001.

[6] Berry D. A., Herzel H., Titze I. R. and Krischer K., "Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions", *J. Acoust. Soc. Am.*, Vol. 95, pp. 3595 - 3604, 1994.

[7] Berry D. A. and Titze I. R., "Normal modes in a contunuum model of vocal fold tissues", *J. Acoust. Soc. Am.*, Vol. 100, pp. 3345 - 3354, 1996.

[8] Beutnagel M., Conkie A., Schroeter J., Stylianou Y. and Syrdal A., "The AT&T next-gen TTS system," *Joint Meeting of ASA, EAA and DAGA*, Berlin, Germany, March 15 - 19, 1999.

[9] Blackburn C. S. and Young S., "A self-learning predictive model of articulator movements during speech production", *J. Acoust. Soc. Am.*, Vol. 107, No. 3, pp. 1659 - 1670, 2000.

[10] Bourlard H. and Kamp Y., "Autoassociation by the multi-layer perceptrons and singular value decomposition," *Biological Cybernetics*, Vol. 59, pp. 291 - 294, 1988.

[11] Carreira-Perpinan M. A. and Renals S., "Dimensionality reduction of electropalatographic data using latent variable models," *Speech Communication*, Vol. 26, pp. 259 - 282, 1998.

[12] Chan R. W. and Titze I. R., "Viscoelastic shear properties of human vocal fold muscosa: theoretical characterization based on constritutive modeling", *J. Acoust. Soc. Am.*, Vol. 107, No. 1, pp. 565 - 580, 2000.

[13] Chomsky N. and Halle M., *The sound pattern of English*, New York: Harper & Row, 1968.

[14] Coker C. H., "A model of articulatory dynamics and control," *Proc. IEEE*, Vol. 64, pp. 452 - 460, 1973.

[15] Conkie A., "Robust unit selection system for speech synthesis", *Joint Meeting of ASA, EAA and DAGA*, Berlin, Germany, March 15 - 19, 1999.

[16] Cranen B. and Boves L., "On the measurement of glottal flow", *J. Acoust. Soc. Am.*, Vol. 84, No. 3, pp. 888 - 900, 1988.

[17] Currie I. G., *Fundamental Mechanics of Fluids*, McGraw-Hill Inc., New York, 1974.

[18] de Vries M. P., Schuttle H. K. and Verkerke G. J., "Determination of parameters for lumped parameter models of the vocal folds using a finite-element method approach", *J. Acoust. Soc. Am.*, Vol. 106, No. 6, pp. 3620 - 3628, 1999.

[19] Dempster A. P., Laird N. M. and Rubin D. B., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, Vol. B 39, pp. 1 - 38, 1977.

[20] Dudley H., Riesz R. R. and Watkins S. A., "A synthtic speaker," *J. Franklin Inst.*, Vol. 227, pp. 739 - 764, 1939.

[21] Dudley H., "The vocoder," *Bell laboratories record*, Vol. 17, pp. 122 - 126, 1939.

[22] Dutoit T., *An introduction to text-to-speech synthesis*, Kluwer Academic Publishers, Boston, 1997.

[23] Engwall O., "Modeling of vocal tract in three dimensions," *Proceedings of Eurospeech*, Budapest, Hungary, pp. 113 - 116, Sept. 1999.

[24] Espy-Wilson C. Y., Boyce S. E., jackson M., Narayanan S. and Alwan A., "Acoustic modeling of American English /r/", *J. Acoust. Soc. Am.*, Vol. 108, No. 1, pp. 343 - 356, 2000.

[25] Everitt B. S., *An introduction to latent variable models, monographs on statistics and applied probability*, Chapman and Hall, London, 1984.

[26] Fant G., "Acoustic analysis and systhesis of speech with applications to swedish," *Ericsson Technics*, Vol. 15, pp. 1615 - 1626, 1959.

[27] Fant G., *Acoustic theory of speech production*, Mouton, the Hague, 1970.

[28] Fant G. and T. V. Ananthapadmanabha, "Speech production", *STL-QPSR*, Vol. 2, pp. 1 - 17, 1982, Speech Transmission Laboatory, Royal Institute of Technology, Stockholm.

[29] Fant G., Lin Q. and Gobl C., "Notes on glottal flow interaction", *Speech Transmission Laboratory - Quarterly Progress and Status Report*, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Vol. 2, pp. 21 - 45, 1985.

[30] Fant G. and Lin Q., "Glottal source - vocal tractacoustic interaction", *STL-QPSR*, Speech Transmission Laboatory, Royal Institute of Technology, Stockholm, Vol. 1, pp. 13 - 27, 1987.

[31] Flanagan J. L., Coker C. H., Rabiner L. R., Schafer R. W. and Umeda N., "Synthetic voices for computers," *IEEE Spectrum*, Vol. 7, No. 10, pp/ 22 - 45, 1970.

[32] Flanagan J. L., "Voices of men and machines," *J. Acoust. Soc. Am.*, Vol. 51, pp. 1375 - 1387, 1972.

[33] Flanagan J. L., *Speech analysis, synthesis, and perception*, Springer-Verlag, New York, 2nd edition, 1972.

[34] Flanagan J. and Rabiner L. R., Editors, *Speech synthesis*, Dowden, Hutchington and Ross Inc, 1973.

[35] Flanagan J. L., Ishizaka K., and Shipley K. L., "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," *Bell Sys. Tech. J.*, Vol. 45, No. 3, pp. 485 - 506, 1975.

[36] Hardcastle W. J. and Marchal A., Editors, *Speech production and speech modelling*, Kluwer Academic Publishers, Boston, 1990.

[37] Harnsberger J. D., "A cross-language study of the identification of non-native nasal consonants varying in place of articulation", *J. Acoust. Soc. Am.*, Vol. 108, No. 2, pp. 764 - 783, 2000.

[38] Harshman R., Ladefoged P. and Goldstein L., "Factor analysis of tongue shapes," *Journal of the Acoustical Society of America*, Vol. 62, pp. 693 - 707, 1977.

[39] Hasegawa-Johnson M. A. and Cha J. S., "CTMRedit: A Matlab-based tool for viewing, editing and three-dimensional reconstruction of MR and CT images," *Proc. BMES/EMBS*, Atlanta, FL, 1999.

[40] Hasegawa-Johnson M. A., "Line spectral frequencies are poles and zeros of the glottal driving impedance of a discrete matched-impedance vocal tract model", *J. Acoust. Soc. Am.*, Vol. 108, No. 1, pp. 457 - 460, 2000.

[41] Hasegawa-Johnson M. A., *Lecture notes in speech production, speech coding, and speech recognition*, University of Illinois at Urbana-Champaign, 2000.

[42] Holmes J. N., "An investigation of the volume velocity waveform at the larynx during speech by means of an inverse filter", *Proc. Speech Communication Seminar*, Royal Institute of Technology, Stockholm, Sweden, Vol. 1 pp. B4, 1962.

[43] Holmes J. N., Mattingly I. G. and Shearme J. N., "Speech synthesis by rule," *Language and speech*, Vol. 7, No. 3, pp. 127 - 143, 1964.

[44] Holmes J. N., "The influence of glottal waveform on the naturalness of speech from a parallel-formant synthesizer," *IEEE Trans. Audio, Elctrroacoust.*, Vol. 21, pp. 298 - 305, 1973.

[45] Holmes J. N., "Formant synthesizers: cascade or parallel," *Speech Communication*, Vol. 2, pp. 251 - 273, 1983.

[46] Holmes W. J., Holmes J. N. and Judd M. W., "Extension of the band-width of the JSRU parallel-formant synthesizer for high quality synthesis of male and female speech," *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, Vol. 1, pp. 313 - 316, 1990.

[47] Holst T., Warren P. and Nolan F., "Categorising [*s*], [*ʃ*] and intermediate electropalographic patterns: Neural networks and other approaches," *European Journal of Disorders of Communication*, Vol. 30, pp. 161 - 174, 1995.

[48] Howe M. S., "The genaration of sound by aerodynamic sources in an inhomogeneous steady flow", *J. Fluid Mechanics*, Vol. 67, part 3, pp. 597 - 610, 1975.

[49] Howe M. S., *Acoustics of Fluid-Structure Interactions*, Cambridge University Press, Cambridge, 1998.

[50] Huang J. and Levinson S. E., "Estimation of articulatory movement and its application to speech synthesis", *Proc. of 138th meeting of the Acoust. Soc. Am*, Columbus, Ohio, November 1-5, 1999.

[51] Huang J. , Levinson S. E., Silmon S. and Davis D., "Articulatory speech synthesis and the analysis of boundary conditions", *Proc. of 139th meeting of the Acoust. Soc. Am*, Atlanta, Georgia, May 30 - June 3, 2000.

[52] Huang J. , Levinson S. E., and Hasegawa-Johnson M. A., "Signal Approximation in Hilbert Space and its application on articulatory speech synthesis', *Proc. ICSLP*, Beijing, China, Oct. 16 - 20, 2000.

[53] Huang T. S. and Netravali N. N., "Motion and structure from feature correspondences: a review", *Proc. of IEEE*, Vol. 82, No. 2, pp. 252 - 268, 1994.

[54] Ishizaka K. and Flanagan J. L., "Synthesis of voiced sounds from a two-mass model of the vocal cords", *Bell Sys. Tech. J.*, Vol. 51, No. 6, pp. 1233 - 1268, 1972.

[55] Kaburagi T., Honda M. and Okadome T., "A trajectory formation model of articulatory movements using a multidimensional phpnemic task," *Proceedings of Eurospeech*, Budapest, Hungary, pp. 121 - 124, Sept. 1999.

[56] Kelley J. L. and Gerstman L. J., "An artificial talker driven from phonetic input," *J. Acoust. Soc. Am.*, Vol. 33, p. 835(A), 1961.

[57] Kelley J. L. and Lochbaum C. C., "Speech synthesis", *Proc. 4th Int. Congr. Acoust.*, Copenhagen, Paper G-42, pp. 1 - 4, 1962.

[58] Klatt D. H., "Acoustic theory of terminal analog speech synthesis," *Proc. IEEE Intl. Conf. on Acoust, Speech, Signal Processing*, Vol. 1, pp. 131 - 135, 1972.

[59] Klatt D. H., "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, Vol. 67, pp. 971 - 995, 1980.

[60] Klatt D., "Review of text-to-speech conversion for English," *J. Acoust. Soc. Am.*, Vol. 82, pp. 737 - 793, 1987.

[61] Koizumi Y., Taniguchi S. and Hiromitsu S., "Two-mass models of the vocal cords for natural sounding voice synthesis", *J. Acoust. Soc. Am.*, Vol. 82, pp. 1179 - 1192, 1987.

[62] Krane M. H., "Aeroacoustic production of unvoiced speech sounds," submitted to Journal of the Acoustical Society of America.

[63] Lawrence W., "The synthesis of speech from signals which have a low information rate", in Jackson W. ed., *Communication Theory*, Butterworths Science Publication, London, pp. 460 - 469, 1953.

[64] Levinson S. E. and Schmidt C. E., "Adaptive computation of articulatory parameters from the speech signal," *J. Acoust. Soc. Am.*, 74(4), pp. 1145 - 1154, Oct. 1983.

[65] Liberman A. M., Ingeman F., Lisker L., Delattre P. and Cooper F. S., "Minimal rules for synthesising speech", *J. Acoust. Soc. Am.*, Vol. 31, pp. 1490 1499, 1959.

[66] Liljencrants J., "Speech synthesis with a reflection-type analog", Ph.D. Thesis, Royal Institute of Technology, Stockholm, Sweden, 1985.

[67] Maeda S., "An articulatory model of the tongue based on a statistical analysis," in Wolf J. J., and Klatt D. H. (Eds.), *Speech communication papers*, Acoustical Society of America, New York, pp. 67 - 70, 1979.

[68] Maeda S., "A digital simulation method of the vocal-tract system", *Speech Communication*, Vol. 1, pp. 199 - 229, 1982.

[69] Makhoul J., "Linear prediction: a tutorial review", *Proc. of IEEE*, Vol. 63, No. 4, pp. 561 - 580, 1975.

[70] Markel J. D. and Gray H., "On autocorrelation equations as applied to speech analysis", *IEEE Trans. Audio Electroacoust.*, Vol. 20, pp. 69 - 79, 1973.

[71] Mermelstein P., "Articulatory model for the study of speech production," *Journal of the Acoustical Society of America*, Vol. 53, pp. 1070 - 1-82, 1973.

[72] Meyer P., Wilhelms R. and Strube H. W. A., "Quasiarticulatory speech synthesizer for German language running in real time", *J. Acoust. Soc. Am.*, Vol. 86, No. 2, pp. 523 - 539, 1989.

[73] Murphy P. J., "Spectral characterization of jitter, shimmer, and additive noise in synthetically generated voice signals", *J. Acoust. Soc. Am.* Vol. 107, No. 2, pp. 978 - 988, 2000.

[74] Nelson W. L., "Physical principles for economies of skilled movements," *Journal of Biological Cybernetics*, Vol. 46, pp. 135 - 147, 1983.

[75] Nguyen N., Hoole P. and Marchal A., "Regenerating the spectral shape of [s] and [$\int$] from a limited set of articulatory parameters," *Journal of the Acoustical Society of America*, Vol. 96, pp. 33 - 39, 1994.

[76] Nguyen N., Marchal A. and Content A., "Modeling of tongue-plate contact patterns in the production of speech," *Journal of Phonetics*, Vol. 24, pp. 77 -97, 1996.

[77] N. Nguyen, "EPG bidimensional data reduction," *European Journal of Disorders of Communication*, Vol. 30, pp. 175 -182, 1995.

[78] O'Brien D. and Monaghan A. I. C., "Concatenative synthesis based on a harmonic model", *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 1, pp. 11 - 20, 2001.

[79] Paget R., *Human speech: some observations, experiments and conclusions as to the nature, origin, purpose and possible improvement of human speech*, Harcourt, London, 1920.

[80] Perkell J. S. and Klatt D. H., eds. *Invariance and Variability in Speech Processes*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[81] Rabiner L. R. and Schafer R. W., *Digital Processing of Speech Signals*, Prentice Hall, Englewood Cliffs, 1978.

[82] Rabiner L. R. and Levinson S. E., "Isolated and connected word recognition - theory and selected applications", *IEEE trans. Communications*, Vol. 29, No. 5, pp. 621 - 659, 1981.

[83] Rabiner L. R. and Juang B. H., *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, 1993.

[84] Rahim M. G. and Goodyear C. C., "Parameter estimation for spectral matching in articulatory synthesis," *Collq. Spectral Est. Tech. Speech Processing*, IEE Press, London, 1989.

[85] Rahim M. G. and Goodyear C. C., "Articulatory speech with the aid of a neural net," *Proc. IEEE Intl. COnf. on Acous., Speech, Signal Processing*, Vol. 1, pp. 227 - 230, 1989.

[86] Rahim M. G. and Goodyear C. C., "Estimation of vocal tract filter parameters using a neural net," *Speech Communication*, Vol. 9, pp. 49 - 55.

[87] Rahim M. G., Kleijin W. B., Schroeter J. and Goodyear C. C., "Acoustic to articulatory parameter mapping usng an assembly of neural networks," *Proc. Int. Conf. on Acoustic., Speech, Signal Processing*, Vol. 1, pp. 485 - 488, 1991.

[88] Rahim M. G., *Artificial neural networks for speech analysis/synthesis*, Chapman & Hall, London, 1994.

[89] Richard G., Liu M., Sinder D., Duncan H., Lin Q., Flanagan J., Levinson S., Davis D. and Slimon S., "Numerical simulations of fluid flow in the vocal tract", *Proc. Eurospeech*, Madrid, Spain, pp. 1297 - 1300, 1995.

[90] Roe D. B. and Wilpon J. G., *Voice Communication between Human and Machines* National Academy Press, Washington D. C., 1994.

[91] Rosenberg A. E., "Effects of pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am.*, Vol. 49, No. 2, pp. 583 - 591, 1973.

[92] Rothenberg M., "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," *J. Acoust. Soc. Am.*, Vol. 53, No. 6, pp. 1632 - 1654. 1973.

[93] Rubin P., Baser T. and Mermelstein P., "An articulatory synthesizer for perceptual research", *J. Acoust. Soc. Am.*, Vol. 70, No. 2, pp. 321 - 328, 1981.

[94] Rubin P., Saltzman E., Goldstein L., McGowan R., Tiede M. and Browman C., "CASY and extensions to the task-dynamic model", *1st ECSA Tutorial and Research Workshop on Speech production Modeling - 4th Speech Production Seminar*, pp. 125 - 128, 1996.

[95] Saltzman E. L. and Munfall K. G., "A dynamical approach to gestural patterning in speech production," *Haskin Laboratories Status Rerport on Speech production*, SR-99/100, pp. 38-68, 1989.

[96] Saltzman E. L., "The task dynamic model in speecg production", in Peters H., Hulstijn W. and Starkweather C. W., eds. *Speech motor control and shuttering*, Elsevier Science Publishers, New York, pp. 37 - 52, 1991.

[97] van Santen J. P. H., Sproat R., Olive J. P., and Hirschberg J. editors, *Progress in speech synthesis*, Springer, New York, 1997.

[98] Schroeter J. and Sondhi M. M., "Dynamic programming search of articulatory codebooks," *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, Vol. 1, pp. 588 - 591, 1989.

[99] Schroeter J., Meyer P., and Parthasarathy S. "Evaluation of improved articulatory codebooks and codebook access distance measures," *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, Vol. 1, pp. 393 - 396, 1990.

[100] Schroeter J. and Sondhi M. M., "Speech coding based on physiological models of speech production," in Furui S. and Sondhi M. M. Eds., *Advances in Speech Signal Processing*, Marcel Dekker Inc., New York, 1991.

[101] Schroeter J. and Sondhi M. M., "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 1, pp. 133 - 150, 1994.

[102] Shadle C. H., "The effect of geometry on source mechanisms of fricative consonants," *Journal of Phonetics*, Vol. 19, pp. 409 - 424, 1991.

[103] Shirai K. and Honda M., "An articulatory model and the estimation of articulatory parameters by nonlinear regression method," *Electron. Comm.*, Japan 59-A, pp. 35 -43, 1976.

[104] Sinder D., Richard G., Duncan H., Flanagan J., Slimon S., Davis D., Krane M. and Levinson S., "Flow visualization in stylized vocal tracts", *Proc. ASVA*, Tokyo, Japan, 1997.

[105] Sinder D., "Speech synthesis using an aeroacoustic model," *Technical report CAIP-TR-236*, CAIP center of Rutgers University, 1999.

[106] Slimon S., Davis D., Levinson S., Krane M., Richard G., Sinder D., Duncan H., Lin Q. and Flanagan J., "Low mach number flow through a constricted, stylized vocal tract", *Proc. Amer. Inst. Aeronautics and Astronautics Conf.*, 1996.

[107] Sokolnikoff I. S., *Tensor Analysis: Theory and Applications to Geometry and Mechanics of Continua*, John Wiley & Sons, New York, 2nd edition, 1964.

[108] Sondhi M. M., "Model for wave propagation in a lossy vocal tract," *J. Acoust. Soc. Am.*, Vol. 55, No. 5, pp. 1070 -1075, 1974.

[109] Sondhi M. M., "Estimation of vocal-tract areas: the need for acoustical measurements," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 27, No. 3, pp. 268 - 273, 1979.

[110] Sondhi M. M. and Schroeter J., "A hybrid time-frequency domain articulatory speech synthesizer", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 35, No. 7, pp. 955 - 967, 1987.

[111] Sorokin V. N., Leonov A. S. and Trushkin A. V., "Estimation of stability and accuracy of inverse problem solution for the vocal tract," *Speech Communication*, Vol. 30, pp. 55 - 74, 2000.

[112] Sorokin V. N., "Determination of vocal tract shape for vowels," *Speech communication*, Vol. 11, pp. 71 -85, 1992.

[113] Stevens K. N., "Airflow and turbulence noise for fricative and stop consonants: static considerations," *Journal of the Acoustical Society of America*, Vol. 50, pp. 1180 -1192, 1971.

[114] Stevens K. N., "On the quantal nature of speech," *Journal of Phonetics*, Vol. 17, pp. 3 - 45, 1989.

[115] Story B. H. and Titze I. R., "Voiced simulation with a body-cover model of the vocal folds", *J. Acoust. Soc. Am.*, Vol. 97, pp. 1249 - 1260, 1995.

[116] Strang G. and Fix G., "A Fourier analysis of the finite element variational method", in *Constructive Aspect of Functional Analysis*, Cremonese, Rome, Italy, pp. 796 - 830, 1971.

[117] Stylianou Y., "Removing phase mismatches in concatenative speech synthesis", 3^{rd} *ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, Australia, Nov. 1998.

[118] Stylianou Y., "Concatenative speech synthesis using a harmonic plus noise model," 3^{rd} *ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, Australia, Nov. 1998.

[119] Stylianou Y., "Applying the harmonic plus noise model in concatenative speech synthesis", *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 1, pp. 21 - 29, 2001.

[120] Taylor P., "Analysis and synthesis of intonation using the Tilt model", *J. Acoust. Soc. Am.*, Vol. 107, No. 3, pp. 1697 - 1714, 2000.

[121] Teager H. M., "Some observations on oral air flow during phonation," *IEEE Trans. on ASSP*, Vol. ASSP-28, pp. 599 - 601, 1980.

[122] Teixeira A., Vaz F. and Principe C., "Effects of source-tract interaction in perception of nasality", *Proc. Eurospeech*, Budapest, Hungary, Sept. 1999.

[123] Teman R., *Navier-Stokes Equations*, North-Holland Publishing, New York, 1977.

[124] Thimm G. and Luettin J., "Extraction of articulators in X-ray image sequences," *Proceedings of Eurospeech*, Budapest, Hungary, pp. 157 - 160, Sept. 1999.

[125] Tipping M. E. and Bishop C. M., "Mixtures of principal component analysis," *IEE Fifth International Conference on Artificial Neural Networks*, IEE, London, England, 1997.

[126] Titze I. R., "Synthesis of sung vowels using a time-domain approach," in Lawrence V. L. ed., *Transcripts of the 11th symp.: Care of the Prof. Voice*, The Voice Foundation, New York, pp. 90 - 98, 1982.

[127] Titze I. R., Ed., *Vocal fold physiology: frontiers in basic science*, Singular Publishing Group, San Diego, 1993.

[128] Titze I. R., *Principles of voice production*, Prentice Hall, Englewood Cliffs, 1994.

[129] Titze I. R. and Story B. H., "Comparison between electroglottography and electromagnetic glottography", *J. Acoust. Soc. Am.* Vol. 107, No. 1, pp. 581 - 588, 2000.

[130] M. Unser and I. Daubechies, "On the approximation power of convolution-based least squares versus interpolation", *IEEE Trans. on Signal Processing*, Vol. 45, No. 7, pp. 1697 - 1711, July 1997.

[131] T. Blu and M. Unser, "Quantitative Fourier analysis of approximation techniques: part I – interpolators and projectors", *IEEE Trans. on Signal Processing*, Vol. 47, No. 10, pp. 2783 - 2795, October 1999.

[132] Watika H., "Direct estimation of the vocal tract shape by inverse filtering of the acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.*, Vol. 21, pp. 417 - 427, 1973.

[133] Wouters J. and Macon M. W., "Control of spectral dynamics in concatenative speech synthesis", *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 1, pp. 20 - 38, 2001.

[134] Wrede R. C., *Introduction to Vector and Tensor Analysis*, John Wiley & Sons, New York, 1963.