Speech Coding and Synthesis Using Parametric Curves

Luis Miguel Teixeira de Jesus

A Thesis Submitted for the Degree of Master of Science by Research

School of Information Systems
University of East Anglia

October 1997

©This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognize that its copyright rests with the author and that no quotation from the thesis, nor any information derived therefrom, may be published without the author's prior written consent.

Abstract

Accurate modelling of co-articulation, the context-sensitive merging of the boundaries between allophones in continuous speech, is vital for natural sounding speech synthesis. This thesis describes research investigating the use of Bézier Curves to form models of the effects of co-articulation in human speech. The trajectories of line spectral pair (LSP) parameters through each diphone are represented by cubic Bézier curve segments, found using the Levenberg-Marquardt curve fitting method. The Bézier curve is found to provide an effective model of the transitions between a majority of speech sounds, however a more complex model is found to be needed where abrupt or complex transitions are required, for instance during plosives.

Contents

Ao Desconcerto do Mundo

Redondilhas

Os bons vi sempre passar
No mundo graves tormentos;
E para mais m'espantar,
Os maos vi sempre nadar
Em mar de contentamentos.
Cuidando alcançar assi
O bem tão mal ordenado,
Fui mao; mas fui castigado.
Assi, que só para mi
Anda o mundo concertado.

in "Obras completas de Luis de Camões", Correctas e Emendadas Pelo Cuidado e Diligencia de J. V. Barreto Feio e J. G. Monteiro. Tomo Terceiro. Hamburgo na Officina Typographica de Langhoff, 1834.

Acknowledgements

Aos meus pais e irmão que sempre acreditaram mesmo nos momentos de maior desespero.

I would like to thank my supervisor Dr. Gavin C. Cawley for his advice, patience and friendship.

I also would like to give thanks to Stephen Cox for his support and encouragement.

The author would like to thank Dr. Andrew Breen at the British Telecommunications Laboratories at Martlesham Heath, Ipswich, for providing the speech data used in this research.

Para os meus pais e irmão.

Chapter 1

Introduction

Speech is the most natural form of communication between humans as it can convey an almost infinite range of thoughts and concepts. With their finite vocabulary humans can produce a large number of utterances each of which will be understood by another person with knowledge of the language that is being used. That person may have never heard some utterances before but will be able to extract its meaning and even gain some understanding of totally novel words from context. Science has been interested in modelling the physiology and acoustics of speech since the seventeenth century. One of the first attempts to synthesize speech is that of Kratzenstein (1769) [38], who used five acoustic resonators, excited by a reed to produce vocalic sounds. Since then a great deal of progress has been made but we can clearly establish some turning points in the history of speech science. The first steps in the acoustic and electric theory lead to the implementation of the first electric synthesizer by Stewart (1922) [55], a departure from the mechanical models used until then. Spectrograms were used to obtain relevant data such as the frequency, bandwidth and amplitude of formants [1] in speech sounds. The development of electronic computers and electronic circuits during 1960s lead to remarkable advances in speech analysis and synthesis techniques. Important advances in digital speech processing technology have recently opened the way to a new generation of analysis and synthesis systems.

Much attention has been devoted in modern systems to the quality of syn-

[1] The peaks of the speech sound spectrum (resonances) are called formants.

thetic speech. The production of naturally sounding speech is now the primary concern, rather than merely intelligible speech. Speech synthesis systems are able to reproduce several types of voice with specific characteristics such as those of female and children speakers. A major factor to further improvements in synthesis is the enormous variability of speech sounds, largely due to the effect of co-articulation caused by the inertia of articulators such as the lips, tongue and jaw. For example, the merging of the boundaries between allophones occurs in "the toucan" ("D@ tuk=n") because the lip rounding for "u" starts while "D@" is still being articulated.

1.1 Research Overview

This research aims to investigate the use of parametric curve fitting techniques to form a data-driven model of the effects of co-articulation, based on cubic Bézier segments. The trajectories of each line spectral pair parameter between a pair of phonemes can be represented by cubic Bézier segments, as shown in Figure 1.1, forming a model of the effects of co-articulation within a particular diphone. The parameters controlling the shape of the Bézier segments are determined using a least squares curve fitting procedure over examples of each diphone extracted from a phonetically annotated corpus of human speech. The aim of this research is to determine the accuracy in representing the trajectory of LSP parameters of Bézier models of co-articulation compared with a conventional linear template based approach.

A single cubic Bézier curve is found to be an effective model of the trajectory for line spectral pair parameters during the transitions between a majority of speech sounds. Where abrupt or complicated transitions are required a more complex model is needed. A suitable strategy must then be developed to blend adjacent Bézier segments together, according to phonetic context, to provide a simple model of the effects of co-articulation that extend beyond diphone boundaries. Initial results of this work were presented at the EuroSpeech'97 conference [30].

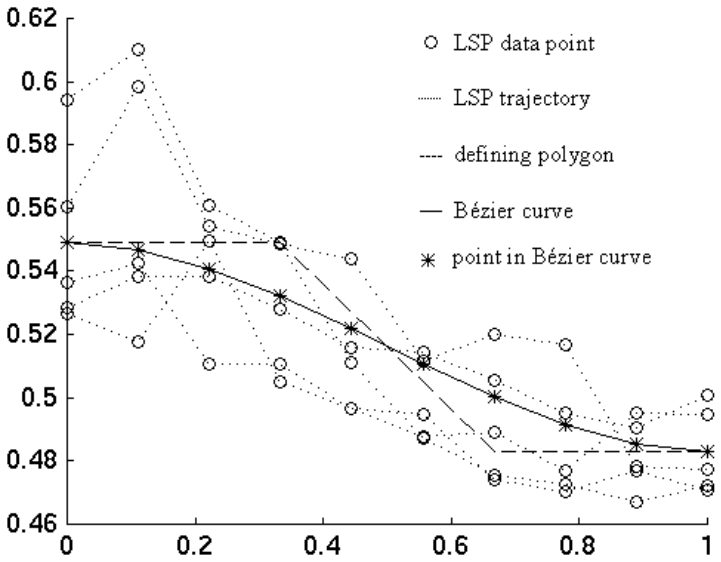

Figure 1.1: The seventh LSP parameter trajectory of diphone "laU" (all examples in the speech corpus) and the cubic Bézier model.

1.2 Thesis Overview

The remainder of this thesis is arranged as follows:

- Chapter 2 describes the human speech production system and introduces the concept of co-articulation. Some vocal tract models and speech synthesis techniques are presented. The chapter also includes a literature review of conventional methods for modelling co-articulation. Case studies of some concatenative synthesis and synthesis by rule systems are also provided.

- Chapter 3 discusses the proposed Bézier model of co-articulation. The reader is also provided with background information on Bézier curves and curve fitting methods.

- Chapter 4 describes the speech corpus and discusses the data preprocessing procedure used in the research presented in this thesis.

- Chapter 5 presents the detailed method used to model diphones using

cubic Bézier segments and describes a basic strategy for blending adjacent diphones. Alternative interpolation templates are also presented. A model based on the Holmes-Mattingly-Shearme scheme is used for comparison.

- Chapter 6 presents results obtained using both Bézier and Holmes-Mattingly-Shearme models. Results of blending adjacent diphones and first attempts to synthesize an example sentence are described.

- Chapter 7 presents the conclusions drawn from the experimental work described in the previous chapters and also presents some suggestions for future work.

- Appendix A describes a set of tools used to visualize the effects of coarticulation in speech parameters (phonemes, diphones and sentences).

- Appendix B presents detailed information about file formats and phonetic notation used in the speech corpus.

Chapter 2

Speech Production by Humans and Machines

2.1 Introduction

This chapter describes the human speech production system and the nature of speech sounds. The mechanisms of phonation and articulation of both voiced and unvoiced sounds are introduced. The manners of articulation (approximant, nasal, plosive, affricate, fricative and vowel) are described in detail.

Since the work presented in this thesis is concerned with forming models of co-articulation in human speech the concept of co-articulation is also introduced. Articulatory, formant and LPC vocal tract models used in speech synthesis systems are presented, including a description of conventional models of the effects of co-articulation.

The chapter also provides a literature review or case studies of concatenative speech synthesis and speech synthesis by rule systems. Diphone concatenation is of particular interest to this work and so systems that use this method are presented in chronological order. The Holmes-Mattingly-Shearme, MITalk and Laureate systems are described with focus on the methods used to generate the raw speech.

2.2 Human Speech Production

Speech can be described as a highly integrated and complex chain of events leading to a meaningful sequence of sounds. Figure 2.1 shows an integrated model of speech production, describing the temporal overlap, the mutual influence and the feedback between the several structures that compose the speech mechanism. A sequence of neural impulses is generated at cortical level and transmitted to the musculature of the breathing mechanism, to the larynx and to the articulators [60]. There is temporal overlap between speech mechanism structures such as the phonation and articulation, i.e. phonation is active while speech sounds are being articulated. The auditory feedback and conscious feedback from muscles provide important information for the brain to monitor the speech production process. Changes in phonation and articulation processes influence the respiration process. A practical example that illustrates the importance of an integrated speech production system is the speech of deaf people, where the feedback mechanism is severely affected and so there is little coordination between the process of phonation and the articulatory gestures. It is the coordination of these multiple events, interacting and overlapping in time, that results in the speech waveform radiated from the mouth. The following sections describe the processes of phonation and articulation that produce the various categories of speech sounds.

2.2.1 The Production of Speech Sounds

Speech is produced by the action of an acoustic filter formed by the vocal/nasal tract on an excitation signal generated within the larynx in the case of voiced speech and by a constriction of the vocal/nasal tract itself in the case of unvoiced speech. We move the articulators to change the acoustic properties of the vocal/nasal tract. The nervous system coordinates many muscles in order to produce movements of the diaphragm, larynx, tongue and lips, as shown in Figure 2.2.

2.2.1.1 Voiced Sounds

The lungs are expanded causing air to flow in to neutralize the negative pressure, then they are contracted causing air to flow out to equalize the positive

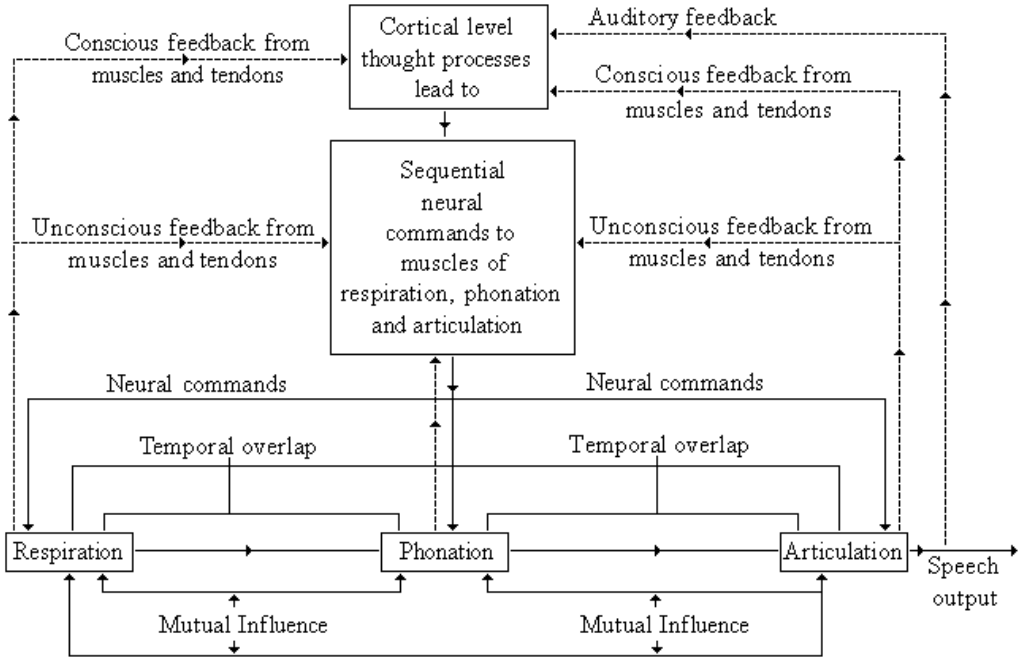

Figure 2.1: A model of speech production. After Zemlin [60].

pressure, providing the energy required to produce the acoustic excitation. The sub-glottal air pressure from the lungs, applied to the lower part of the vocal folds, opens the glottis, the folds come together again because of their inherent elasticity and sudden pressure drop between the folds due to the Bernoulli principle as the air streams through the open glottis and the cycle then repeats. Periodic sounds are created by this rapid opening and closing of the vocal folds, known as phonation.

The excitation signal for voiced sounds is produced in the larynx, which is located just above the trachea. The structure of the larynx, shown in Figure 2.3, encloses two sets of folds that lie behind the thyroid cartilage. The true vocal folds are used for phonation and the ventricular folds or false vocal folds form a second constriction. The horizontal space between the true vocal folds is called the glottis. The vocal fold tension and elasticity can be altered to effect changes in the frequency of vibration (fundamental frequency). Their opening can adjusted from close together to open wide and their length can

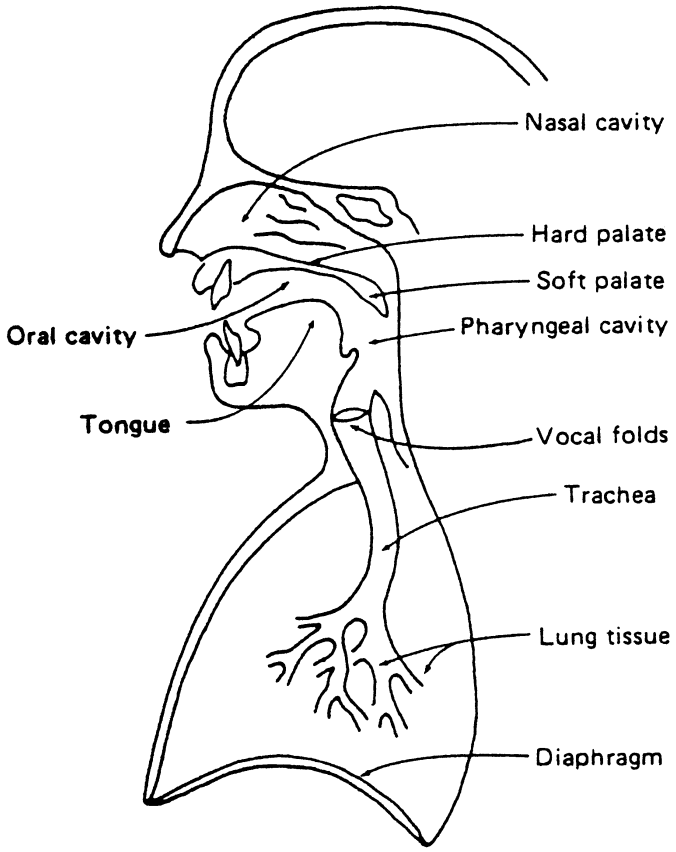

Figure 2.2: The speech mechanism. From Zemlin [60].

be varied.

Different voice qualities or timbres are produced by various modes of vocal fold vibration.

- The vocal folds are separated (abducted) in order to allow the passage of sufficient air from the lungs to produce voiceless speech sounds in the oral cavity and during normal breathing, as shown in Figure 2.4.
- In order to create voiced sounds the vocal folds are brought together (adducted) and set into vibration.

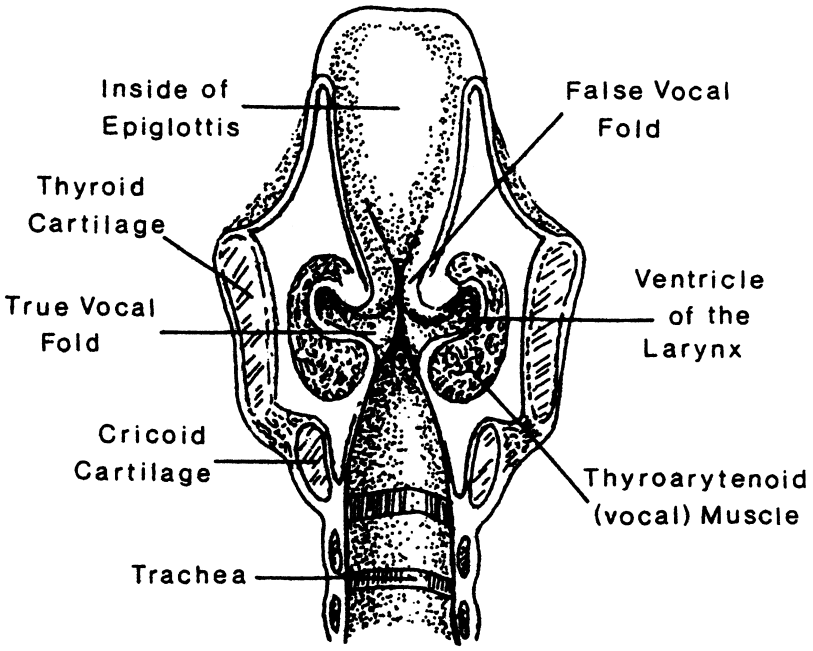

Figure 2.3: Frontal section of the larynx. From Borden, Harris and Raphael [5].

- A breathy voice is achieved by failing to adduct the vocal folds sufficiently for full voicing.
- For speech sounds requiring both periodic and aperiodic sound sources, the vocal folds are partially adducted (e.g. voiced consonants).

A hoarse voice is caused by irregularities in the vocal folds. Vocal fry or creaky voice is due to extremely low frequency phonation.

The intensity of the voice is primarily controlled by the sub-glottal pressure and so it can be varied by altering the amplitude of the sound pressure waves. During continuous speech, fundamental frequency, quality and intensity, are constantly changed by the phonation process. Different stresses are given to certain speech sounds to show emotion. They are generated by prosodic effects such as the rising pitch at the end of sentences.

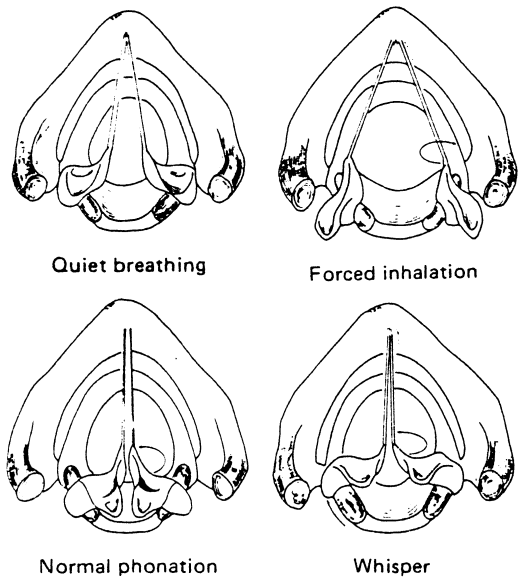

Figure 2.4: Various glottal configurations. From Zemlin [60].

2.2.1.2 Unvoiced Sounds

Continuous aperiodic sound waves are generated by partially adducting the vocal folds or by forming constrictions in the vocal tract, causing the air stream to become turbulent. For example, the constriction used to produce phoneme "f" in the word "fear" is located at the front of the mouth. The phoneme "s" in the word "sing" is produced with a constriction at the roof of the oral cavity.

The periodic sound source can be combined with the continuous aperiodic sound source producing mixed sound sources such as the ones used to generate voiced fricatives.

2.2.1.3 Articulation and the Vocal Tract

Articulation refers to movements of the pharynx, tongue, lips and jaw to vary the properties of the acoustic filter formed by the vocal tract, that acts on the excitation signal generated within the larynx, to produce the spectral properties we recognize as the basic speech sounds of a given language. The

nasal tract can be acoustically coupled to the vocal tract by lowering the velum which allows the air to circulate through the nasal cavities.

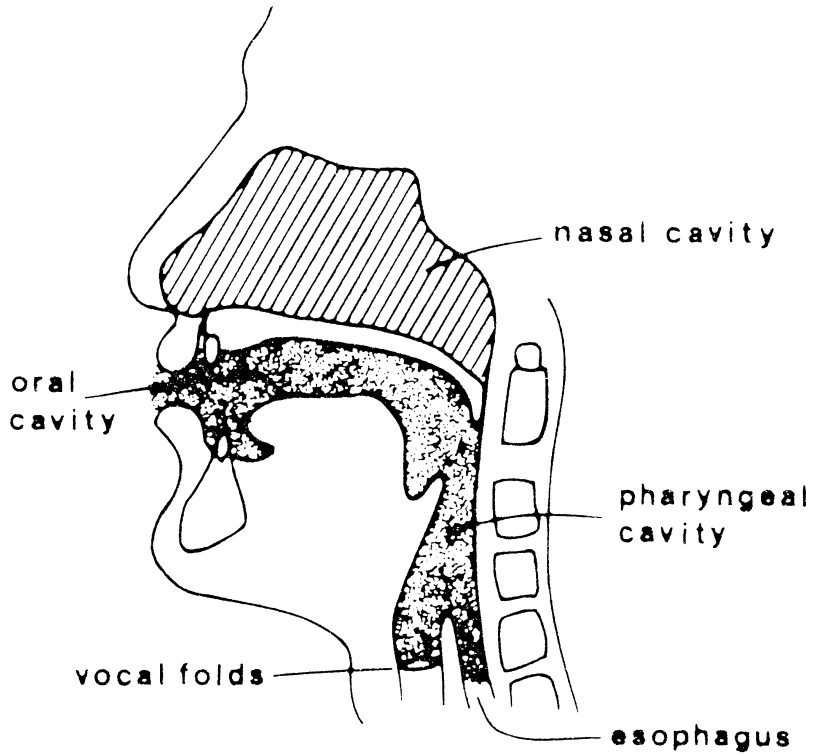

Figure 2.5: The cavities of the vocal and nasal tract. From Borden, Harris and Raphael [5].

The vocal tract includes all of the air passages above the larynx from the glottis to the lips, as shown in Figure 2.5. It is possible to alter shape and cross-section of the vocal tract by positioning the articulators into the desired configuration. The roof of the oral cavity consists of the hard palate, the soft palate or velum and the uvula (a small fleshy mass that hangs from the back of the velum), as shown in Figure 2.7. The upper surface of the tongue constitutes the floor of the oral cavity, which can be used to alter the dimension of the vocal tract. It has three parts: the tip or anterior edge; the blade (part that normally lies before the smooth part of the hard palate);

the body (part after the blade). The lips are also used to vary the length of the vocal tract. They can have four basic shapes: closed, spread, neutral or rounded.

Formant frequencies (resonances in the radiated spectrum) are primarily determined by the shape and length of the vocal tract [60]. Figure 2.6 shows how the glottal spectrum is shaped by the resonant characteristics of the vocal tract. The formant frequencies cannot be attributed solely to a particular resonant cavity. They are determined by the acoustic response of the vocal tract as a whole to the voiced source. Acoustically, the critical feature is the front cavity length and not the tongue position or lip protrusion. Speech sounds requiring small mouth openings tend to have low first formant frequencies (vocal tract resonances) [5]. For example, typical frequencies for the first three formants of the phoneme "i" would be $F_1 = 300Hz$, $F_2 = 2045Hz$ and $F_3 = 2960Hz$ [31]. When a speech sound is articulated with the tongue at the back of the oral cavity or the lips rounded the frequency of the second formant is usually lowered [5]. The third formant is influenced by the position of the tongue tip. For example, typical frequencies for the first three formants of the phoneme "A" would be $F_1 = 700Hz$, $F_2 = 1220Hz$ and $F_3 = 2600Hz$ [31].

Figure 2.7 shows the possible places of articulation for English speech sounds, including the lips (labial or bilabial), the gums (alveolar), the hard palate (palatal), the soft palate (velar) and the glottis (glottal). Manner of articulation describes the degree of constriction and the way in which the constriction is formed in the vocal tract [10]. A strong and forceful articulation is known as fortis and a weak articulation is known as lenis. Fortis is also used to account for aspiration in plosives. There are six manners of articulation: approximant (liquid or glide), nasal, plosive, affricate, fricative and vowel (monophthong or diphthong).

2.2.2 Vowels

Vowels are voiced sounds: the articulatory gestures that shape the vocal tract for vowel production are characterized by the point of major constriction, degree of constriction and lip rounding. The major constriction is considered to be the highest point of the body of the tongue [60].

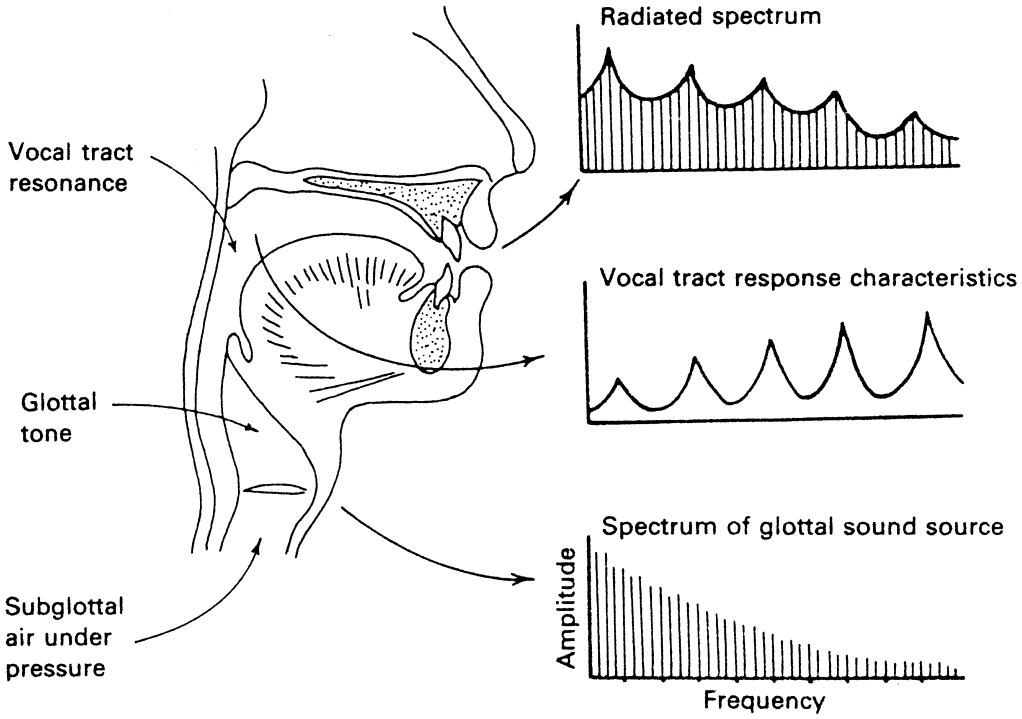

Figure 2.6: The production of a neutral speech sound. Spectrum of the glottal sound source, vocal tract transfer function and radiated spectrum. After Zemlin [60].

Figure 2.8 shows the vowel quadrilateral, indicating the articulatory positions of the vowels, labelled using the SAMPA phonetic alphabet (see Appendix B), relative to the "cardinal" vowels. Vowels are classified according to the articulatory position of the tongue relative to the palate, i.e. according to height, and as being toward the front of the oral cavity or toward the back. The cardinal vowels ("i", cardinal 1; ... ; "A", cardinal 5 ; ... ; "u", cardinal 16) are a set of idealized reference sounds whose perceptual quality is defined independently of any specific language. The vowels "i", "A" and "u" represent articulatory extremes. A discussion of articulation and the acoustics of these speech sounds is included in this section, following the detailed

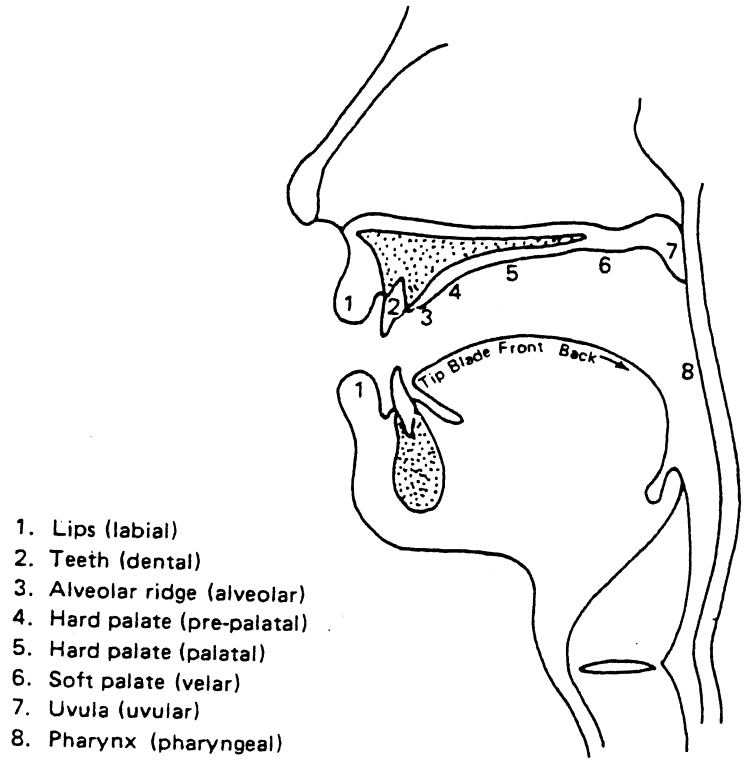

Figure 2.7: Articulators and places of articulation. From Zemlin [60].

description given by Borden, Harris and Raphael [5].

The front close unrounded vowel "i" (cardinal 1) is produced by elevating the tongue toward the alveolar ridge, as shown in Figure 2.9. The oral cavity is reduced and the pharyngeal space enlarged. The first formant is very low and the second and third formants are relatively high.

The back open unrounded vowel "A" (close to cardinal 5) presents a larger oral cavity and a smaller pharyngeal cavity than the vowel "i". The size of the oral cavity is increased by lowering the jaw and depressing the tongue, as shown in Figure 2.10. The first formant is relatively high due to the small pharyngeal cavity and the second formant is low owed to the large oral cavity.

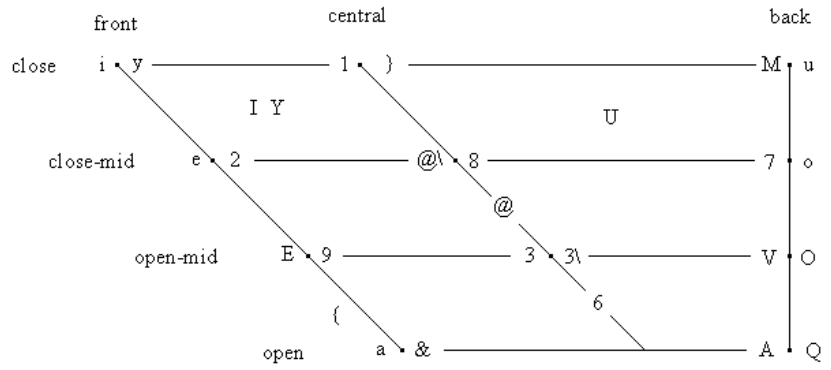

Figure 2.8: The vowel quadrilateral. Where symbols appear in pairs, the one on the right represents a rounded vowel.

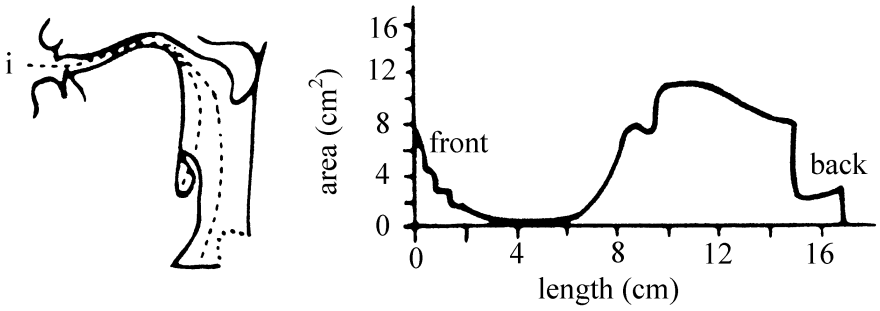

Figure 2.9: Vocal tract configuration and cross sectional area for the vowel "i". The abscissa indicates distance from the lips. After Fant [16].

The back close rounded vowel "u" (cardinal 16) is articulated with the dorsum of the tongue raised toward the roof of the mouth in the area of the juncture of the hard palate and the velum, as shown in Figure 2.11. The tongue is moved forward enlarging the pharyngeal cavity and producing a low first formant frequency. The rounded lip position increases the size of the vocal tract, thus lowering the second formant frequency.

Diphthongs are moving vowel articulations. Articulatory movement, particularly of the tongue, occupies a substantial portion of a diphthong. For diphthongs "I@", "E@" and "U@" the tongue moves toward the centre from

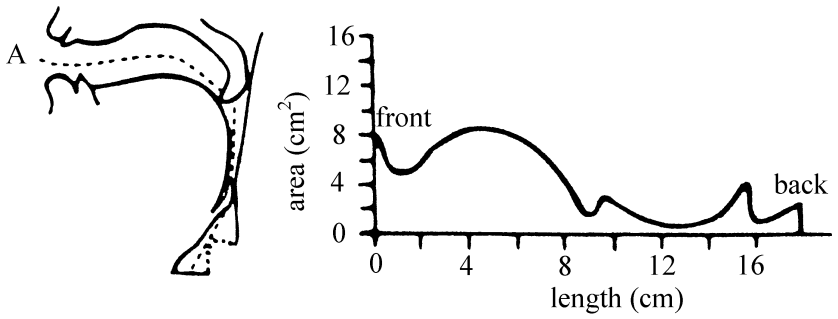

Figure 2.10: Vocal tract configuration and cross sectional area for the vowel "A". The abscissa indicates distance from the lips. After Fant [16].

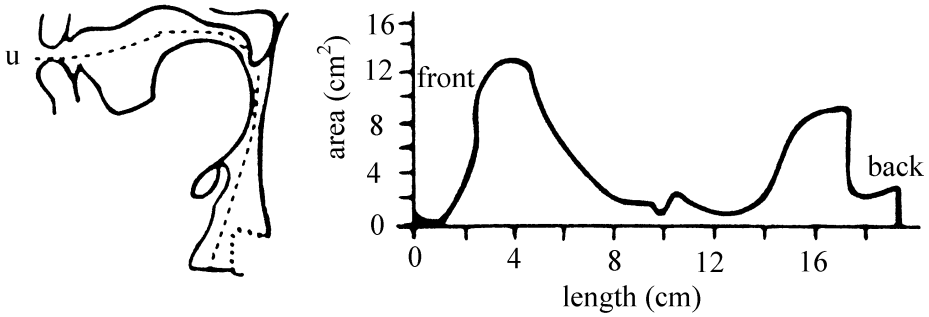

Figure 2.11: Vocal tract configuration and cross sectional area for the vowel "u". The abscissa indicates distance from the lips. After Fant [16].

the articulatory positions for "I", "E" and "U". Diphthongs "eI", "aI" and "OI" are produced by moving the tongue forward and up from the "e", "a" and "O" positions, to form the vocal tract cavities appropriate for "I". Diphthongs "@U" and "aU" are produced with back tongue movement from the "@" and "a" positions and lip closure.

2.2.3 Approximants

The approximants "l", "r", "w" and "j" are produced with a relatively open vocal tract and the spectrum exhibits well defined formants. They usually occur in the periphery of syllables but can also serve as nuclei of syllables. The phoneme "=l" as in "bott<u>l</u>e" is an example of an approximant that be-

cause of phonetic context is syllabic and consequently serves as vowel.

The approximants "l" and "r", known as liquids, are produced in syllable initial position by raising the tongue toward the alveolar ridge. Differences in tongue tip configuration and position ("l", front close; "r", central) create the distinctions between the two sounds. The liquid "l" is also known as a lateral because the air flows around the sides of the tongue. The tongue tip adjustments are acoustically reflected in the frequency of the third formant.

The approximants "w" and "j", known as glides, are produced by movements of the tongue and lips that change the vocal tract shape from the initial position to the next vowel articulation. The glide "w" has two places of articulation: labial and velar. The tongue blade approximates the palate to produce the glide "j".

2.2.4 Nasals

The velum is lowered opening the entrance to the nasal branches required to produce the nasal sounds "m", "n" and "N". The addition of nasal branches creates both nasal resonances and nasal antiresonances. They introduce new poles and zeros in the vocal tract transfer function. The longer resonator lowers the formant frequencies and the antiresonances attenuate the amplitudes of nasals.

• The voiced bilabial nasal "m" is produced with the lips closed, Figure 2.12. The air flow from the vocal folds is resonated in the pharyngeal cavity, the oral cavities and the nasal cavities.

• For the voiced alveolar nasal "n" the blade or tip of the tongue touches the alveolar ridge, with the back side of the tongue touching the upper molars.

• The voiced velar nasal "N" is produced with the tongue dorsum touching the soft palate, allowing less of the oral cavity to resonate as a side branch of the vocal tract.

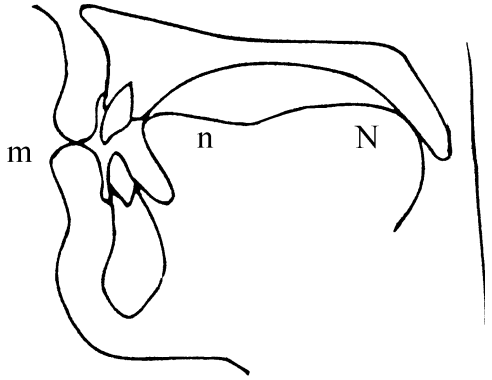

Figure 2.12: Place of articulation for the nasals: bilabial, alveolar and velar. After Borden, Harris and Raphael [5].

2.2.5 Fricatives

The fricative sounds are characterized by a restricted air flow caused by constrictions formed by the articulators. This generates aperiodicity whether or not phonation accompanies their articulation. Fricatives are the product of one (voiceless) or two (voiced) sound sources modified by vocal tract. This creates the possibility of using a single articulation to produce two distinctive sound sources.

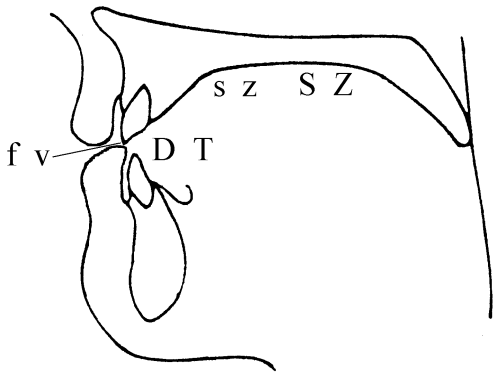

Figure 2.13: Place of articulation for the fricatives: labio-dental, dental, alveolar and palato-alveolar. The tongue is omitted for clarity. After Borden, Harris and Raphael [5].

• The labio-dental fricatives "f" (voiceless) and "v" (voiced) are produced by bringing the lower lip close to the interior edges of the upper central incisors, Figure 2.13.

• The dental fricatives "D" (voiceless) and "T" (voiced) are formed by approximating the tip of the tongue to the upper incisors.

• The alveolar fricatives "s" (voiceless) and "z" (voiced) are produced by forming a constriction between the alveolar ridge and the tongue.

• The palato-alveolar fricatives "S" (voiceless) and "Z" (voiced) are produced by forming a constriction in the pre-palatal area and by rounding the lips.

• The glottal fricative "h" is formed by constricting the air at the glottis. The vocal tract takes the shape of whatever vowel that follows. Although it is usually voiceless it can also become voiced when located between voiced segments.

2.2.6 Plosives

Transient aperiodic sound waves are generated at various locations of the vocal tract by blocking the flow of air completely for a brief period of time and then releasing the built up pressure abruptly. The articulators form the occlusion (closing phase), maintain the blockage of air flow (hold or closure phase) and then break the occlusion discharging the built up air pressure (release phase). The locations of the occlusions ("b" and "p" bilabial; "d" and "t" alveolar; "g" and "k" velar) are shown in Figure 2.14. For example, the phoneme "b" in the word "bear" is produced by closing the lips. The constriction used to generate the phoneme "g" in the word "gear" is located at the velum. There are fortis/lenis pairs for each place of articulation, for example the phoneme "p" in the word "pen" (fortis) and in the word "happy" (lenis).

There is a silent gap as the result of the period in which there is no flow of air out of the vocal tract and a noise burst at the moment of release, more intense for the voiceless plosives than for the voiced plosives. The first formant frequency rises rapidly following the initial plosives and falls rapidly

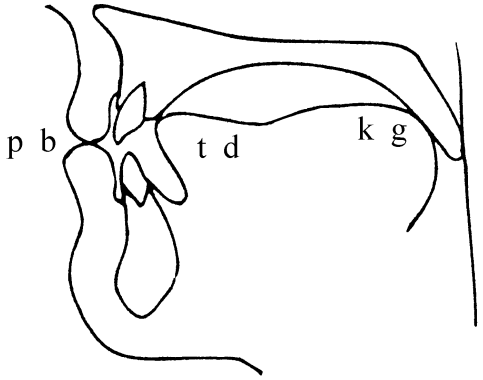

Figure 2.14: Place of articulation for the plosives: bilabial, alveolar and velar. After Borden, Harris and Raphael [5].

before the final plosives.

A phenomenon called voice onset time (VOT) is an important cue for the voiced-voiceless distinction of plosives. It is defined as the time interval between the articulatory burst release and the instant the vocal fold vibration begins [60]. If $VOT \geq 25ms$, the phoneme will be perceived as voiceless. If $VOT \leq 20ms$, it is perceived as voiced.

2.2.7 Affricates

The affricates "tS" and "dZ" are the combination of an alveolar closure and a continuous release. The articulators (tongue tip or blade and alveolar ridge) come together to form a closure and then, instead of coming fully apart, they separate only slightly, so that a fricative is made at the same place of articulation [35].

2.3 Co-articulation

Co-articulation can be defined as the variability in the speech sounds corresponding to a given phoneme, due to phonetic context. For a given language, there are a finite number of elementary speech sounds, known as phonemes, produced by articulatory gestures. The boundaries between phonemes in hu-

man speech are not distinct, instead there is a gradual transition from one speech sound to the next, largely due to the physical inertia of the articulators. For example, the lips do not come apart at the end of "p" in "apt" because the closure for "t" is anticipated [35]. Also the pronunciation of "k" before the front vowel "i" as in "key" is different from the pronunciation of "k" before the back vowel "O" as in "caw" [35]. This effect, known as co-articulation, must be reproduced in synthetic speech in order to produce a natural voice quality. We can produce intelligible speech without modelling the effects of co-articulation but it would sound too synthetic and unbearable to listen to for a long time.

2.4 Models of Speech Production

Fant [16] developed a popular acoustic model of speech production. The speech wave is the response of the vocal tract acoustic filter formed by the vocal tract, to voiced excitation generated in the larynx or unvoiced excitation generated within the vocal tract itself. The speech wave may be uniquely specified in terms of source and filter characteristics, as shown in Figure 2.15:

$$P(f) = U(f)R(f),$$

where

$$U(f) = S(f)T(f).$$

U(f) is the volume velocity at the lips. The amount of air flowing through the glottis is known as volume velocity. It is determined by the rate at which the vocal folds open and close, subglottal air pressure and area of glottal opening. R(f) is the radiation characteristic resulting from the conversion of the volume velocity passing through the lips to pressure in the sound field P(f). U(f) does not take into account the flow to pressure conversion (radiation) function at the lip boundary, which is included in R(f) [12]. S(f) is the glottal sound source volume velocity and T(f) is the vocal tract transfer function.

Most speech synthesis systems use source filter models as the one described above. The following sections will describe implementations of articulatory, formant and LPC models of speech production.

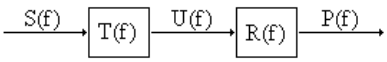

Figure 2.15: Acoustic model of speech production.

2.4.1 Articulatory Models

Articulatory models are intended to directly model the physiological structure of the articulators. In such a modelling system, locations and shapes of articulatory organs are used as control parameters for speech synthesis. Articulatory parameters directly encode the glottal area and the air pressure in the lungs, larynx, vocal tract and nasal tract, and the resulting mechanical articulator movements. The model assumes that a nonlinear function transforms the articulatory parameters representing the location of the articulatory organs into the acoustic features of speech. The vocal tract excitation by periodic vibration of the vocal folds can be reproduced by synthetic glottal pulses or by a two-mass model of the vocal folds, as shown in Figure 2.16.

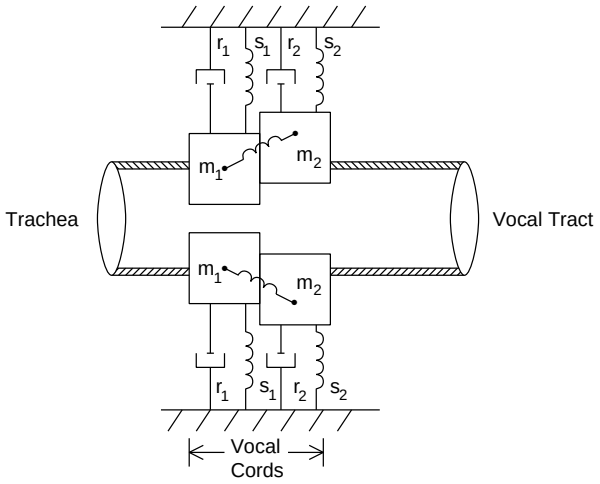

Figure 2.16: Two-mass, two-dimensional model of the vocal folds. After Ishizaka and Flanagan [28].

Mermelstein [43] represented the vocal tract outline by means of variables

specifying the positions of the jaw, tongue, lips and velum, as shown in Figure 2.17. The vocal tract area function is computed at regular intervals from mid-sagittal x-ray tracings using the grid shown in Figure 2.18. The vocal tract outline is sampled at intervals of 0.5cm, where it is approximately straight, and at intervals of 10° around the back of the tongue. Synthetic speech is generated by concatenating the individual responses of each vocal tract section to a periodic excitation signal.

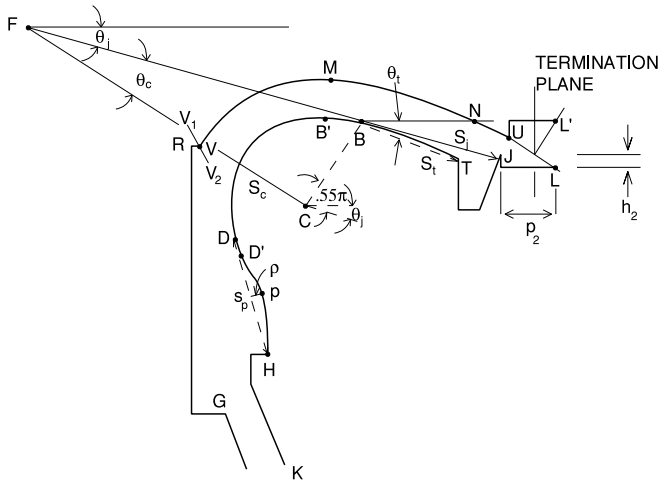

Figure 2.17: Model generated vocal tract outline. After Mermelstein [43].

Articulatory synthesizers are useful speech research tools. The articulatory model uses simple control parameters, directly encoding the effects of coarticulation. The complexity of articulatory models and the limited supply of articulatory data from natural speech present a major obstacle to the commercial application of articulatory synthesis [38].

2.4.2 Formant Models

Formant models attempt to replicate the spectral properties, i.e. the formant structure, of the acoustic speech signal, without modelling the underlying physical processes involved. The formant models are simpler and require less calculation than articulatory models. A formant synthesizer normally uses a shaped glottal pulse train as the sound source for voiced speech and a noise

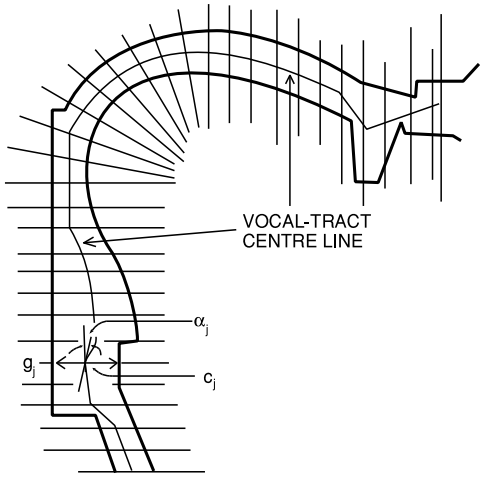

Figure 2.18: Grid system for conversion of mid-sagittal dimension to vocal tract cross sectional area. After Mermelstein [43].

source to produce a suitable spectrum for unvoiced speech sounds. The vocal tract is modelled by a digital filter with several resonances on its frequency response. The most common configurations are:

- parallel configuration - the resonators that model the vocal tract transfer function are connected in parallel. Each resonator is preceded by an amplitude control that determines the relative amplitude of a spectral peak (formant) for both voiced and voiceless sounds;
- cascade configuration - the voiced sounds are synthesized using a set of cascade resonators. The output of a resonator is fed into the input of the next. Amplitudes are automatically correct as consequence of cascade calculation.

Holmes [24] [27] proposed a parallel synthesizer configuration that modelled all types of speech sounds with individual amplitude control, which provided a better model for vocal tract variations.

Klatt [31] [32] proposed a cascade/parallel approach that used a cascade configuration to obtain the relative amplitudes of the desired formant peaks without individual control of each formant amplitude. Klatt concluded that

better results could be obtained when producing fricatives and plosives by using the parallel configuration. Since both cascade and parallel branches were used to produce speech, the synthesizers overall structure is quite complex.

The synthesizer described by Klatt, shown in Figure 2.20, is usually used in its cascade/parallel configuration but it can also work in a parallel configuration when individual control of formant amplitudes in vowels is required.

A formant resonator, shown in Figure 2.19, can easily be implemented on a digital computer using a second order discrete time band pass filter. The sampled output y(n) is obtained from the input x(n) through the equation:

$$y(n) = Ax(n) + By(n-1) + Cy(n-2).$$

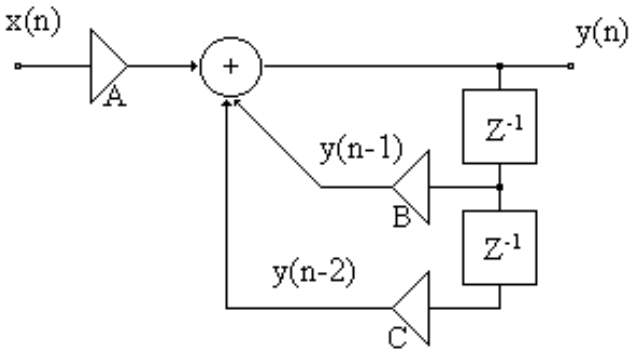

Figure 2.19: Resonator block diagram. After Klatt [31].

The constants A, B and C are related with the parameters used to specify the input and output resonator characteristics (resonant frequency F and resonant bandwidth BW) through the equations:

$$C = -e^{-2\pi BWT},$$

$$B = 2e^{-\pi BWT} cos(2\pi FT),$$

$$A = 1 - C - B,$$

where T=1/sampling frequency.

When the resonant frequency F=0 we obtain a low-pass filter with a -12dB/octave slope and a 3dB frequency of BW/2. This resonator can be used to model the natural glottal impulse reproduced by the synthesizers voicing source.

The antiresonators introduce two transfer function zeros (antiresonances or antiformants) used to model the voicing source spectrum and to reproduce the effects of nasalization in the cascade configuration.

The antiresonator output y(n) is related to the input x(n) through the equation:

$$y(n) = A'x(n) + B'y(n-1) + C'y(n-2).$$

The constants A', B' and C' are defined through the equations:

$$C' = -CA,$$

$$B' = -B/A,$$

$$A' = 1/A,$$

where A, B and C are obtained substituting the antiresonance central frequency F and the antiresonance bandwidth BW in the resonator equations.

The synthesizer block diagram is illustrated in Figure 2.20. The resonator is represented by the prefix r and the amplitude control is represented by the prefix a. Each resonator rn has a resonant frequency control parameter fn and a resonant bandwidth control parameter bn.

2.4.3 Linear Predictive Models

Speech analysis and synthesis methods are concerned with redundancies in the natural speech signal. The coding efficiency and the quality of synthesized speech depend on the accuracy of extracted feature parameters. Methods based on linear prediction of the speech signal [39] have been widely used. The following sections present a discussion of linear predictive coding (LPC) [51], partial autocorrelation (PARCOR) and line spectral pair (LSP) [56] methods.

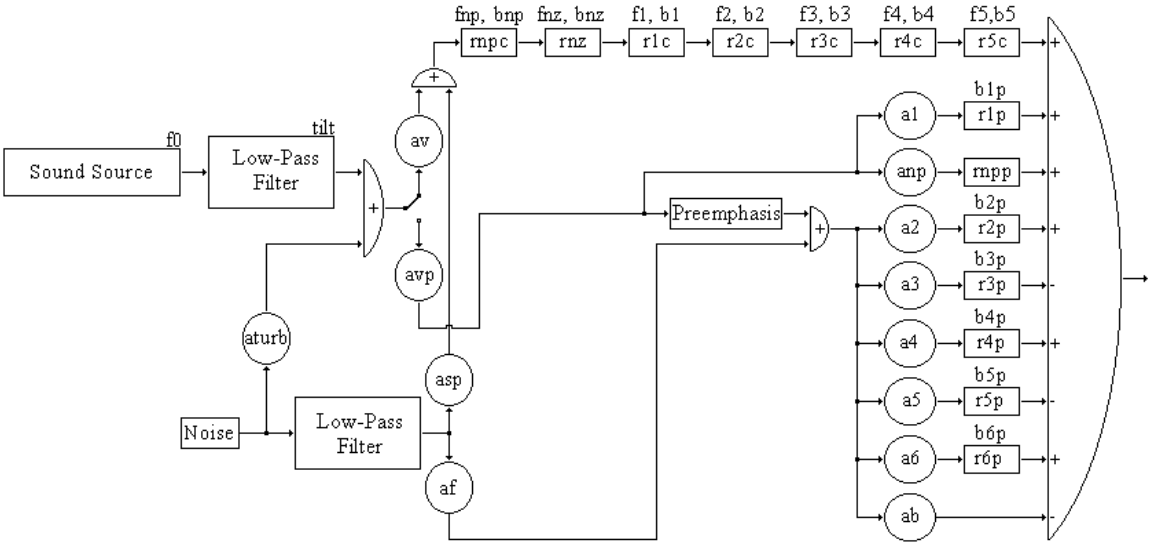

Figure 2.20: The Klatt formant synthesizer. rnpc is a nasal resonator (cascade branch). rnz is a nasal antiresonator (cascade branch). r1c, r2c, r3c, r4c and r5c are the cascade branch resonators. rnpp is a nasal resonator (parallel branch). r1p, r2p, r3p, r4p, r5p and r6p are the parallel branch resonators. After Klatt [31].

2.4.3.1 Linear Predictive Coding (LPC)

Although speech is a continuously time varying process it is possible to develop linear models that are locally time invariant for describing important speech events. The linear prediction (LP) of speech signals is a linear equivalent model for speech production. A given speech sample can be approximated as a weighted sum (or linear combination) of a small number of past samples. Let *x(n)* be a discrete signal (sampling period ΔT seconds). The linear predictive model of a speech signal is given by the equation

$$x(n) - \sum_{i=1}^{p} \alpha_i x(n-i) = \sigma e(n),$$

where *e(n)* is the excitation source signal and σ its RMS value. $\sigma e(n)$ is regarded as a prediction residual, i.e. the difference between x(n) and the

linearly predicted signal. The LPC synthesis filter, shown in Figure 2.21, is an all pole time varying filter of the form

$$H(z^{-1}) = \frac{\sigma}{A_p(z^{-1})},$$

where

$$A_p(z^{-1}) = 1 + \sum_{i=1}^{p} \alpha_i z^{-1}.$$

α_i are the LPC parameters and p is the predictor order. The filter parameters (short term predictor coefficients) are determined by an *autocorrelation method* or a *covariance method*. The basic approach is to find a set of predictor coefficients α_i that minimize the mean squared prediction error over a short segment of sampled speech, resulting in the following set of linear equations

$$\sum_m x_n(m-k)x_n(m) = \sum_{i=0}^{p} \alpha_i \sum_m x_n(m-k)x_n(m-i),$$

where $x_n(m) = x(n+m)$ is the short term speech segment at time n. If the short term covariance of $x_n(m)$ is defined as

$$\phi_n(k,i) = \sum_m x_n(m-k)x_n(m-i),$$

where $1 \leq k \leq p$ and $0 \leq i \leq p$, then the set of linear equations can be expressed as

$$\sum_{i=0}^{p} \alpha_i \phi_n(k,i) = \phi_n(k,0).$$

The covariance method assumes that $x_n(m)$ is zero outside of the interval $0 \leq m \leq N-1$ (window of length N) and so the m summation limits are

$$\sum_{m=0}^{N-1} x_n(m-k)x_n(m-i).$$

The autocorrelation method uses a weighting window function for defining $x_n(m)$ and so the m summation limits are

$$\sum_{m=0}^{N-1+p} x_n(m-k)x_n(m-i).$$

The filter parameters α_i can be determined by solving the set of linear equations, once the covariance values $\phi_n(k, i)$ are calculated for the interval defined by the covariance or autocorrelation method.

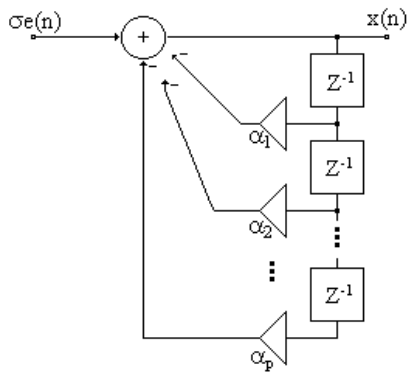

Figure 2.21: LPC synthesis filter. After Sugamura and Itakura [56].

LPC coefficients are known to be inappropriate for quantization because of their relatively large dynamic range and possible filter instability problems. An alternative representation know as partial autocorrelation was thus proposed.

2.4.3.2 Partial Autocorrelation (PARCOR)

Different sets of parameters representing the same spectral information, such as the PARCOR coefficients, were proposed for quantization. The coefficients for the PARCOR lattice filter, shown in Figure 2.22, provide a parametric representation of speech that can be quantized efficiently. The PARCOR feature parameters can be divided into two groups: the excitation source parameters (fundamental frequency, power and voiced or unvoiced information) and the spectral parameters that represent the vocal tract frequency transmission characteristics according to the articulatory movements.

The speech synthesis system includes a voicing source and a noise source that excite a time varying filter composed of lattice sections. The synthesis filters are always stable as long as the absolute value of each PARCOR coefficient

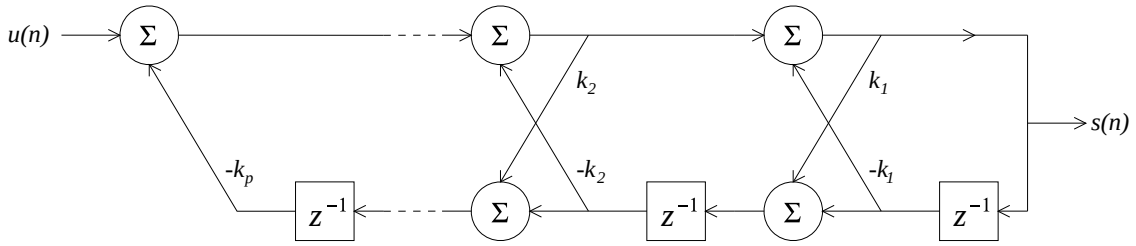

Figure 2.22: PARCOR lattice synthesis filter.

k_i is less than one.

When the analysis frame period is long the quality of the synthesized speech degrades. Non-uniform bit allocation and non-linear transform quantization of PARCOR coefficients have been used to improve the quality of speech synthesized with a lower bit rate [56]. The PARCOR method is superior to any other previously developed methods due to the simplicity of maintaining filter stability, but if the bit rate falls bellow 2400 bps the synthesized speech quality rapidly deteriorates.

2.4.3.3 The LSP Transformation

In order to alleviate problems in the PARCOR scheme, the line spectral pair (LSP) transformation of the LPC polynomial was introduced by Itakura [29]. The LSP coefficients are obtained from two orthogonal polynomials which are related to the LPC prediction polynomial. The LSP parameters are widely used in speech coding due to the desirable properties such as bounded range, strong intra-frame and inter-frame correlation and simple checking of filter stability. Furthermore, LSP representation allows frame to frame interpolation with smooth spectral changes and great robustness against quantization effects.

During the LSP quantization procedure the LPC parameters are converted into a set of quantized line spectral frequencies. Line spectrum pair (LSP) parameters (ω_1, θ_1),... , ($\omega_{p/2}$, $\theta_{p/2}$) are an alternative representation to the linear prediction (LP) parameters (α_1,... , α_p). w and θ are angles in the unit circle whose dimension is in radians ($z = e^{jw}$).

The starting point for deriving the line spectrum pair (LSP) parameters is the all pole prediction filter of order p $A_p(z^{-1})$, which is decomposed into two p+1 order polynomials, one having an even symmetry $P(z^{-1})$ and the other having an odd symmetry $Q(z^{-1})$. This can be accomplished by taking a difference and sum between $A_p(z^{-1})$ and its conjugate function as follows

$$P(z^{-1}) = A_p(z^{-1}) - z^{-(p+1)}A_p(z) =$$

$$1 + (\alpha_1 - \alpha_p)z^{-1} + ... + (\alpha_p - \alpha_1)z^{-p} - z^{-(p+1)},$$

and

$$Q(z^{-1}) = A_p(z^{-1}) + z^{-(p+1)}A_p(z) =$$

$$1 + (\alpha_1 + \alpha_p)z^{-1} + ... + (\alpha_p + \alpha_1)z^{-p} + z^{-(p+1)}.$$

The LPC analysis filter, reconstructed by the use of these two filters, is

$$A_p(z^{-1}) = \frac{P(z^{-1}) + Q(z^{-1})}{2}.$$

The roots of these polynomials lie on the unit circle in the z plane. $P(z^{-1})$ has a real zero at z=-1, $Q(z^{-1})$ has a real zero at z=1, all other zeros occur in complex conjugate pairs and alternate around the unit circle as shown in Figure 2.23. Each zero pair corresponds to a pole pair in the forward model. The arguments of these zeros are referred to as LSP parameters (only the angle is recorded as the modulus is 1). $P(z^{-1})$ and $Q(z^{-1})$ correspond to lossless models of the vocal tract with the glottis completely closed and completely open, respectively.

The LSP synthesis filter is an all pole filter with identical properties to the LPC filter shown in Figure 2.21 with

$$A(z^{-1}) = \sum_{i=1}^{p} z^{-1}.$$

It can be shown that a necessary and sufficient condition for the stability of the synthesis filter is the above interleaved ordering property of the LSP frequencies. Figure 2.24 shows an example of an all-pole digital filter with

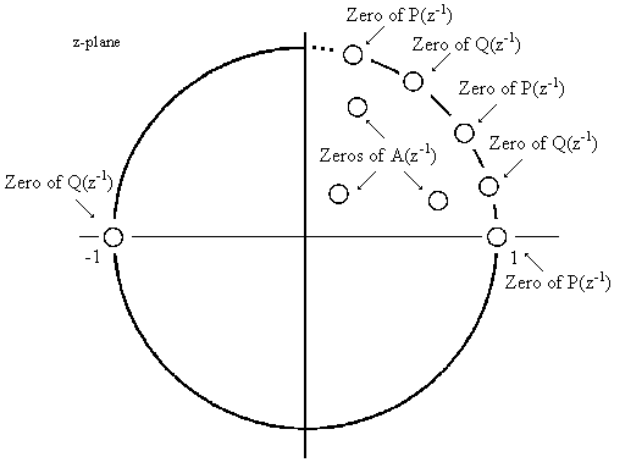

Figure 2.23: The zeros of $A(z^{-1})$, $P(z^{-1})$ and $Q(z^{-1})$. After Deller, Proakis and Hansen [12].

the above transfer function in the negative feedback loop.

There is some correspondence between the LSP frequencies and the formant structure of speech sounds. The LSP frequencies are related to the frequencies of LPC poles and these are in turn related to the formant frequencies. A zero of $A(z^{-1})$ close to the unit circle corresponds to a formant in the model, corresponding to a zero of $P(z^{-1})$ and a zero of $Q(z^{-1})$ close in frequency. In Figure 2.25 the LSP frequencies corresponding to a resonant pole are close together around the formant centre frequency, i.e. parameters concentrate when there is a formant. During a period of silence the LSP parameters are placed at almost equal intervals along the frequency axis.

The quality of synthesized speech is related to the information rate. A lower information rate causes deterioration of synthetic speech. Therefore it is crucial to define the relation between information rate and speech quality. The LSP parameters have been shown to be superior to PARCOR coefficients (Sugamura and Itakura [56]) as to their quantization and interpolation properties as a function of spectral distortion. LSP uniform bit allocation is nearly optimum and LSP parameter interpolation characteristics are superior to that of PARCOR coefficients. The LSP frequency range is limited

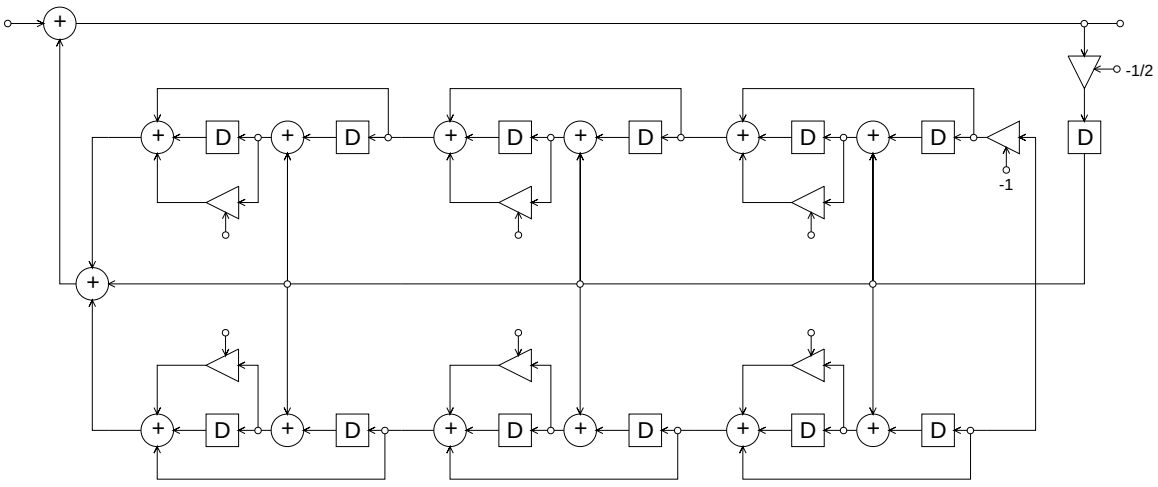

Figure 2.24: Sixth order direct form LSP synthesis filter (p is odd). After Sugamura and Itakura [56].

to a fraction of the total range allowing to improve coding efficiency. There is a strong frame to frame correlation of the LSP frequencies, as well as the correlation within a frame. Differential coding makes use of this property to make yet another reduction in bit rate.

2.5 Speech Synthesis

Speech synthesis is concerned with the generation of speech by computers. Many techniques have been developed ranging from relatively simple systems for creating speech utterances by concatenating speech waveforms, to complex synthesis by rule systems. A brief discussion of these two approaches followed by a description of concatenative speech synthesis and speech synthesis by rule systems forms the remainder of this section.

2.5.1 Concatenative Speech Synthesis

Speech synthesis systems based on the principle of concatenating speech sounds have been widely used since the 1930s. An early example is the British Post Office talking clock which used analogue recording and play-

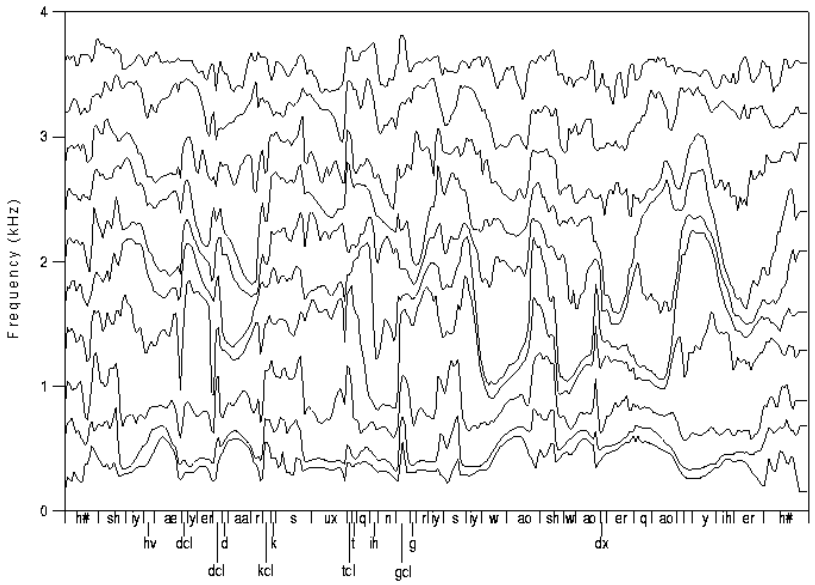

Figure 2.25: Line spectrum pair parameters for the sentence "she had your dark suit in greasy wash water all year". After Cawley [6].

back systems. When only a limited vocabulary is required, whole words have been used to produce high quality speech. Words can be easily concatenated because co-articulation effects are captured within the individual speech units. The size of the basic speech unit is a compromise between storage requirements and speech quality. If the size of the unit is increased additional co-articulation effects are included in the speech unit, i.e. more articulatory gestures are captured in the speech unit, resulting in high levels of naturalness in synthetic speech.

- The smallest synthesis unit is the phoneme. Only about 70 allophones are required for unrestricted vocabulary speech synthesis, which include the grossest effects of co-articulation.

- The diphone is a natural unit for speech synthesis. It is defined as the transition between two consecutive phonemes which corresponds to the second half of one phoneme followed by the first half of the next. The co-articulatory influence between consecutive phonemes does not extend much further than halfway into the next phoneme. This results

in small concatenation discontinuities and so it's possible to produce good quality speech with around 1200 units.

- The syllable consists of a vowel nucleus and two adjacent consonants clusters (consonant nucleus and two adjacent vowel clusters syllables are also used). Around 10000 syllables are necessary to reproduce most natural speech transitions although co-articulation across syllable boundaries is not included.

- The demi-syllable is defined as half of a syllable, either the initial consonant plus half of the vowel, or the second half of the vowel plus the postvocalic consonant. Around 800 initial demi-syllables and 1200 final demi-syllables are required to produce high quality speech.

A technique for blending the speech units is usually included in concatenative synthesis systems in order to smooth the transition between adjacent speech sounds. Text-to-speech systems based on concatenation of speech units also incorporate prosodic units.

2.5.1.1 Concatenation of Dyads

One of the earliest attempts to synthesize speech by this method is that of Peterson, Wang and Sivertsen [49] [59]. Firstly they defined a set of all segments involving a single articulatory sequence pair and all conditions of prosody associated with that sequence. They used these discrete segments of recorded sentences, known as dyads, to produce continuous speech by manually joining them together. The beginning and ending of a segment was considered to be at the most stable position in each phoneme. A segment inventory for speech synthesis of 43 phonetic units divided into 6 groups (syllabics, glides, nasals, fricatives, plosives and silence) was produced. These units represented all the allophones in the synthesized sentence used to test this method. The collection of an inventory involving 8500 segments was suggested as a suitable approach to reproduce the dynamics of American English speech.

2.5.1.2 Diphone Method of Segment Assembly

Dixon and Maxey [13] presented a system for generating speech signals using the diphone method of segment assembly. The system consisted of a cas-

cade/parallel formant synthesizer controlled by parameters extracted from a set of stored speech segments. The speech segmentation method was based on a transcription scheme consisting of 25 consonants, 10 vowels, 5 diphthongs and the schwa. A detailed description of several principles to govern the decision as to whether or not a particular diphone is needed in the segment library was devised, the perceptual relevance being the primary guideline.

The diphones in the example phrase "I'm sorry, you've reached this office by mistake. Please consult your directory and dial again." were divided into the following 16 classes: plosive-vowel, vowel-plosive, fricative-vowel, vowel-fricative, glide-vowel, vowel-glide, nasal-vowel, nasal-plosive, nasal-fricative, vowel-affricate, affricate-plosive, fricative-plosive, vowel-vowel, plosive-glide, fricative-glide and plosive-plosive. Dixon and Maxey discussed each one of these categories, presenting considerations on initial and final diphones, syllabic consonants, vowel reduction and prosodic features. Parameters used for synthesis and the evaluation methods are also described.

The system produced intelligible speech. The authors provide a detailed analysis of a suitable segment inventory for diphone synthesis and a deep study of phoneme transitions. They also describe modifications in the normal relationship existing between diphone names and phonetic events.

2.5.1.3 Speech Resynthesis from Phoneme LPC-Derived Area Functions

Olive and Spickenagel [47] resynthesized continuous speech represented by steady state regions of the various phonemes and transition regions between them defined by straight lines of LPC-derived area functions, as shown in Figure 2.26. The exact location of phonetic boundaries was manually adjusted in an attempt to improve the quality of synthesized speech. The authors proposed the use of this method to automatically derive the parameters in text-to-speech systems.

2.5.1.4 Rule Synthesis from Dyadic Units

Olive [46] suggested the use of a new unit that specifies only the transitions between the phonemes in a dyad. The steady states can then be obtained by

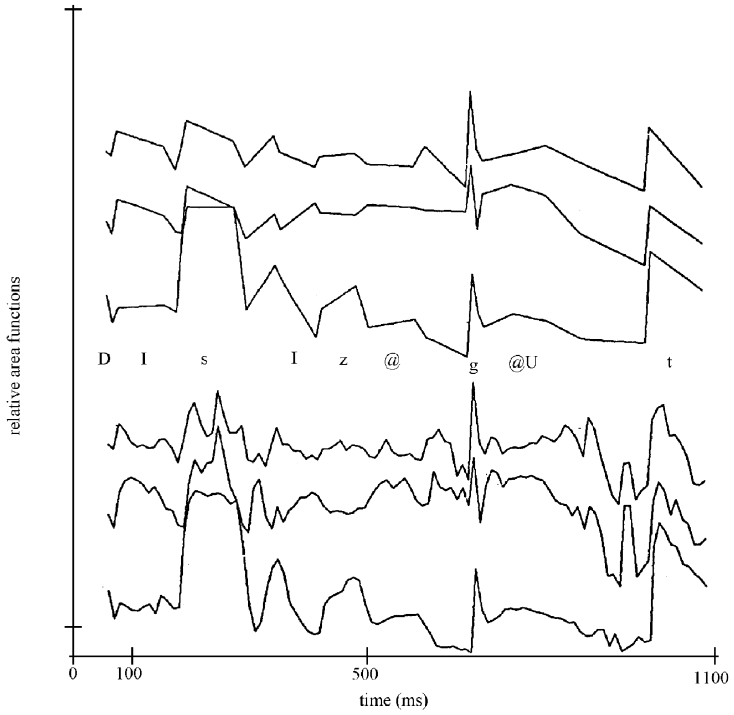

Figure 2.26: A comparison of smoothed and unaltered area functions of the first, sixth and twelfth area functions for the sentence "This is a goat". After Olive and Spickenagel [47].

connecting, with straight lines, the end points of adjacent transitions. Only the end points of transitions need to be specified and stored. The transitions are obtained by interpolation of end points. Olive described a synthesis by rule scheme based on these segmental concepts.

The system included a dictionary containing segmental data specified in terms of 12 LPC derived log area parameters. The dictionary described end points of 11 vowels and 22 consonants. Affricates and diphthongs are defined as the combination of two phonemes.

Olive also uses a pronouncing dictionary to do the orthographic to phonemic

transformation and to provide prosodic data such as durational values for the different phonetic segments. Vowel and consonant concatenation rules are specified differently. The system also incorporated a set of prosodic rules imposing a specific amplitude function for the utterance and a fundamental frequency contour. This complete rule synthesis scheme produced high quality speech.

2.5.1.5 The PSOLA Synthesis

Stella, Charpentier and Moulines [54] [9] [44] [45] described a text-to-speech system based on diphone concatenation using a multipulse representation of the LPC excitation signal. Firstly the diphone time waveforms are sampled at 16 KHz, low-pass filtered and coded using a 16 bit ADC. The voiced portions are hand labelled period by period and the voiceless portions are segmented into a number of nominal length windows. A pitch synchronous overlap and add (PSOLA) analysis is performed, in which the short-term signal is made to correspond to one pitch period of the voiced speech signal, both at analysis and synthesis stages. The unvoiced short-term signals are not converted to the frequency domain since they will only need time scaling. The frequency domain modifications algorithm performs linear interpolation of the source component and useful envelope component modifications, as shown in Figure 2.27. Time scale prosodic modifications of natural speech are also included (time warping function represented in Figure 2.28).

High quality prosodic modifications of speech are obtained using a time domain formulation of the PSOLA technique. The short term signals are given by

$$s_n(m) = h_n(t_n - m)s(m),$$

where $s(m)$ is the original signal, $s_n(m)$ is the short-term signal, $h_n(t_n - m)$ is the analysis window and t_n is the pitch period.

The synthetic signal $\hat{s}(m)$ is obtained by means of the least squares overlap and add synthesis procedure shown in Figure 2.29 and defined as

$$\hat{s}(m) = \frac{\sum_{n=-\infty}^{+\infty} \hat{s}_n(m)\hat{h}_n(m - \hat{t}_n)}{\sum_{n=-\infty}^{+\infty} \hat{h}_n^2(m - \hat{t}_n)}.$$

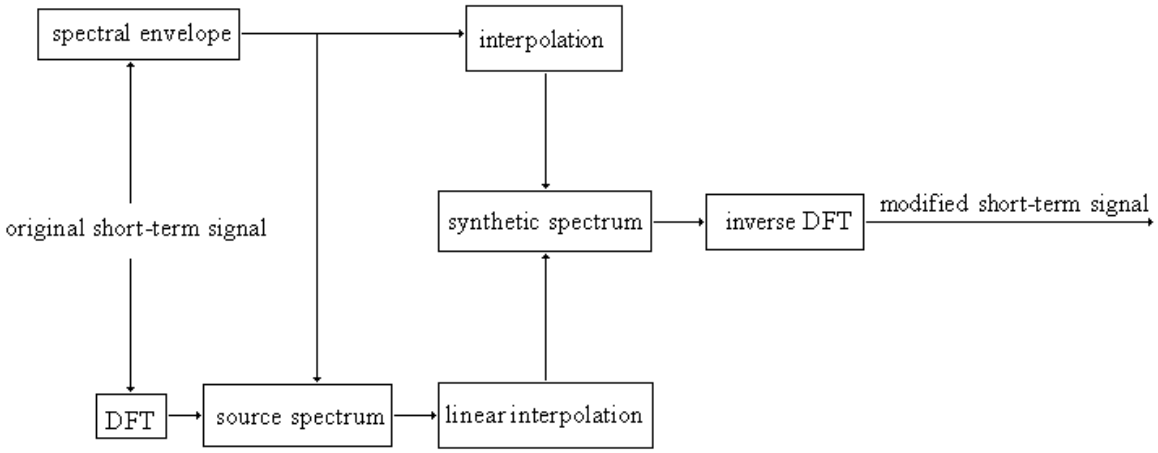

Figure 2.27: Block diagram of the PSOLA frequency domain modifications procedure. After Charpentier and Stella [9].

The modified synthesis short-term signal $\hat{s}(m)$ is produced by altering both the rate of articulation (time scale modification) and the fundamental frequency (pitch scale modification). $\hat{t}_n$ is the new pitch period.

The final step before producing synthetic speech is to concatenate diphones and to reduce distortions due to discontinuities in boundaries. At the boundary between two successive diphones the smoothing of the transition is obtained by time averaging the two boundary short-term signals and an energy correction on the second diphone to eliminate a possible energy mismatch. Short-time phase smoothing combines the use of pitch alignment of the excitation signal and time domain splicing eliminating the unnatural discontinuities at the boundaries.

The system also uses a prosodic module providing F0 and duration values for the sentence at synthesis time. The synthesis of speech, with a variable number of samples equal to the current synthesis pitch period, is possible by use of the least squares overlap and add scheme.

Two dictionaries (male and female speakers), consisting of 1200 diphone waveforms labelled with pitch marks, were used to test the text-to-speech

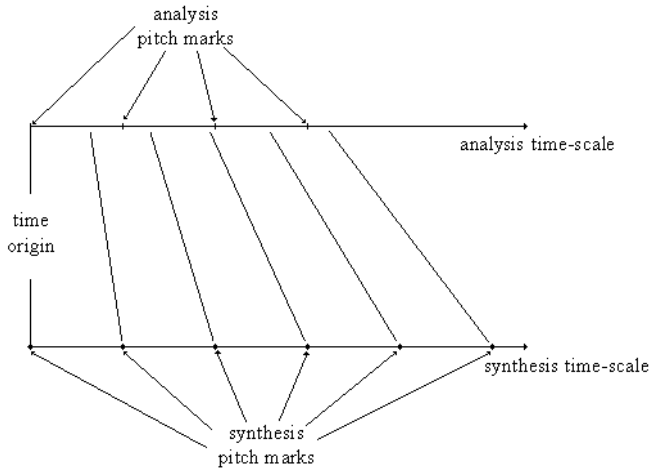

Figure 2.28: PSOLA time scale modification scheme. After Charpentier and Stella [9].

system shown in Figure 2.30. The high quality prosodic manipulations implemented by this system result in speech with better quality than conventional LPC synthesis.

2.5.2 Speech Synthesis by Rule

Rather than assembling short segments of human speech, truly synthetic speech can be generated by rule, as in unlimited vocabulary text-to-speech systems consisting of a phonological component which pre-processes plain text (i.e. converts the input text into a phonetic transcription) and a phonetic component which generates the speech sounds corresponding to each speech unit. Systems such as the Joint Speech Research Unit synthesiser [37] and MITalk [1], attempt to model the effects of co-articulation by interpolating formant parameters with a piecewise linear template, using interpolation parameters tabulated for each parameter for each allophone. The Holmes-Mattingly-Shearme [26] algorithm provides an early example of this approach. A set of ad-hoc rules are used to modify the interpolation parameters, where necessary, according to phonetic context is often required to achieve acceptable speech quality. Compiling and fine-tuning the tables of interpolation parameters for each allophone involves a great deal of painstak-

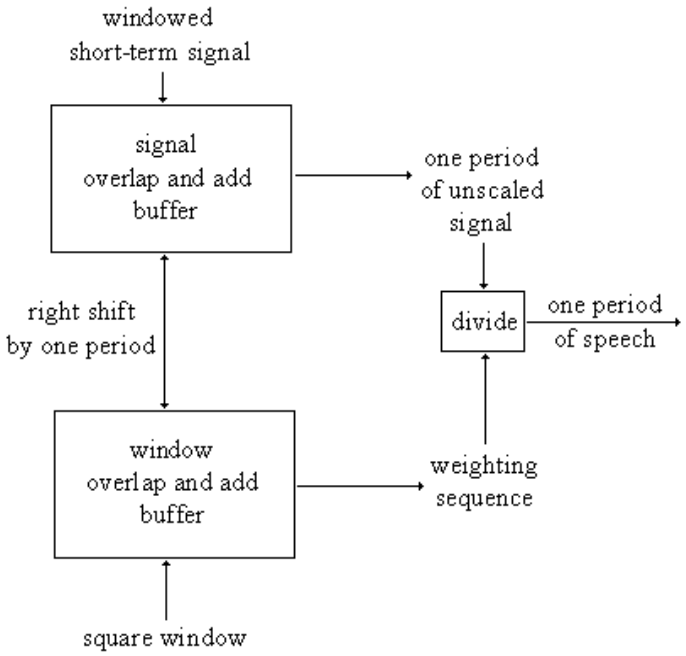

Figure 2.29: Least squares overlap and add synthesis procedure. After Charpentier and Stella [9].

ing manual comparison of the spectra of natural and synthetic utterances. Revoicing in a text-to-speech synthesis system is therefore a costly operation.

2.5.2.1 The Holmes-Mattingly-Shearme Algorithm

The system described by Holmes, Mattingly and Shearme [26] consists of an electronic analog synthesizer and a computer program that calculates the synthesis parameters. Information entered in an input table supplies specifications for transitions between the speech sounds in the sentence to be synthesized.

The computer program processes the information contained in the input file and produces new values (every 10ms) for the nine parameters that control the raw synthesizer [23] [53] [24] [27]. The input sentence consists of phonetic symbols, prosodic modifiers and F_0 values. The collection of phoneme

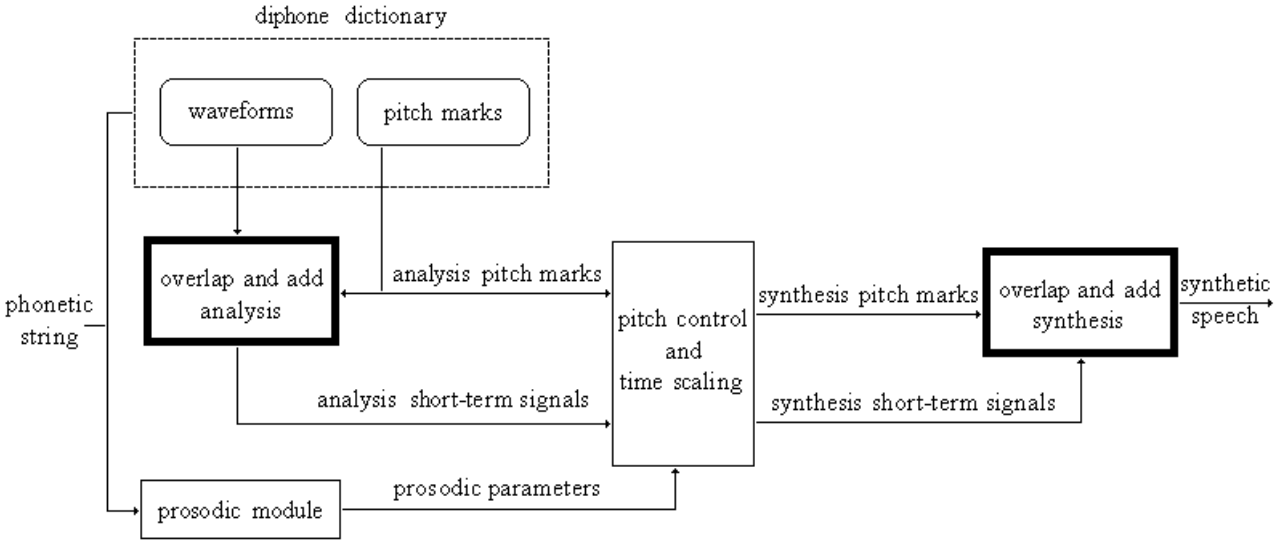

Figure 2.30: Block diagram of the PSOLA diphone synthesis procedure. After Charpentier and Stella [9].

input tables include the following additional information: duration, steady-state parameter values and transition conditions. These tables were compiled manually from direct observation of natural speech spectrograms.

The formant parameters linear interpolation algorithm uses the template shown in Figure 2.31 to model smooth transitions between the tabulated target values. The system uses a simple interpolation strategy modified by a set of ad hoc rules to model complex parameter transitions.

The transition values of F_0 are calculated by linear interpolation. The values for F_1, F_2, F_3, A_1, A_2, A_3 and $A_{frication}$ are determined for an initial transition, a steady state period and a final transition. The interpolation procedure uses the following set of parameters to control the transition: the steady state value, the fixed contribution value, the percentage value (steady state value - fixed contribution to the boundary values), the external transition duration and the internal transition duration.

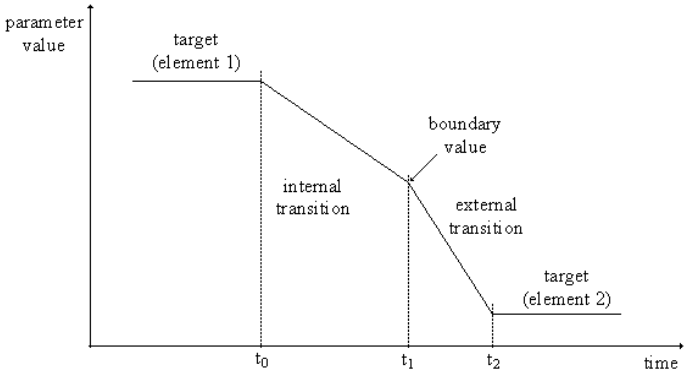

Figure 2.31: Template used for interpolation of parameter transitions (Holmes-Mattingly-Shearme scheme). After Holmes [25].

In addition a rank is assigned to each phonetic element. The rank is considered to be low if the transition characteristics are influenced by adjacent phonemes, i.e., higher ranked segments dominate lower ranked ones. The table of the higher rank element is used to determine the nature of the transition and the table of the lower rank element is used only to provide parameter targets. Where two consecutive elements have the same ranks, the first one is considered dominant [25].

The consonant table entries are used to determine the transition type. Both vowel-consonant and consonant-vowel transitions are defined by the consonant characteristics. The transition values are calculated by linear interpolation between steady state values and boundary values given by

boundary value = fixed contribution value of dominant sound + (steady state value of dominated sound $\times$ percentage value/100).

The algorithm reproduces a variety of transition paths between adjacent phonemes. For example, the transitions between two vowels are modelled by straight line templates as the one shown in Figure 2.31. Stops are represented by a closure phase, the start of release (high energy) and a release phase (low energy). These sounds would be represented by a sequence of two or more phonetic elements, each having its own table of transition rules.

2.5.2.2 The MITalk System

The MITalk text-to-speech system [1] is the result of over 25 years of research developed by a vast team of engineers and linguists. The entire system architecture was carefully coordinated from individual contributions focusing on particular issues such as speech analysis and processing, linguistics and speech synthesis. The aim of this system was to provide high quality speech from unrestricted English text. The system consisted of an analysis phase followed by a synthesis phase both to be described functionally in the sections bellow.

The success of this system is largely due to an efficient knowledge transfer from analysis to rule development allowing to choose strategically important descriptive problems [17].

2.5.2.2.1 Analysis There is a considerable amount of pre-processing of the input to be done before converting text to speech. Figure 2.32 shows, in the form of a block diagram, the initial stages of text processing.

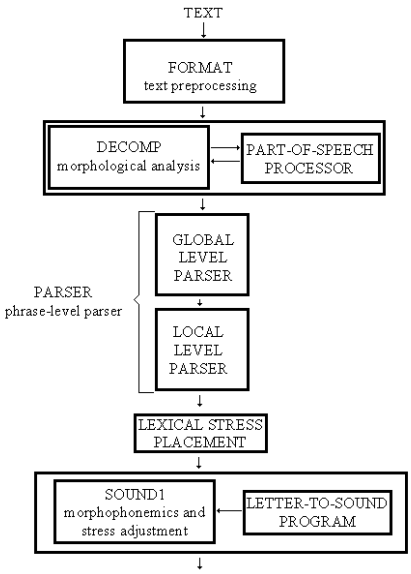

Figure 2.32: Analysis blocks of the MITalk system.

The first module FORMAT of the MITalk system performs the conversion of the original unrestricted English text to be analysed to a sequence of words and punctuation marks recognizable by later modules.

In an initial analysis phase text is converted to a narrow phonetic transcription consisting of phonetic symbols and prosodic markers. The minimum syntactic unit of language is the morpheme used to determine word pronunciation. The MITalk system uses a morph lexicon containing 12000 entries, sufficient to analyse ten times that number of English words, giving correct morph analysis, pronunciation and parts of speech [1]. A dictionary is also used to provide exception rules for pronunciations. The morphs used in MITalk determine pronunciations providing high quality phonetic segment label sequences that are used for synthesis.

The morphological analysis in the MITalk systems is provided by the DECOMP module. The input records contain either a word or a pronunciation mark provided by the output of the preceding block and the output consist of a sequence of decomposed word records (morph spelling, morph pronunciation, morph type and parts of speech). The decomposition process uses alternative algorithms to the concatenation of morph spellings to provide an improved analysis.

The PART-OF-SPEECH PROCESSOR is part of the DECOMP module in the text-to-speech system. It computes a part of speech set for each word in the input, given the morph decomposition and the parts of speech morphs.

The PARSER is able to handle arbitrary text in real time and supplies a surface structure parse, providing information for F0TARG as shown in Figure 2.33. The input file contains information the parser reads from DECOMP on the words in the text, one sentence at a time. It then attempts to find phrases in the sentences. The parsing module can be divided in two levels:

- the GLOBAL LEVEL PARSER reflecting the parsing strategy which has been found to give the best phrases;
- the LOCAL LEVEL PARSER which interprets the augmented transition network (ATN) grammar used to recognize the phrases.

The output of the PARSER is a series of nodes representing words in a phrase.

The stress rules implemented in the module LEXICAL STRESS PLACEMENT were written to take advantage of known parts of speech. They include 3 stress levels (primary stress, less than primary stress and no stress) and suffix dependent stress categories.

The module SOUND1 checks the contexts in which the pronunciation at morph boundaries occur and changes the pronunciation. It also adjusts the lexical stress and performs letter to sound conversion. The pronunciation for each word is constructed by concatenating the pronunciations of its components. The input to SOUND1 contains word and morph pronunciation information from DECOMP and phrase and part of speech information produced by PARSER. The output is a set of phonetic symbols, stress marks, syllable and morph boundaries for each word.

The LETTER-TO-SOUND PROGRAM converts letter strings into stressed phonetic segment label strings. It processes words which were not segmented by DECOMP. This module stipulates a pronunciation for words not analyzable by the lexical analysis algorithm. The rules were developed by a process of extensive statistical analysis of English words.

2.5.2.2.2 Synthesis The MITalk system synthesis routines are used in unrestricted text-to-speech applications to produce natural sounding speech.

The phonological component PHONOL is divided into two modules as shown in Figure 2.33. PHONO1 uses information from the PARSER to specify syntactic markers and PHONO2 selects an appropriate allophone for each phoneme. PHONOL also includes rules for inserting pauses in various text locations. The input to PHONOL is the set of data resulting from text analysis. The input to PHONO1 consists of phonetic symbols, a lexical stress pattern and syntactic information. The output is a set of symbols for each sentence.

The input to the prosodic component PROSOD is a phonological representation of the sentences (output of PHONOL) and the output consists of a

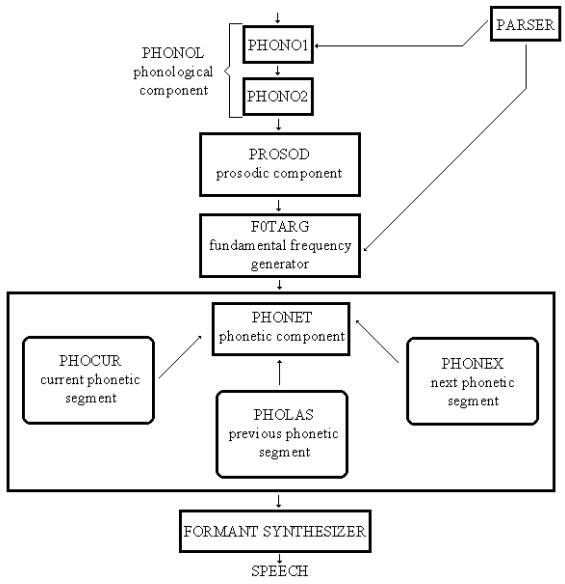

Figure 2.33: Synthesis blocks of the MITalk system. After Allen, Hunnicutt and Klatt [1].

string of phonetic segments with additional information about stress and duration. The factors that influence segmental durations (percentage increase or decrease in duration of the segment) are modelled by the following formula

$$DUR = \frac{(INHDUR - MINDUR) \times PRCNT}{100} + MINDUR,$$

where INHDUR is the inherent duration, MINDUR is the minimum duration and PRCNT is the percentage of shortening determined by applying a set of rules. The resulting durations are determined in part by a variable that controls the speaking rate SPRATE. The rules account for 84% of the observed total variance in the segmental durations.

The fundamental frequency generator F0TARG uses information from both PARSER and PROSOD components to generate two F0 target values for each phonetic segment (one at onset and one at mid point value). This allows the use of interpolation techniques to calculate F0 values for each segment. The input file contains information about the phrase structure of each sentence used to determine the declination lines, to calculate the amount of

excursion from the declination line through each phrase and to insert continuation lines. Lexical markers, syllable division and parts of speech are also included in the input file to F0TARG. They are used to determine the F0 maximums and provide information needed to determine the relative hight of the peaks. The output from F0TARG contains two target values for each segment.

The phonetic component PHONET produces a set of target values for most of the synthesizer control parameters for various phonetic segment types. A smooth transition between target values is calculated from time constants computed by rule and the parameter value at the segment boundary. These constants are calculated from values of the previous phonetic segment PHOLAS, the current phonetic segment PHOCUR and the next phonetic segment PHONEX as shown in Figure 2.33.

In a consonant to vowel transition some consonants present high variance but the vowel spectrum is rather invariant and so a locus theory equation with two variables can be used to predict the value of the first formant at voicing onset from a known vowel target. This forms the basis to a synthesis algorithm for predicting CV formant transitions. The vowel and consonant formant values are first defined in terms of strait line segments and then the formant values at the CV boundaries are determined by a locus theory equation.

The general algorithm used in the MITalk system for predicting control parameters is:

1. draw the target value for the first segment;
2. draw the target value for the next segment;
3. smooth the boundary between the segments using one of the templates in Figure 2.34 (note that DISCON does no smoothing);
4. go to 2 unless there are no more segments.

The transitions between target values are determined by control parameters computed by rules. The MITalk system normally uses a smooth template

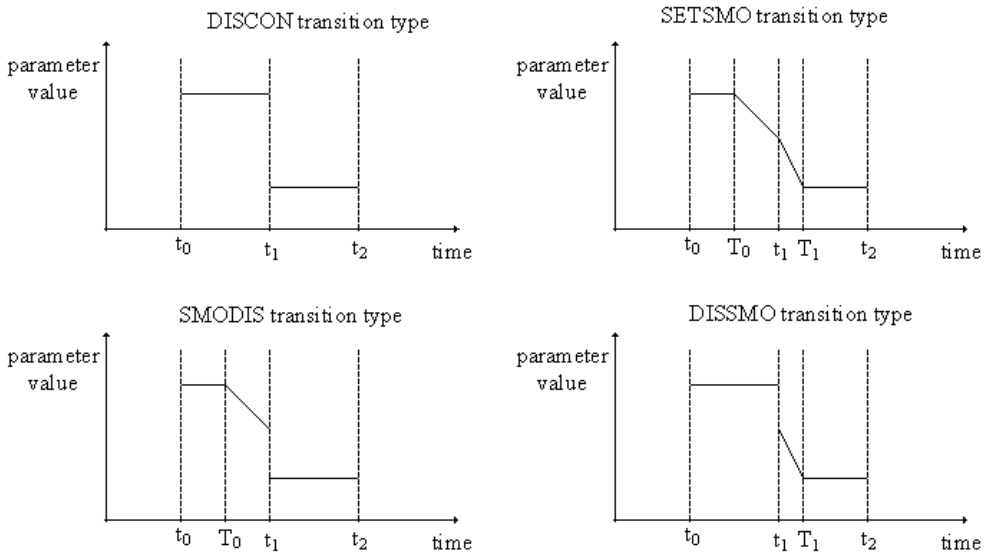

Figure 2.34: Templates for smoothing adjacent phonetic segment targets. After Allen, Hunnicutt and Klatt [1].

(SETSMO) to model the transitions, but must resort to partially discontinuous templates (SMODIS and DISSMO) for some nasal and plosive sounds. The discontinuous template (DISCON) is only used for parameters relating to the excitation signal, e.g. pitch and amplitude of voicing.

The final module is a FORMANT SYNTHESIZER described in [31] and [32]. This produces the actual output speech waveform controlled by the PHONET output parameters specifying the actions of the software synthesizer.

2.5.2.3 The Laureate Text-to-Speech System

The Laureate text-to-speech system [48], developed at the BT Laboratories, performs text analysis, prosody synthesis and speech sound selection, as shown in Figure 2.35. The system searches its database for the appropriate triphones [1] but will use diphones or even monophthongs if these are more

[1] Speech unit that contains the transitions from steady state portions of neighbouring phones into and out of a phone [14].

adequate to model transitions. Once the speech units are selected, smooth concatenation of speech fragments is ensured by pitch synchronous interpolation of the analysis and synthesis filter parameters. Synthetic prosody is applied in the form of a segmental duration modification algorithm and pitch modification algorithm. The system allows different models of natural speech to be integrated in the same architecture. Laureate provides a set of standard linguistic representations (text analysis) and production models (prosody and speech generation) [15] [14].

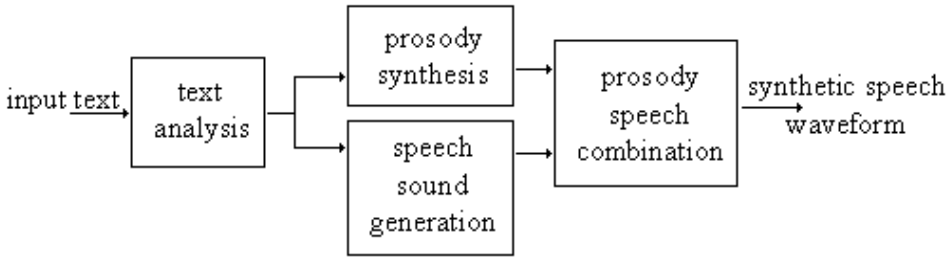

Figure 2.35: The Laureate text-to-speech conversion system. After Page and Breen [48].

The Laureate system consists of two sections: the core components and the satellite components, as shown in Figure 2.36.

- The core contains the linguistic object which is a dynamic relational database of information that only allows certain linguistic relationships between records (e.g. words contain syllables).
- The satellite components are external linguistic structures which access the linguistic object through a formalized linguistic interface, that ensures the linguistic data can be used by other system defined structures.
- The text normalization block produces linguistic units with a variable length (from single words to paragraphs).
- The pitch assignment component is used to produce accent information needed to generate a contour.

- The realization component is used to generate the actual pitch values as defined by the accent information produced by the linguistic object.

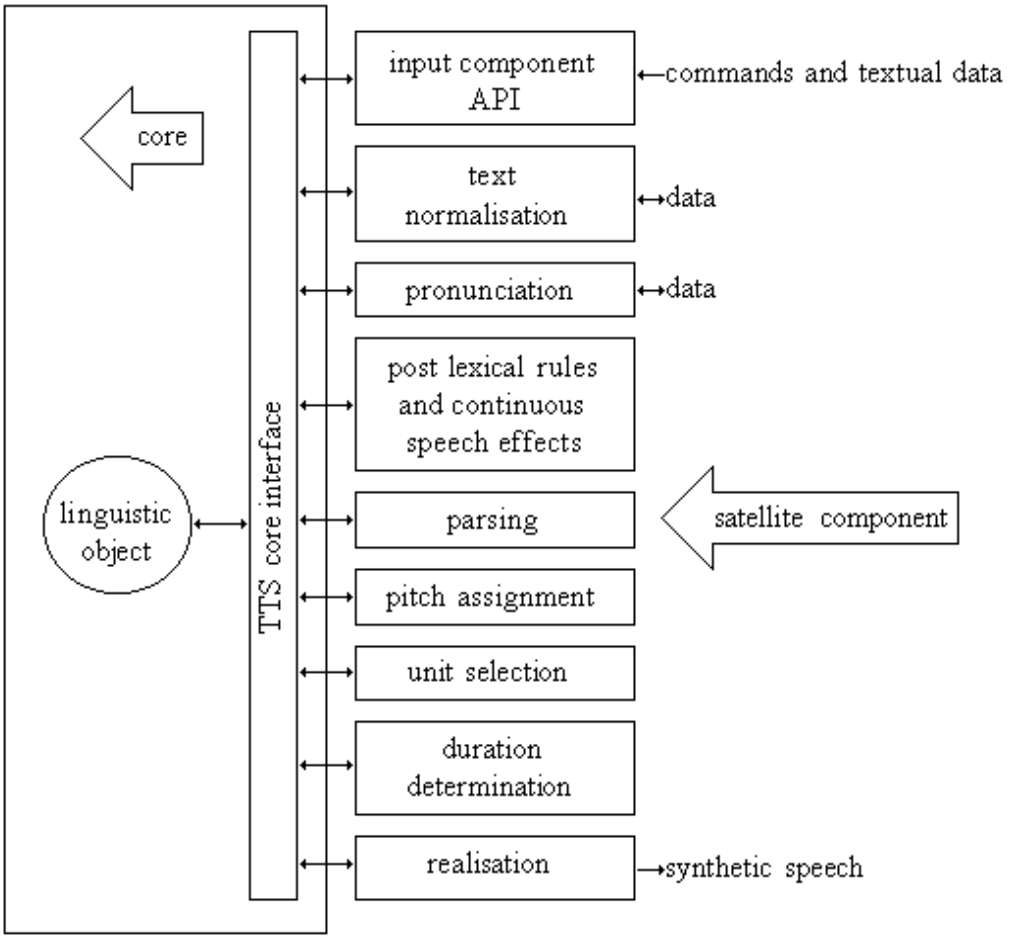

Figure 2.36: The Laureate architecture. After Page and Breen [48].

The Laureate architecture provides data store separate from the linguistic object, making possible the use of language specific data. Its modular design allows for isolated component changes. If a new structure (different voices, accents and languages) is required, it can be incorporated into the system with minimal modifications of the existing ones. A number of software tools,

such as an abbreviation dictionary, an acronyms dictionary and a pronunciation dictionary, can be adjusted to meet the demands of application specific data. The Laureate system has been easily integrated into a number of different BT prototype systems.

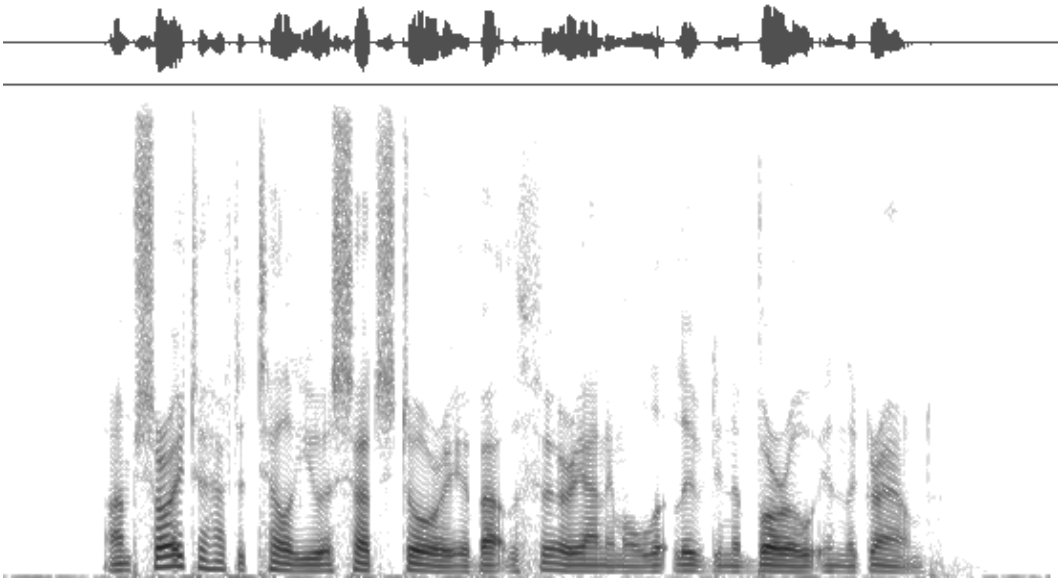

Figure 2.37: Signal and Spectrogram of the sentence "I'm sorry to have to tell you the sad story about the furry animal that lived in the burrow in the ground" spoken by an English northern male.

The Laureate text-to-speech system produces natural sounding speech. The natural and synthesized spectrograms of an example sentence, not used for training the speech models, is shown in Figure 2.37 and Figure 2.38. This illustrates the synthesizer ability to reproduce the spectral contours of natural speech.

The Laureate text-to-speech system is described more fully by Page, Breen, Edgington, Lowry, Jackson and Minnis [48] [15] [14].

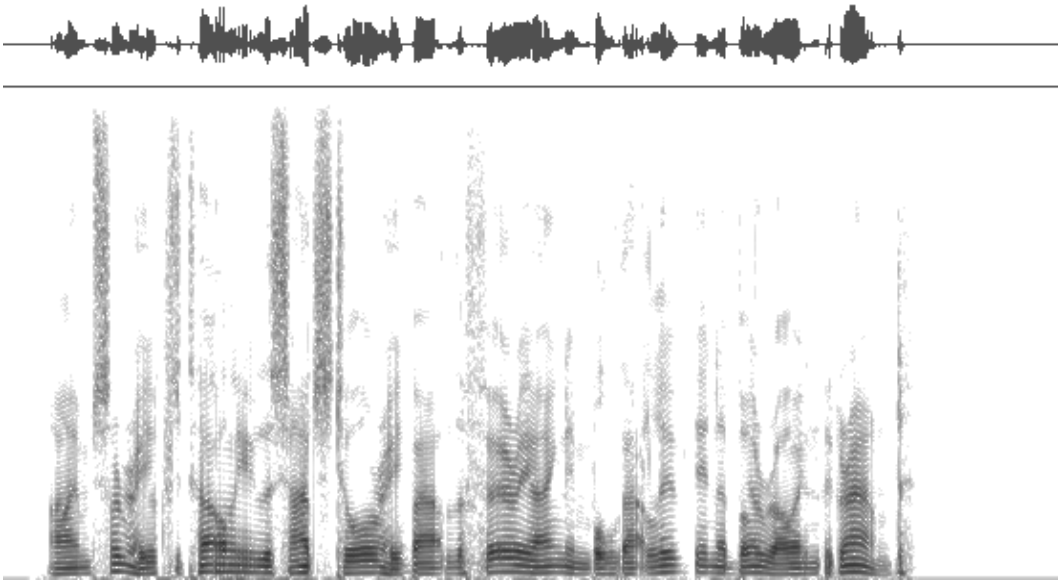

Figure 2.38: Signal and Spectrogram of the sentence "I'm sorry to have to tell you the sad story about the furry animal that lived in the burrow in the ground" synthesized by the Laureate text-to-speech system (speaker pc16k).

2.6 Models of Co-articulation

A literature review of methods for modelling co-articulation is presented in the following sections. Co-articulation effects must be reproduced in synthetic speech in order to produce a natural voice quality.

2.6.1 Neural Speech Synthesis

Cawley and Green [7], and Cawley and Noakes [8] proposed an alternative approach to generate the formant contours for the Holmes parallel formant synthesizer. They used a multi-layer backpropagation network to generate the control parameters corresponding to a sequence of allophone tokens. A neural network was firstly trained to model co-articulation in the words "pit", "pat", "pot", "put", "putt", "bit", "bat", "bot", "but" and "butt". Figure 2.39 shows the frequency contour of the first formant for the word "pit". The formant data was later extended to a list of allophones provided by the JSRU rule system from a corpus of 25 English sentences. The network considered

two allophones at a time, producing formant data for the diphone formed by their adjacent halves. The quality of the synthesized speech was similar to that of the JSRU formant synthesizer. A large amount of training data would be required to produce naturally sounding speech for an unlimited vocabulary.

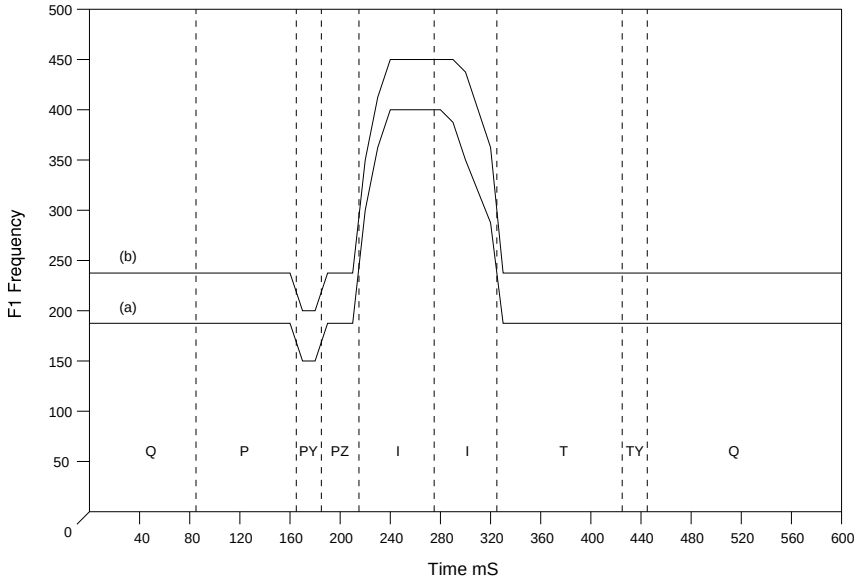

Figure 2.39: (a) Frequency contour produced by the JSRU synthesizer. (b) Frequency contour reconstituted using a cubic spline from samples generated by the network (displaced by 50Hz). After Cawley and Green [7].

2.6.2 Speech Coding Using B-Spline Curves

A first attempt at speech coding using a speech production model with continuously varying parameter values is that of Gouveia [20]. B-spline segments were used to represent LP parameter waveforms, found using a curve fairing method. The average SNR [2] introduced by using 4 cubic B-spline curves to

[2]Let $s(n)$ denote a noise free speech signal at time n and $\hat{s}$ the corresponding processed signal [12]. The resulting SNR measure (in dB) is obtained as
$\mathrm{SNR} = 10\log_{10}\frac{\sum_n s^2(n)}{\sum_n [s(n)-\hat{s}(n)]^2}$.

fit a 100ms segment was 10.07dB. This was not a significant improvement to the conventional LP model with fixed parameter values (SNR=9.84dB). The coding was also experimentally evaluated by performance analysis of the LP vocoder. It was possible to produce intelligible speech.

Speech coding by time-varying models did not present any clear benefits. This was due firstly to the poor LP coding method and secondly to the lack of knowledge about the nature of parameter trajectories in each segment of speech. The computational requirements were found to be enormous, introducing significant delays in coding.

2.6.3 Interpolation of LSP Coefficients Using Recurrent Neural Networks

Kohata [33] used a recurrent neural network to interpolate LSP parameters in order to reduce the bit rate in speech coding and increase the duration of interpolation to 100ms. The segmentation method partitioned the LSP coefficients at stationary targets and defined transition regions in the middle of frames. The lengths of signals were normalized. A low spectral distortion at a bit rate of 287 bit/s was obtained after coding and interpolation of the LSP parameters. Speech was synthesized using the LSP coefficients and raw residual excitation signals.

2.6.4 Modelling Co-articulation in Speech Recognition

Sun [57] is currently developing a method to model speech transitions by a sequence of feature vectors interpolated among a set of anchor points corresponding to the target phonetic units. Transitions are modelled by smoothing spline based trajectories derived from the neighbouring target units. Sun suggested that it is more important to model the target positions precisely than the paths of the intermediate positions (co-articulation effects).

The speech signal x(t) is modelled by a smooth interpolation function plus a random part e(t) as follows

$$x(t) = f(t; y(s_1), ..., y(s_k), t(s_1), ..., t(s_k)) + e(t).$$

x(t) is a sequence of speech feature vectors (Mel-frequency cepstral coefficients and their differences) calculated from a window at time t=1,...,T. $(y(s_1), ..., y(s_k))$ is the sequence of target feature vectors corresponding to the underlying sequence of target phonetic units $(s_1, ..., s_k)$ of utterance x(t). $(t(s_1), ..., t(s_k))$ is the sequence of knots for interpolation.

Results from a phoneme classification experiment were presented. Context dependent models were compared to this new context independent model. The distributions of the target phonetic units are more concentrated than those calculated by hidden Markov models (HMM).

2.7 Summary

This chapter provided a description of the human speech production system. The characteristics of different speech sounds and the concept of co-articulation were introduced. The articulatory, formant and LPC vocal tract models were described. Since concatenative speech synthesis systems use an approach related to the one used in this work a literature review of such systems is provided. A discussion of speech synthesis by rule systems is also included concentrating on the methods used to blend adjacent speech units and model co-articulation.

The chapter is particularly concerned with the description of existing models of co-articulation because the work presented in this thesis uses a new model of co-articulation to be described in the following chapter.

Chapter 3

Bézier Model of Co-articulation

3.1 Introduction

This chapter presents an introductory discussion of the Bézier model of co-articulation and some applications of this technique. It also provides background information on Bézier curves and curve fitting techniques used to form the proposed Bézier model.

3.2 The Bézier Model of Co-articulation

Conventional interpolation methods have used straight line templates to model parameter transitions [26]. Figure 3.1 shows the Holmes-Mattingly-Shearme based template (described in Section 5.2.2) used to fit the phoneme "p3". Since the curve to which they are required to fit is of a smooth nature it is to be expected that a closer and more natural fit might be obtained using a parametric curve, as shown in Figure 3.2, rather than piecewise linear interpolation. However, some transitions for instance are quite abrupt and discontinuous, requiring a model involving more than one curve segment.

A good model of co-articulation is vital for natural sounding speech. The Bézier model of co-articulation includes the following features:

- The Bézier curve shape is well suited to the nature of transitions.

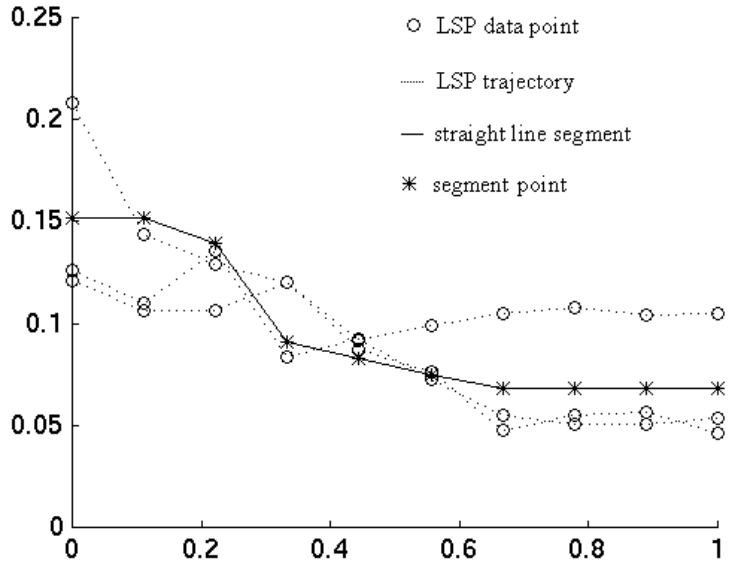

Figure 3.1: The second LSP parameter trajectory of diphone as in "p3" (all examples in the speech corpus) and the straight line template model.

- The mathematical simplicity of the model makes it ideal for the curve fitting technique.
- Since the curve is contained between two anchor points and the "speed" of the transition is determined by two control points, the Bézier curve parameters have a meaningful interpretation.

3.2.1 Speech Coding Using Bézier Curves

The polygon points defining a Bézier curve may form an attractive representation of speech parameters, providing a considerable amount of data compression, as it exploits the inter-frame redundancy in the speech parameters. The cubic curve passes through two data points (B_0 and B_3) located at the phonetic targets and two control points (B_1 and B_2) that determine the gradient of the segment. This is an extremely intuitive way of controlling the nature of the transition according to phonetic context.

Vector quantization [22] of Bézier curves could be used to form a "phonetic

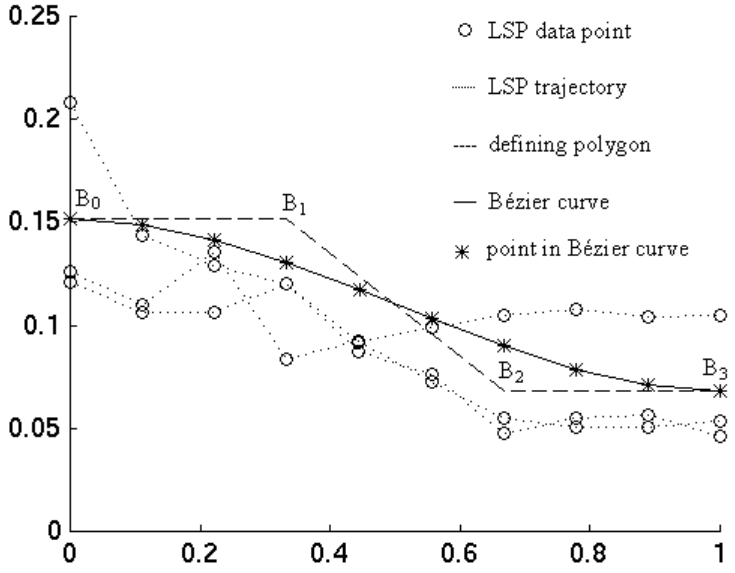

Figure 3.2: The second LSP parameter trajectory of diphone as in "p3" (all examples in the speech corpus) and the parametric curve (cubic Bézier curve) model.

vocoder". This technique would exploit the underlying structure of the input vectors (B_0, B_1, B_2 and B_3) for the purpose of data compression. Diphones abut during the steady state conditions in the middle of each phoneme and so the Bézier models of each diphone are merged at these points, as to ensure a smooth transition between each diphone.

3.2.2 Speech Parameters

This section deals with the choice of speech parameters used. Linear predictive coding attempts to model the spectral properties of a signal using a fixed number of poles. Some of these poles are used to model the broad spectral envelope of the speech signal. For voiced speech, the available poles are engaged in modelling formants. The spectra of unvoiced sounds do not generally exhibit well defined formant structure and so the remaining poles tend to model spurious features.

Formant parameters directly encode the perceptually important spectral fea-

tures of speech and so they exhibit a very low spectral sensitivity. Formant coding would clearly be the optimal coding method for use in this work. However formant analysis of speech is computationally intensive and has not yet been automated, and so linear predictive coding provides an automatic and computationally efficient coding technique.

3.3 Bézier Curves

A parametric curve is defined by a set of coordinate points represented as a function of a single parameter [52]. The parameter values determine the position vector of a point in the curve. Let t be the curve parameter, then the curve coordinates are defined as

$$x = x(t)$$

and

$$y = y(t).$$

The position vector of a point on the curve is given by

$$P(t) = [x(t) \quad y(t)]\,.$$

The parametric form is axis independent because a point is specified by a single value. The curves beginning and end points are fixed by the parameter range, which is extremely useful if, for example, we normalize the parameters to the range $0 \leq t \leq 1$.

A Bézier segment is a parametric curve, defined by a polygon (a quadrilateral in the case of the cubic curves), as shown in Figure 3.3. The polygon consists of two data points (B_0 and B_3) and one or more control points (B_1 and B_2); the curve passes through each of the data points, while the control points determine the initial direction of the curve at each data point. These curves were first defined, by Bézier [3], in terms of an initial point and a series of incremental vectors constituting successive sides of the defining polygon. Forrest [18] later developed the widely accepted formulation of the Bézier curve in terms of polygon vertices and recognized that the basis functions were Bernstein polynomials.

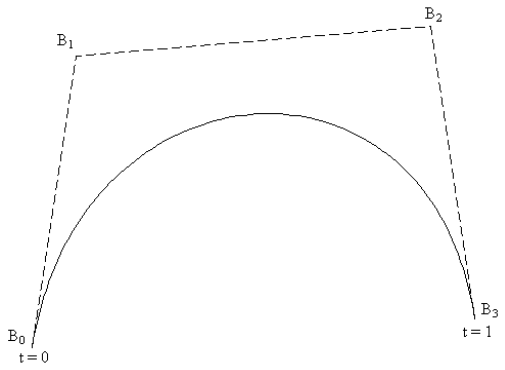

Figure 3.3: Cubic Bézier curve and its four point defining polygon. B_0 and B_3 are data points. B_1 and B_2 are control points.

The following discussion is based on the notation used by Rogers and Adams [52].

3.3.1 The Mathematical Model

This section describes the mathematical model and the parametric representation of a Bézier curve [4]. Let $P(t)$ denote a Bézier curve of order n. Let B_i be the coordinates of the i^{th} vertex of the defining polygon. Then a Bézier curve is defined by

$$P(t) = \sum_{i=0}^{n} B_i J_{n,i}(t) \qquad 0 \leq t \leq 1,$$

where

$$J_{n,i}(t) = \binom{n}{i} t^i (1-t)^{n-i}.$$

$J_{n,i}(t)$ denotes the basis function (a Bernstein polynomial), as shown in Figure 3.4, associated with the i^{th} vertex of a n-sided polygon, with

$$\binom{n}{i} = \frac{n!}{i!(n-i)!}.$$

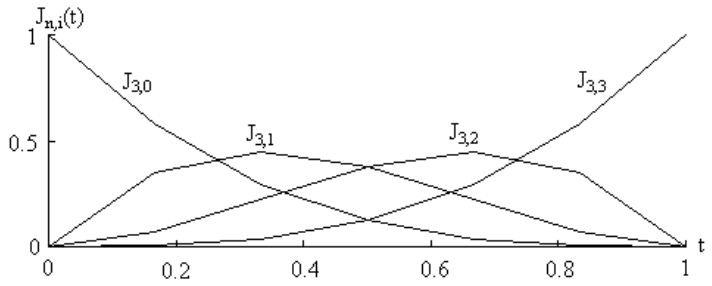

Figure 3.4: Basis functions associated with a cubic Bézier curve (n=3).

For computational reasons it is often convenient to let n=N-1 and i=I-1, the equation then becomes

$$P(t) = \sum_{I=1}^{N} B_I J_{N,I}(t),$$

where

$$J_{N,I}(t) = \binom{N-1}{I-1} t^{I-1}(1-t)^{N-I},$$

with

$$\binom{N-1}{I-1} = \frac{(N-1)!}{(I-1)!(N-I)!}.$$

3.3.2 The Properties of Bézier Curves

The basis functions of Bézier curves are Bernstein polynomials [19]. Among the interesting features of these functions, we note the following:

- the curve follows the shape of the defining polygon;
- the first point on the Bézier curve and the first point on its defining polygon are coincident

  $$P(0) = B_0;$$

- the last point on the Bézier curve and the last point on its defining polygon are coincident

$$P(1) = B_n;$$

- the tangent vectors at the ends of the curve have the same direction as the respective polygon sides;
- the curve is restricted to the bounding polygon (convex hull property);
- the maximum value of each blending function occurs at t=i/n and is given by

$$J_{n,i}\left(\frac{i}{n}\right) = \left(\frac{n}{i}\right)\frac{i^i(n-i)^{n-i}}{n^n};$$

- for any given value of the parameter t, the summation of the basis functions is precisely one

$$\sum_{i=0}^{n} J_{n,i}(t) = 1.$$

3.3.2.1 Continuity Conditions Between Adjacent Bézier Curves

The conditions for first derivative continuity between two Bézier polynomials may be obtained from the expressions for end derivatives. Let P(t) denote the Bézier segment defined by vertices B_i and Q(t) an adjacent Bézier segment defined by vertices C_i both of degree 3. If we pretend to concatenate the two curves with a smooth transition, as shown in Figure 3.5, first derivative continuity can be assured by

$$P'(1) = Q'(0)$$

where

$$Q'(0) = 3(C_1 - C_0),$$

and

$$P'(1) = 3(B_3 - B_2).$$

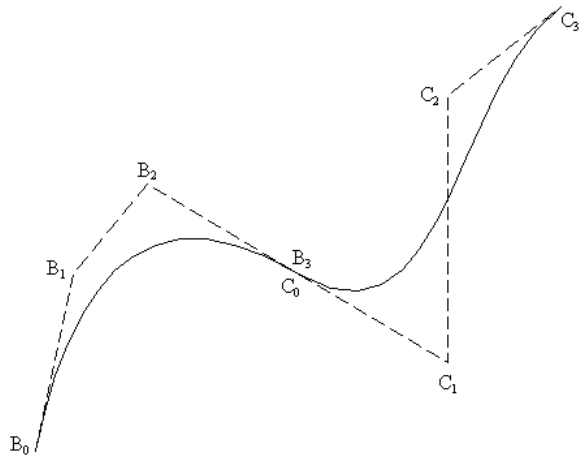

Figure 3.5: First derivative continuity.

Since $C_0 = B_3$ the previous equations yield

$$C_1 = (B_3 - B_2) + B_3.$$

Thus the tangent vector directions and magnitudes at the joint are the same if the vertices B_2, $B_3 = C_0$ and C_1 are collinear and $B_3 = C_0$ is the mid point, that is

$$C_1 - C_0 = B_3 - B_2 = C_0 - B_2$$

or

$$C_1 + B_2 = 2C_0 = 2B_3.$$

3.3.3 Increasing Flexibility

It is often found that a particular curve segment is not sufficiently flexible to adopt a desired shape. This difficulty may be resolved by using the technique described by Bartels, Beatty and Barsky [2]. The Bézier curve is progressively subdivided into two new Bézier curves as shown in Figure 3.6, that combined give us the desired shape, shown in Figure 3.7. This is accomplished by calculating the mid-points of the lines, as shown in Figure 3.8, thus increasing the number of defining polygon points.

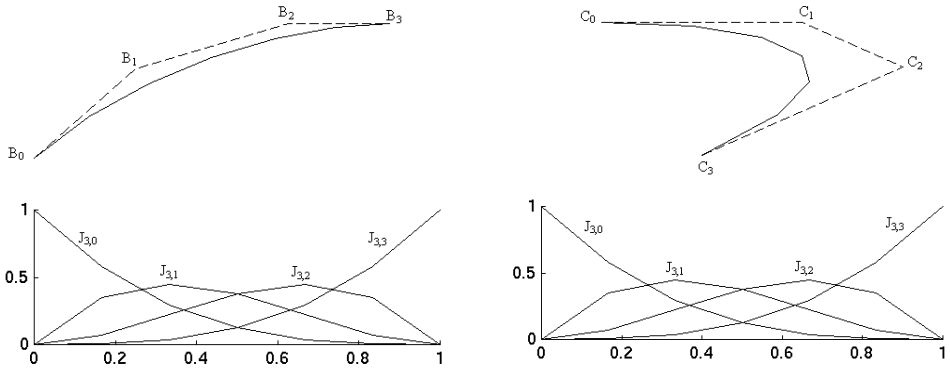

Figure 3.6: Bézier curves and basis functions.

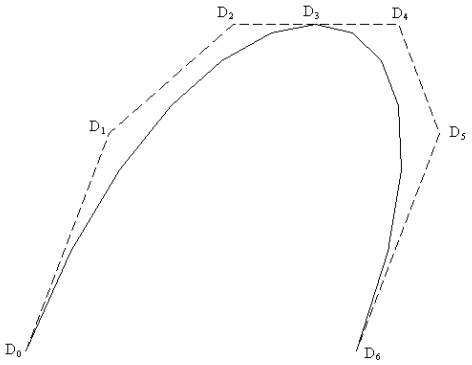

Figure 3.7: Bézier curve with additional flexibility.

To obtain mid-points $((x_{12}, y_{12}), (x_{23}, y_{23}), (x_{34}, y_{34}), (x_{123}, y_{123}), (x_{234}, y_{234})$ and $(x_{1234}, y_{1234}))$ of the lines defined by initial four points $((x_1, y_1), (x_2, y_2), (x_3, y_3)$ and $(x_4, y_4))$ we proceeded as follows

$$x_{12} = \frac{x_1 + x_2}{2}, \quad y_{12} = \frac{y_1 + y_2}{2}$$

$$x_{23} = \frac{x_2 + x_3}{2}, \quad y_{23} = \frac{y_2 + y_3}{2}$$

$$x_{34} = \frac{x_3 + x_4}{2}, \quad y_{34} = \frac{y_3 + y_4}{2}$$

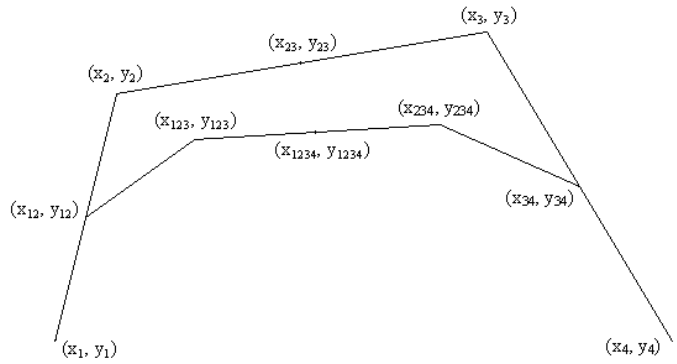

Figure 3.8: Mid-point subdivision.

$$x_{123} = \frac{x_{12} + x_{23}}{2}, \quad y_{123} = \frac{y_{12} + y_{23}}{2}$$

$$x_{234} = \frac{x_{23} + x_{34}}{2}, \quad y_{234} = \frac{y_{23} + y_{34}}{2}$$

$$x_{1234} = \frac{x_{123} + x_{234}}{2}, \quad y_{1234} = \frac{y_{123} + y_{234}}{2}$$

The two new Bézier curves are drawn using the polygons defined by the points $((x_1, y_1)$, (x_{12}, y_{12}), (x_{123}, y_{123}) and $(x_{1234}, y_{1234}))$ and $((x_{1234}, y_{1234})$, (x_{234}, y_{234}), (x_{34}, y_{34}) and $(x_4, y_4))$.

3.3.4 Matrix Representations

The matrix forms for curve generation are very compact to write, simple to program and clear to understand. The equation for a Bézier curve can be expressed in a matrix form

$$P(t) = [F][G],$$

where

$$[F] = [J_{n,0} \quad J_{n,1} \quad ... \quad J_{n,n}],$$

and

$$[G]^T = [B_0 \quad B_1 \quad ... \quad B_n].$$

Cohen and Riesenfeld [11] generalized this representation to

$$P(t) = [T][N][G],$$

where

$$[T] = [t^n \quad t^{n-1} \quad ... \quad t \quad 1],$$

and

$$[N] = \begin{bmatrix} \binom{n}{0}\binom{n}{n}(-1)^n & \binom{n}{1}\binom{n-1}{n-1}(-1)^{n-1} & ... & \binom{n}{n}\binom{n-n}{n-n}(-1)^0 \\ \binom{n}{0}\binom{n}{n-1}(-1)^{n-1} & \binom{n}{1}\binom{n-1}{n-2}(-1)^{n-2} & ... & 0 \\ \binom{n}{0}\binom{n}{1}(-1)^1 & \binom{n}{1}\binom{n-1}{0}(-1)^0 & ... & 0 \\ \binom{n}{0}\binom{n}{0}(-1)^0 & 0 & ... & 0 \end{bmatrix},$$

where the individual terms are given by

$$(N_{i+1,j+1})_{i,j=0}^{n} = \begin{cases} \binom{n}{j}\binom{n-j}{n-i-j}(-1)^{n-i-j}, & 0 \le i+j \le n \\ 0, & other \end{cases}$$

[N] can be decomposed into a sometimes more convenient form

$$[N] = [C][D],$$

where

$$[C] = \begin{bmatrix} \binom{n}{n}(-1)^n & \binom{n}{1}\binom{n-1}{n-1}(-1)^{n-1} & ... & \binom{n}{n}\binom{n-n}{n-n}(-1)^0 \\ \binom{n}{n-1}(-1)^{n-1} & \binom{n}{1}\binom{n-1}{n-2}(-1)^{n-2} & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \binom{n}{1}(-1)^1 & \binom{n}{1}\binom{n-1}{0}(-1)^0 & ... & 0 \\ \binom{n}{0}(-1)^0 & 0 & ... & 0 \end{bmatrix},$$

and

$$[D] = \begin{bmatrix} \binom{n}{0} & \dots & 0 \\ \vdots & \binom{n}{1} & \vdots \\ \vdots & \ddots & \vdots \\ 0 & \dots & \binom{n}{n} \end{bmatrix}.$$

3.4 Curve Fitting

The research described in this thesis aims to represent diphone transitions by fitting existing speech data to a model depending nonlinearly [34] on a set of adjustable parameters using the Levenberg-Marquardt method. The following sections introduce this curve fitting procedure.

3.4.1 The Method of Least Squares

Curve fitting means to fit a function to a set of given data. Gauss proposed a widely used procedure to fit a straight line through a set of graph points. The method of least squares defines a straight line model

$$y(x) = a + bx,$$

to be fitted through a set of N data points (x_i, y_i), so that the sum of the squared vertical distances from the points to the straight line is minimum. The sum of squares is defined as

$$\chi^2(a, b) = \sum_{i=1}^{N} (y_i - a - bx_i)^2 .$$

A necessary condition for χ^2 to be minimum is that its partial derivatives $\frac{\partial \chi^2}{\partial a}$ and $\frac{\partial \chi^2}{\partial b}$ are equal to zero. These are the components of a gradient vector. Finding the values a and b that minimize χ^2 is often difficult and so a solution by iteration is generally used, where the search process starts at some point and moves toward a point where χ^2 is minimum.

3.4.1.1 The Gradient Method

Cauchy introduced a method to find a minimum of a real valued function $f(x_1, ..., x_N)$ by repeatedly computing the minimum of a function g(t). Let X_0 be a minimum and x the start point. The method of steepest descent (or gradient method) calculates the minimum of f closest to x along a straight line in the direction of maximum decrease

$$-\nabla f(x_1, ..., x_N),$$

where ∇f is the gradient of f. That is, the method determines the value t and the next approximation of X_0

$$z(t) = x - t\nabla f(x_1, ..., x_N),$$

at which the function

$$g(t) = f(z(t))$$

has a minimum.

Convergence can sometimes be slow and so more refined methods, such as the ones presented in the following sections, have been proposed.

3.4.2 The Method of Damped Least Squares

The standard methods for solving least squares problems may fail to improve the initial solution in applications involving the approximation of one function by another. Levenberg [36] proposed a solution to this problem by an extension of the standard method which ensured improvement of the initial solution. In this new method the absolute values of the increments of the parameters are limited or damped in order to improve the first order Taylor approximations and to minimize simultaneously the sum of the squares of the approximating residuals under these damped conditions. The *method of damped least squares* solved, with a comparatively rapid rate of convergence due to greater freedom given to individual values, types of problems with higher degree of complexity than those to which the least squares method was usually applied.

3.4.3 The Maximum Neighbourhood Method

The first algorithms for the least squares estimation of nonlinear parameters usually used one of the two methods: Taylor series expansions with corrections of the several parameters calculated at each iteration on the assumption of local linearity, or steepest descent (gradient). Marquardt [40] proposed a new method to overcome numerous problems posed by early methods, namely, divergence of successive iterates and slow convergence after the first few iterations. The *maximum neighbourhood method* performs an optimum interpolation between the Taylor series method and the gradient method.

Various modifications of the steepest descent method compensate partially for slow convergence due to poor conditioning of the error criterion. In Taylor series and gradient methods it is necessary to control the step size carefully once the direction of the correction vector has been established. In the maximum neighbourhood the direction and step size are determined simultaneously. By this algorithm we always obtain a feasible neighbourhood and we almost always obtain the maximum neighbourhood in which the Taylor series give an adequate representation for all purposes.

The maximum neighbourhood method combines the best features of its predecessors while avoiding their most serious limitations. It is possible to converge from an initial guess which may be outside the region of convergence of other methods and rapidly close in on the converged values after their neighbourhood has been determined.

3.4.4 The Levenberg-Marquardt Method

Given a set of observations $(\mathbf{x_i}, \mathbf{y_i})$ (this work uses 12 separate vectors (x_1, y_1), ..., (x_{12}, y_{12})), data (LSP parameters) is fitted to a model that depends nonlinearly on adjustable parameters a_j,

$$y(x_i; \mathbf{a}) = \sum_{j=1}^{M} a_j X_j(x_i),$$

where $X_1(x_i), ..., X_M(x_i)$ are Bézier basis functions and M is the number of data points.

We choose a merit function that measures the agreement between the data and the model with a particular choice of parameters. Best-fit parameters are then determined by the iterative minimization of the cost function

$$\chi^2(\mathbf{a}) = \sum_{i=1}^{M} \left[\frac{y_i - y(x_i; \mathbf{a})}{\sigma_i} \right]^2 ,$$

where σ_i is the measurement error (standard deviation). The minimum of the merit function $\chi^2(\mathbf{a})$ occurs where the derivative of χ^2 with respect to all M parameters $\mathbf{a}$ is equal to zero.

The gradient of χ^2 with respect to the parameters $\mathbf{a}$ is given by

$$\frac{\partial^2 \chi^2}{\partial a_k \partial a_l} = 2 \sum_{i=1}^{M} \frac{1}{{\sigma_i}^2} \left[\frac{\partial y(x_i, ; \mathbf{a})}{\partial a_k} \frac{\partial y(x_i, ; \mathbf{a})}{\partial a_l} - [y_i - y(x_i; \mathbf{a})] \frac{\partial^2 y(x_i; \mathbf{a})}{\partial a_l \partial a_k} \right] ,$$

where

$$\frac{\partial y(x_i, ; \mathbf{a})}{\partial a_k} \frac{\partial y(x_i, ; \mathbf{a})}{\partial a_l}$$

is the Jacobian and

$$\frac{\partial^2 \chi^2}{\partial a_k \partial a_l}$$

is the Hessian (second derivative).

To jump from the current approximation a_l to the next approximation a_k the *inverse Hessian method* is used

$$\sum_{l=1}^{M} \alpha_{kl} \delta_{a_l} = \beta_k ,$$

where

$$\beta_k = -\frac{1}{2} \frac{\partial \chi^2}{\partial a_k} = -\frac{1}{2} \nabla \chi^2(a_k),$$

and

$$\alpha_{kl} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[\frac{\partial y(x_i; \mathbf{a})}{\partial a_k} \frac{\partial y(x_i; \mathbf{a})}{\partial a_l} \right] .$$

This set is solved for increments δ_{a_l} that added to the current approximation, give the next approximation ($a_l + \delta_{a_l} = a_k$).

If this is a poor local approximation the *steepest descent method* (step down the gradient) can still be used

$$\delta_{a_l} = constant \times \beta_l,$$

with a small constant.

The components of the Hessian matrix, even if they are not usable in any precise fashion, give some information about the order of magnitude scale of the problem.

The Levenberg-Marquardt algorithm uses the *steepest descent method* when far from the minimum and the *inverse Hessian method* as the minimum is approached, Figure 3.9.

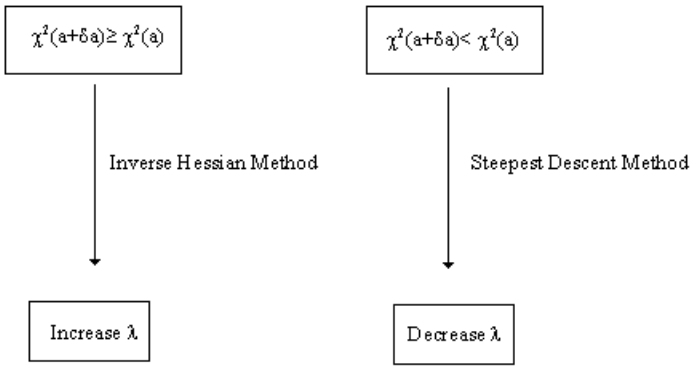

Figure 3.9: Switching between the inverse Hessian method and the steepest descent method. After Grace [21].

The main difficulty is defining a strategy to control the order of magnitude and set the scale of the constant in the previous equation. The scalar λ controls both line search direction and sets the scale of the constant, which we can write as

$$constant = \frac{1}{\lambda \alpha_{ll}}.$$

This algorithm is outlined in [50] as follows:

1. given an initial guess for the set of fitted parameters a_l compute $\chi^2(a_l)$;
2. pick a small value for λ ($\approx$0.001);
3. solve $\sum_{l=1}^{M} \alpha_{kl}\delta_{a_l} = \beta_k$ for δ_{a_l} and evaluate $\chi^2(a_l + \delta_{a_l})$;
4. if $\chi^2(a_l + \delta_{a_l}) \geq \chi^2(a_l)$ increase λ by a substantial factor ($\approx$10) and go back to 3;
5. if $\chi^2(a_l + \delta_{a_l}) < \chi^2(a_l)$ decrease λ by a substantial factor ($\approx$10), update the trial solution ($a_k = a_l + \delta_{a_l}$) and go back to 3.

When $\chi^2(a_l)$ decreases by negligible amount ($\approx$0.001) the iteration process is stopped.

Alternative implementations of the Levenberg-Marquardt algorithm are documented in [21] and [42].

3.5 Summary

During this chapter, the reader has been provided with background information on Bézier curves and curve fitting techniques. The Bézier model of co-articulation, and its use in speech coding and synthesis are also discussed. This provides an introduction to the chapter where the major design decisions are presented (see Chapter 5).

Chapter 4

Data Pre-processing

4.1 Introduction

This work uses an annotated speech corpus which provides information about the recorded speech data, defining the start and end points of the speech units. An LSP analysis is performed on sampled speech data. All of the phoneme pair examples in the speech corpus are recorded sequentially in a data file and their limits registered in a new annotation file.

The main problem in pre-processing the speech data is that the start and end of the frames recorded in the speech corpus annotation files, do not necessarily coincide with the start and end of the frames in the LSP data files. The next stage in data pre-processing is then to calculate the actual diphone data to be fitted using the Bézier model and the new diphone annotation files. This task is particularly important due to the frame alignment problems that make curve fitting more difficult. The LSP speech data is then normalized and resampled, so that the best fit can be found over all of the examples in the diphone inventory. This chapter discusses all of the above procedures in great detail, describing the speech corpus and the data pre-processing procedure.

4.2 The Speech Corpus

The speech corpus used in this work consists of 239 phonetically balanced sentences of neutrally articulated English speech, from a male speaker of a received pronunciation (RP) accent. The corpus provides raw speech data, in the form of 16 bit linear samples at a sampling frequency of 16KHz and time-aligned phonetic transcription, based on the SAMPA phonetic alphabet for each sentence (see Appendix B).

A large corpora containing phonetically transcribed speech such as this, recorded by a single speaker with a very clear voice, is ideal for research in speech synthesis. The problem of using speech databases with several speakers is that the effects of variability amongst speakers are far more prominent than the variability due to co-articulation itself. We are interested in a corpus with a rich variety of allophones that will cover all of the possible speech unit combinations in order to model co-articulation in a wide range of contexts.

A twelfth order pitch synchronous [41] line spectral pair analysis of each sentence in the corpus was performed, using a default frame length of 30ms during unvoiced speech and a default interval of 10ms for unvoiced speech. The interval between frames varies during voiced periods according to pitch.

4.3 Compiling the Phoneme Pair Inventory

Phonemes are those sounds which serve to differentiate words. The term phoneme pair, used in following sections, represents two consecutive phonemes occurring in the speech corpus. The term diphone is used to define the second half of one phone followed by the first half of the next.

The phonemes, for each sentence, were divided to form an inventory of all phoneme pairs occurring in the corpus, listed in descending order of frequency. New data, annotation and parameter files are produced, as shown in Figure 4.1. An inventory of 1472 different diphones was formed and divided into 25 categories. Each digit of the numeric code representing each category corresponds to the broad phonetic type of one of the phonemes forming the diphone, as shown in Table 4.1, for example the diphone t E@ belongs

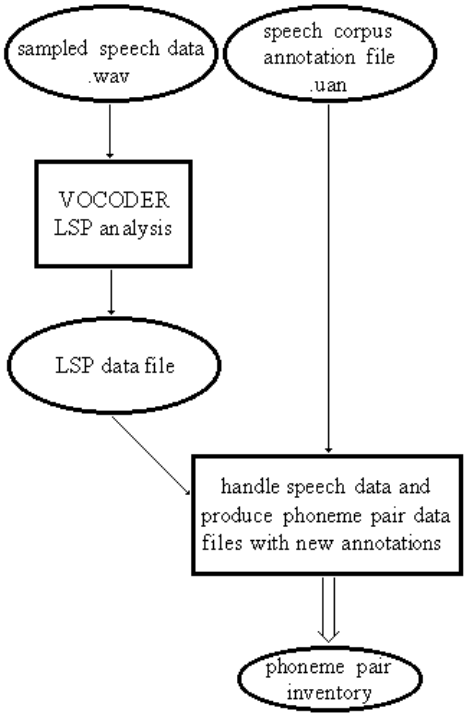

Figure 4.1: Block diagram showing the modules used to produce the LSP parameters and form the phoneme pair inventory.

to category 35 as the initial phoneme is a plosive (group 3) and the final phoneme is a vowel (group 5). Figure 4.2 shows the number of examples for each diphone category. For some diphone categories there are few examples available in the speech corpus (e.g. 11 and 21). This is reflected in the reliability of the results.

4.4 The Realignment Process

As the frame boundaries and phonetic boundaries do not necessarily coincide, the frames of LSP data must be resampled to minimize this registration error. The start and end of the phoneme, in the annotation file, do not necessarily coincide with any frame start and end in the data file. A frame is

Table 4.1: Table of the numeric code assigned to each phonetic category of the set of phonemes represented by the SAMPA alphabet.

Code	Category	Allophones
1	Approximants	=l, l, r, w, j
2	Nasals	m, =m, n, =n, N
3	Plosives	b, p, d, t, g, k
4	Affricates and Fricatives	tS, dZ, v, f, D, T, S, s, z, Z, h
5	Vowels	I, E, {, V, Q, U, @, i, 3, u, A, O, I@, E@, U@, eI, aI, OI, @U, aU

considered to be part of a certain phoneme if it is contained in the interval defined by the annotation file.

The first step in pre-processing the speech data is to calculate the mid points of the phonemes in the phoneme pairs (according to the annotation files), as shown in Figure 4.4. Then the y coordinate of the first point to be resampled (x=0) is interpolated using the coordinates of the point just before the start of the diphone, as shown in Figure 4.3. A similar interpolation process is used to calculate the y coordinate at x=1. Diphone data can then be selected and a diphone annotation file is created as shown in Figure 4.4.

4.5 Diphone Normalization

Since the different examples of a particular diphone available in the inventory do not necessarily have the same length, the duration of each diphone was normalized individually, as shown in Figure 4.4. This allows us to use the curve fitting procedure over all of the examples of a particular diphone. To obtain the new coordinates the start of the diphone is subtracted from the mid point of the frame and this is divided by the length of the diphone. In accordance with the definition of diphone, the start of the diphone is consid-

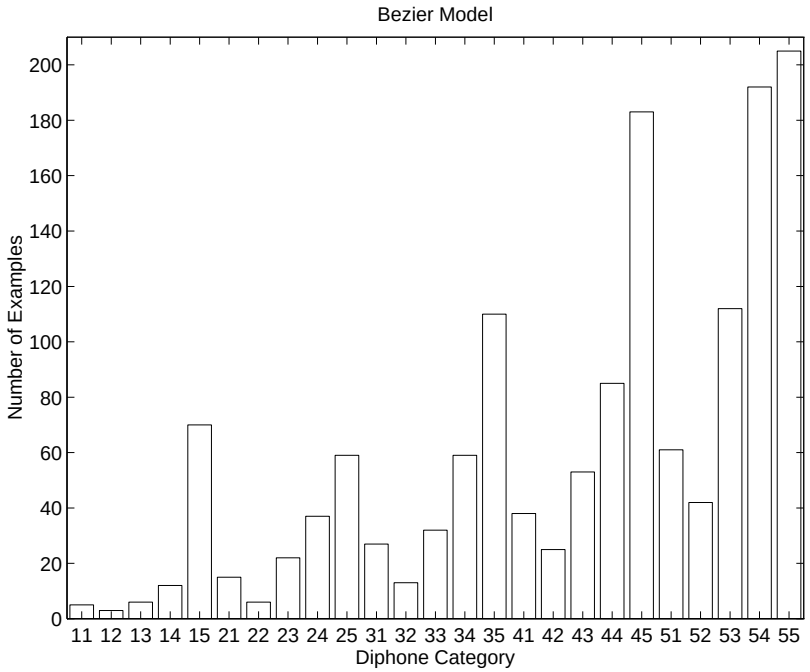

Figure 4.2: Bar chart showing the number of examples for each diphone category (total of 1472 examples).

ered to be the mid point of the first phoneme and the end of the diphone is placed at the mid point of the second phoneme in the corresponding phoneme pair.

After normalization the x-coordinates are contained in the interval defined by the anchor points (B_0 and B_3) located at t=0 and t=1 (see Chapter 3). The actual LSP parameter values (LSP1, ..., LSP12) are also normalized to lie in the range $0 \rightarrow 1$. After this amplitude normalization the coordinate y=0 corresponds to the minimum value amongst all of the diphone examples and the coordinate y=1 represents the maximum value in this set of data.

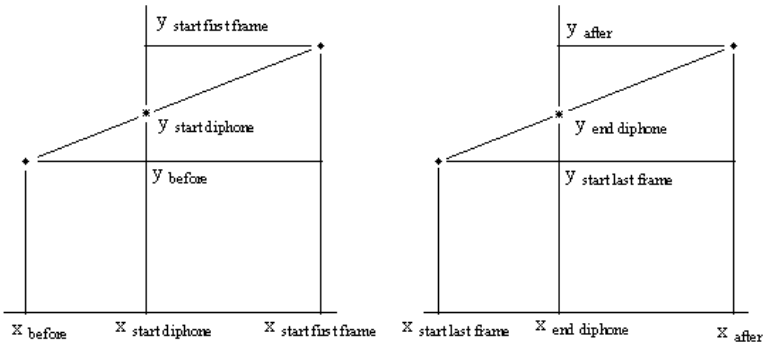

Figure 4.3: Interpolation at x=0 and x=1 (x_{before} and y_{before} are the coordinates of the point just before the start of the diphone).

4.6 Resampling

The frames of line spectral pair parameters were resampled, using simple linear interpolation, as shown in Figure 4.4, so each diphone is represented by ten equally spaced frames of speech parameters. Figure 4.5 illustrates the resampling procedure. The new points (x_{s0},y_{s0}), ..., (x_{s9},y_{s9}) are calculated as follows

$$y_{s1} = \frac{(y_2 - y_1)(x_{s1} - x_1)}{x_2 - x_1} + y_1;$$

...

$$y_{s8} = \frac{(y_{12} - y_{11})(x_{s8} - x_{12})}{x_{12} - x_{11}} + y_{12}.$$

Linear interpolating is used to determine the LSP parameter frames. This has been derived from simple geometrical considerations described in Figure 4.6. Figure 4.7 shows the effect of normalizing and resampling the first LSP parameter of diphone "En" (all examples in diphone inventory).

4.7 Summary

In this chapter, the speech corpus data pre-processing procedures have been presented. A LSP analysis of a phonetically annotated speech corpus is performed, providing data for a newly annotated phoneme pair inventory. The

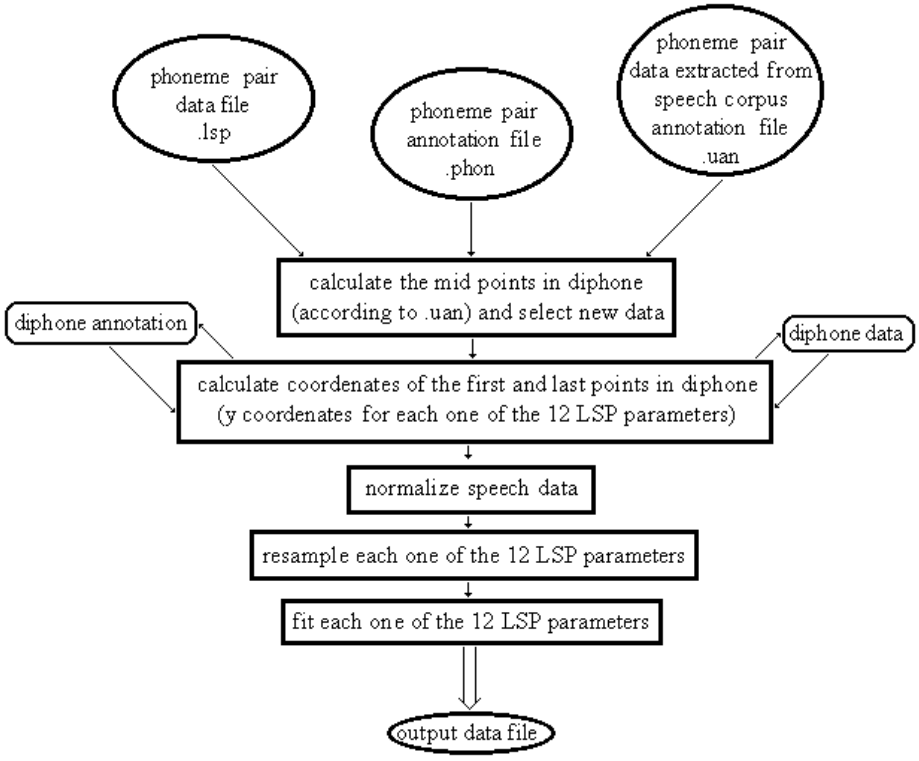

Figure 4.4: Block diagram showing the data pre-processing and fitting modules.

phoneme pair data and annotation files, and the speech corpus annotation files, are used to produce new annotation and data files with the diphones to be fitted using the Bézier model. Alignment problems between the two sets of data are solved by normalizing and resampling the diphone LSP parameters. The pre-processing of speech data proved to be a crucial and meticulous task, essential to all of the parameter fitting methods described in the following chapter.

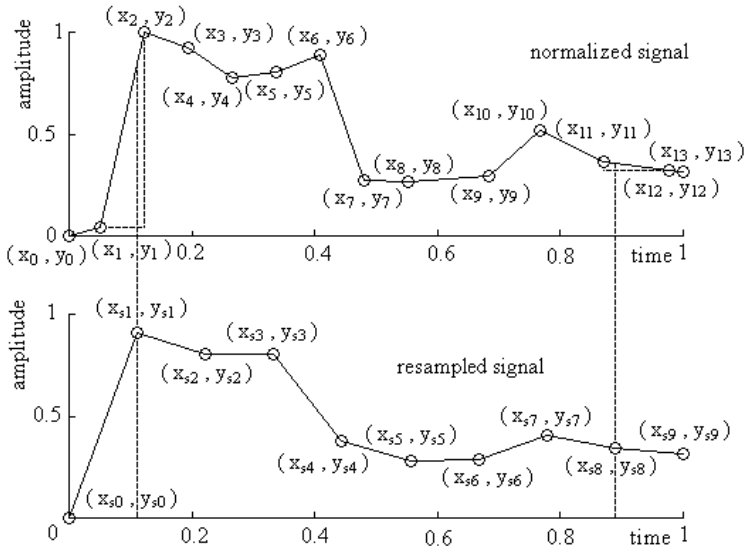

Figure 4.5: Resampling the first LSP parameter in the first example of diphone "p@" (interpolation in the interval $0 < x < 1$).

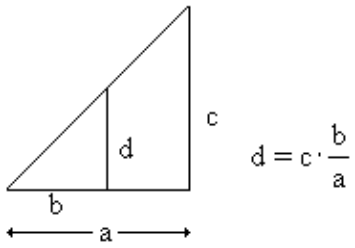

Figure 4.6: Geometrical considerations for resampling.

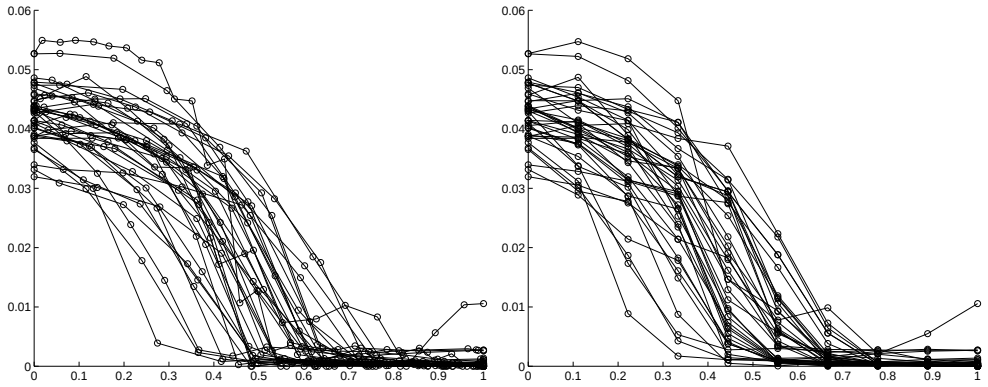

Figure 4.7: Results of normalizing and resampling diphone “En”.

Chapter 5

Method

5.1 Introduction

This chapter discusses the Bézier model of co-articulation and the major design features used in the fitting procedure. A basic strategy for blending adjacent diphones is described. A model based on the Holmes-Mattingly-Shearme scheme is presented for comparison.

5.2 Modelling Diphones

5.2.1 Cubic Bézier Model

The Levenberg-Marquardt algorithm was used to find a single cubic Bézier segment forming the best fit in terms of the RMS error to all available examples of each diphone. The root mean square (RMS) error was recorded for each line spectral pair parameter. This sum of squared errors (deviations of the curve from the data points, as shown in Figure 5.1) can be expressed as

$$RMS = \sqrt{\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(b_j - y_{ij})^2},$$

where N is the number of examples and M is the number of points for each example. b is a vector containing the coordinates of the Bézier curve and y is the vector of data points to model.

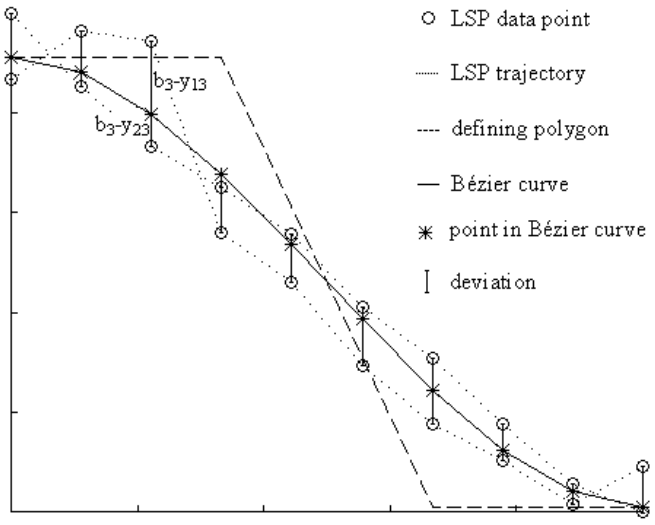

Figure 5.1: Deviations of a Bézier curve from the data points in diphone "ll" (N=2 and M=10). There are two examples of this diphone in the inventory.

In order to reduce the number of parameters controlling the model of each diphone, the constraint that the initial and final gradient of the Bézier segment is zero was imposed, as shown in Figure 5.2, such that

$$y_0 = y_1$$

and

$$y_2 = y_3.$$

This constraint seems reasonable if the diphones are regarded as joining during steady state conditions at the centre of each phoneme. This means that each diphone is represented by only four parameters y_0, x_1, x_2 and y_3 for each LSP parameter providing some amount of data compression.

5.2.2 The Holmes-Mattingly-Shearme Based Method

The accuracy of the Bézier models was compared with an existing template, similar to that used for interpolation of parameter transitions in the Holmes-

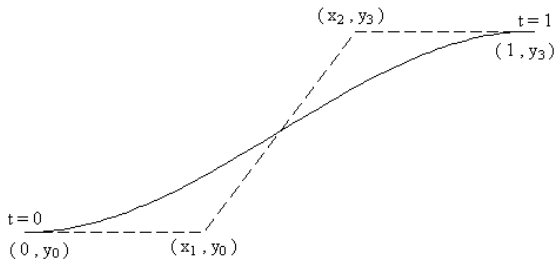

Figure 5.2: Constrained cubic Bézier segment used to model transitions in line spectral pair parameters between adjacent phonemes.

Mattingly-Shearme scheme.

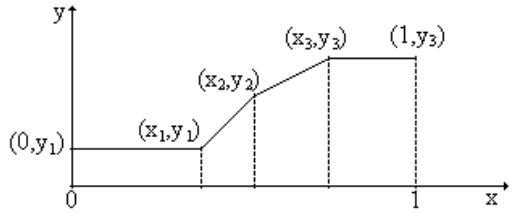

Figure 5.3: Holmes-Mattingly-Shearme like scheme template.

A simple template was used to fit the LSP parameter transitions of all diphones in the inventory using the Levenberg-Marquardt algorithm to minimize the the resulting RMS error. Figure 5.3 illustrates the constraints imposed to the segment boundary points. The template is a set of 4 straight line segments.

5.3 Blending Adjacent Diphones

Each word in the sentence "look out of the window and see if it's raining" ("lUk aUt @v D@ wInd@U =n si If Its reInIN") is fitted diphone by diphone using cubic Bézier segments with a number of curve points equal to the number of data points.

The curve points are calculated assuming an initial value to a four point defining polygon and adjusting new values to minimize the distance (error) between speech data and calculated Bézier curve points. We find the least squares minimum to the function that returns the Euclidean distance

$$d_{Euclid}(x, n) = \| x - n \| .$$

The joint of two adjacent Bézier curves is placed at the mid-points of the phonemes to ensure maximum stability.

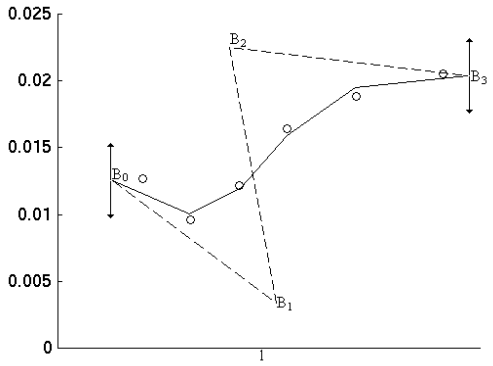

Figure 5.4: Fitting data in the word "look (lUk)": from start of phoneme "l" to mid-point of phoneme "l".

Figure 5.4, Figure 5.5 and Figure 5.6 illustrate the three different constraints imposed to the data points when fitting the word "lUk". The silence $\rightarrow$ "l" transition is not shown. The data points (B_0 and B_3) of the first Bézier curve have fixed x coordinates: the start of the first phoneme and the mid-point of the first phoneme (annotation file values), as shown in Figure 5.4. The curve is fitted to the data that lies in this interval (non-linear constrained minimization). The next Bézier curve ends at the mid-point of the following phoneme, as shown in Figure 5.5. The last Bézier curve is applied between the mid-point of the last phoneme and the end of the last phoneme, as shown in Figure 5.6.

When there is a reduced number of frames in the phoneme, that does not justify subdivision at mid-point, we use specific criteria. The number of Bézier

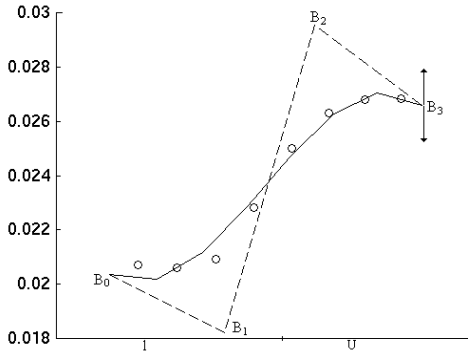

Figure 5.5: Fitting data in the word "look (lUk)": from mid-point of phoneme "l" to mid-point of phoneme "U".

segments is reduced accordingly.

In order to produce a smooth transition between Bézier segments, we need only ensure that the end point of each segment coincides with the start point of the next and that the necessary first-derivative continuity conditions apply:

$$B_3 = B_{0(next)},$$

$$B_{1(next)} = 2B_{0(next)} - B_0.$$

5.4 The Sigmoidal Logistic Curve

Since the LSP parameters abut near the start and end of the diphones, a sigmoidal logistic function improves the accuracy of the model. This could be an alternative parametric curve used to form models of co-articulation in human speech. The sigmoidal logistic function, shown in Figure 5.7, can be defined as

$$g(t) = \frac{c}{1 + e^{-at+b}} + d,$$

where a scales time (used to obtain a steeper transition), b controls the horizontal translation (time), c controls the overall hight and d controls the vertical displacement.

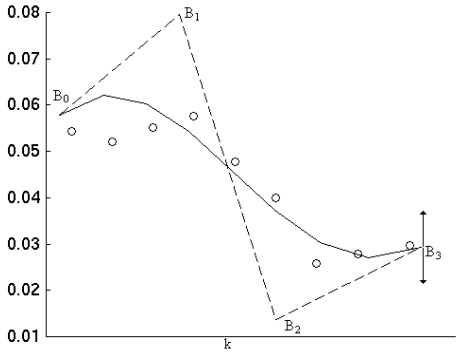

Figure 5.6: Fitting data in the word "look (lUk)": from mid-point of phoneme "k" to end of phoneme "k".

5.5 Summary

This chapter presents a detailed method for modelling diphones using cubic Bézier segments and a strategy for blending adjacent diphones. Also presented are an alternative interpolation template and a Holmes-Mattingly-Shearme like scheme for comparison with the Bézier model. These methods were used to model phonemes, diphones and a sentence, of which results are discussed in the next chapter.

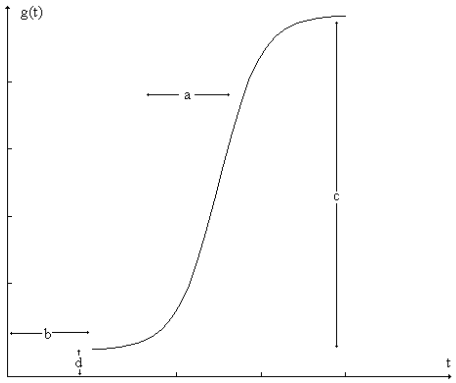

Figure 5.7: Sigmoidal logistic curve.

Chapter 6

Results

6.1 Introduction

This chapter presents the results obtained using both Bézier and Holmes-Mattingly-Shearme models of co-articulation. A discussion of the results of modelling diphones using cubic Bézier segments is included. The accuracy of modelling individual phonemes is also described. The chapter presents results of blending adjacent diphones and discusses first attempts to synthesize an example sentence using this strategy.

6.2 Modelling Phonemes

6.2.1 Bézier Cubic Model

Figure 6.1 shows the average root-mean-square error for each phoneme (all examples in the speech corpus). When abrupt transitions are required it can be seen that the average RMS error of the model is relatively high (e.g. phonemes "D", "dZ" and "tS") when compared with smooth vowel transitions (e.g. phonemes "i", "3" and "u").

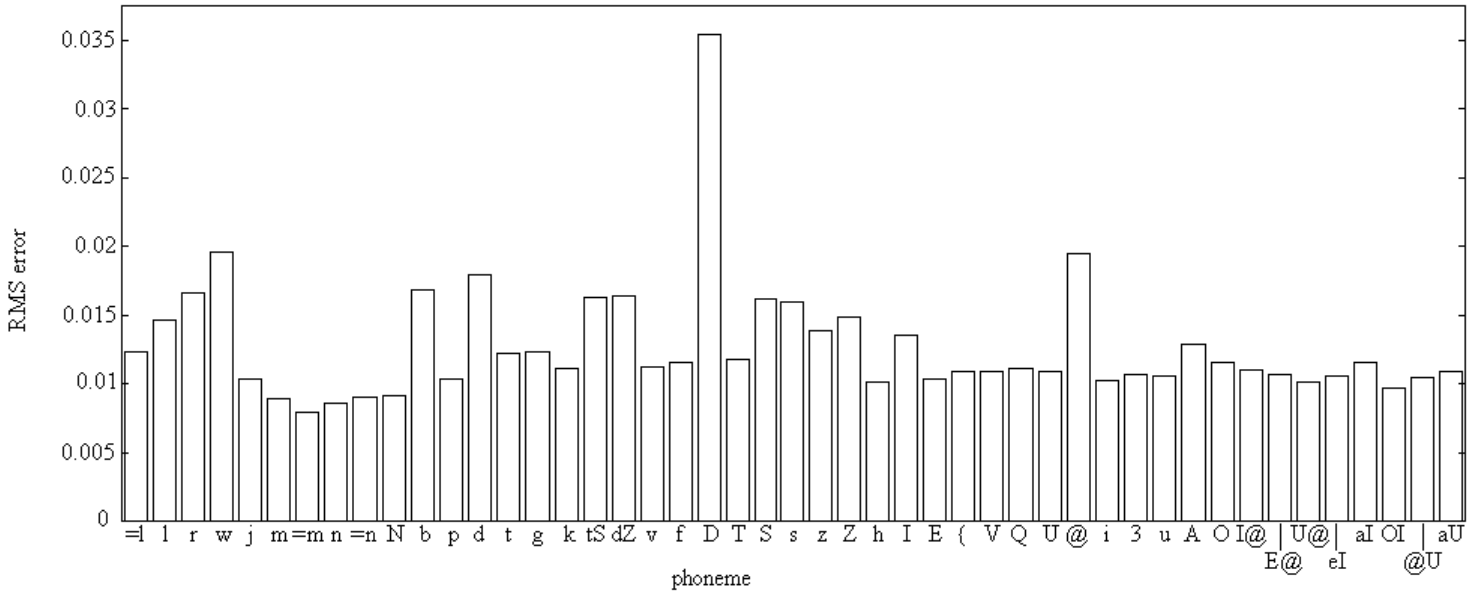

Figure 6.1: Bar chart showing the root mean square error for the Bézier models of each phoneme (all examples in speech corpus).

The phoneme LSP parameters are recorded in 47 data files and their limits registered in new annotation files. These files contain all of the examples available in the speech corpus. The data was normalized and resampled using similar procedures to those described for diphones. The Levenberg-Marquardt algorithm was used to find a single cubic Bézier segment forming the best fit to all available examples of each phoneme and the root mean square (RMS) error was recorded for each line spectral pair parameter. This experiment allows us to have a clear idea of what sort of accuracy to be expected when modelling individual categories of phonemes.

6.3 Modelling Diphones

6.3.1 Bézier Cubic Model

The diphone "j3" and the result of encoding it using the Bézier model of co-articulation are represented in Figure 6.2.

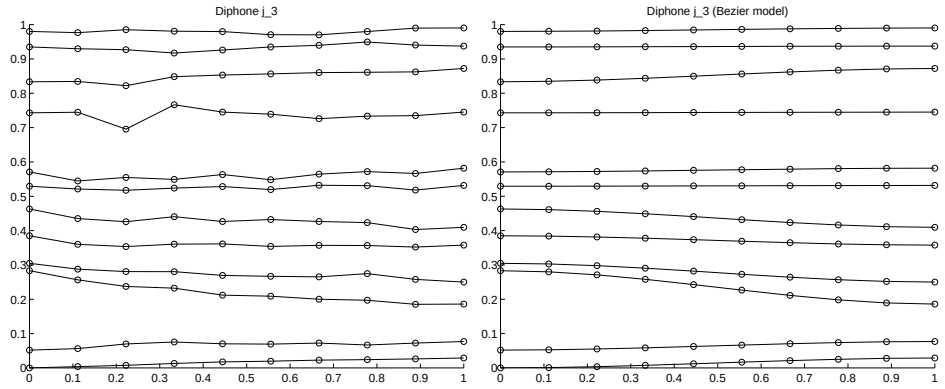

Figure 6.2: The LSP parameters of diphone "j3" and the Bézier model.

Figure 6.3 shows the average root-mean-square error for each of 25 diphone categories.

The accuracy of the Bézier model for some diphones is better than others. Close examination of the data has highlighted the following trends:

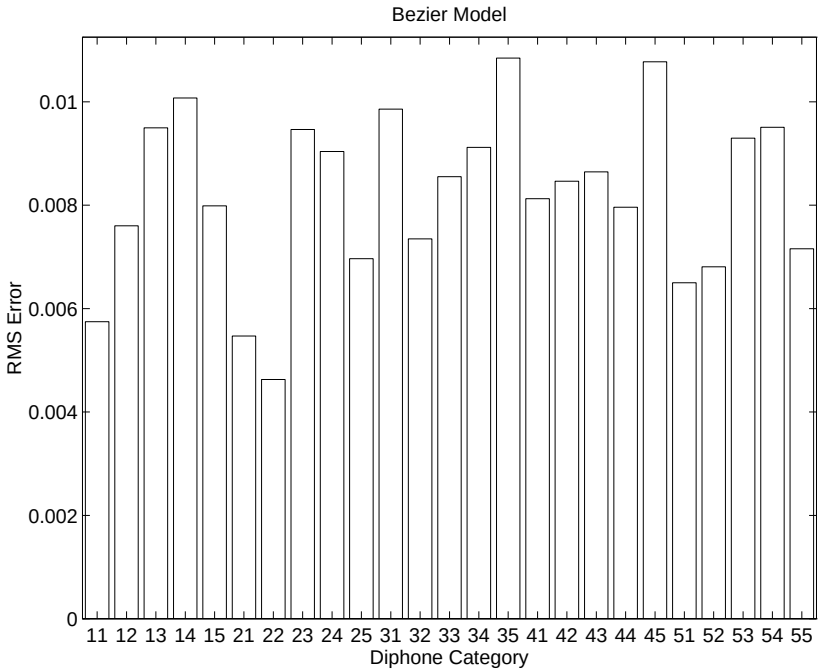

Figure 6.3: Bar chart showing the root mean square error for the Bézier models of each diphone category (all examples in diphone inventory).

Abrupt transitions

- Nasal-plosive (category 23) and nasal-fricative (category 24) diphones are both characterized by a nasal resonance followed by a change to oral emission. The resulting models present a large RMS error due to high variance in plosives and fricatives.

- Plosive-approximant (category 31) diphones are affected by the phonetic context. When a plosive follows an approximant (category 13) the first phoneme terminates while its formants are being moved toward the appropriate loci for the plosive. Plosives have complex parameter transitions and so the RMS error in these categories is very high.

- When affricates or fricatives follow vowels (category 54) the transitions involve strong characteristics of plosive-like closures after which the affricate signal is heard. The RMS error is high.

- In category 43 (affricates or fricatives followed by plosives) the second sound is formed while still producing the first. The plosive is such a precise articulatory gesture that the articulators have to be in a specific position in order to produce the utterance.

High variance

- In fricative-approximant clusters (category 41) the approximant is less affected by the preceding consonant than it is in plosive-approximant (category 31) diphones. The average RMS error is lower in category 41 than in category 31.

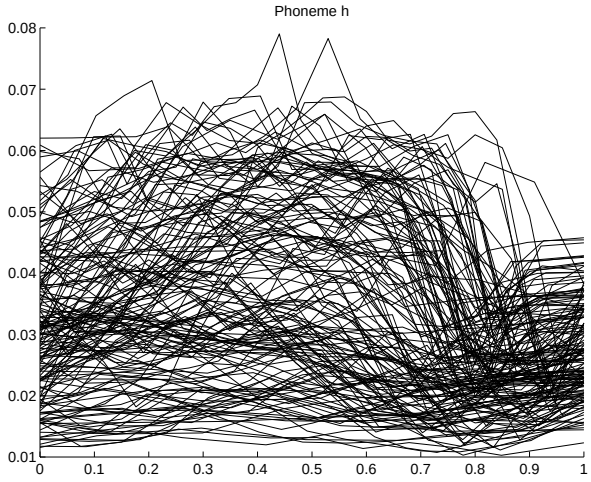

Figure 6.4: Displaying the first LSP parameter of phoneme "h" (all examples in speech corpus).

- Affricates or fricatives followed by vowels (category 45) involve a transitional signal unique to each vowel context. These sounds have a naturally high variance due to being highly co-articulated (e.g. voiced "h" also known as "h" vowel, as shown in Figure 6.4). This affects the accuracy of the model.

- When two consonants occur contiguously, each belonging to different classes but having the same place of articulation, the latter consonant takes vowel characteristics (e.g. "pri<u>s</u><u>o</u><u>n</u>" and "<u>c</u><u>o</u><u>n</u>sult"). This might be

another source of variability in classes that involve consonant-consonant clusters.

Vowels

- Vowel-vowel diphones (category 55) consist of a smooth transition from one vowel to another with perceptually important duration differences between contiguous allophones. The RMS error is relatively low.

6.3.2 Holmes-Mattingly-Shearme Like Model

In this section the accuracy of the cubic Bézier model is compared with that of a straight line template model.

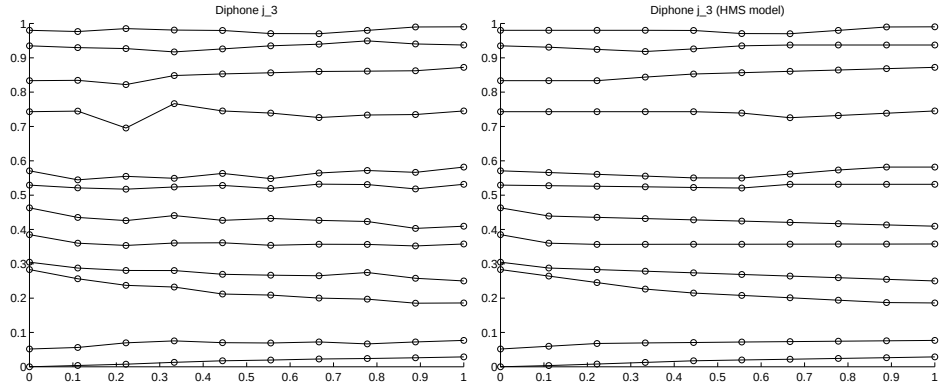

Figure 6.5: The LSP parameters of diphone "j3" and the Holmes-Mattingly-Shearme like model.

The diphone "j3" and the result of encoding it using this technique are represented in Figure 6.5. A straight line template, as described in Section 5.2.2, was used to fit each one of the 12 LSP parameters. It can be seen that the basic LSP parameter trajectory is respected but it is still difficult to model abrupt transitions.

Figure 6.6 shows the average root-mean-square error for each of 25 diphone categories. The diphone category discussion presented for the Bézier model

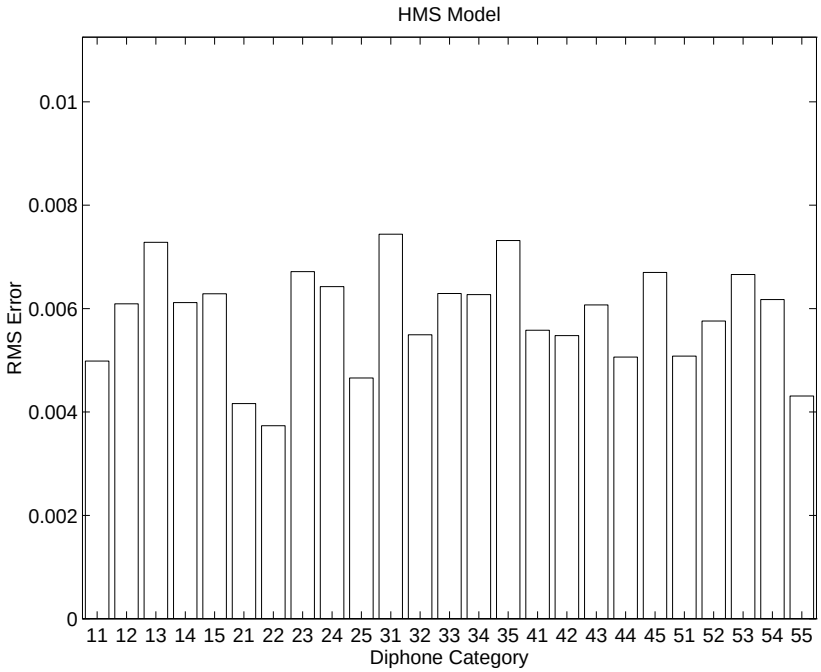

Figure 6.6: Bar chart showing the root mean square error for the Holmes-Mattingly-Shearme models of each diphone category (all examples in diphone inventory).

can still be maintained for these results.

Figure 6.7 shows the Bézier model and the Holmes-Mattingly-Shearme like model of the LSP parameters of diphone "D@". They present remarkable similarities.

In Figure 6.8 the accuracy of the two models is compared by analysis of the RMS error for each diphone category in the speech corpus. A most important result here is that the straight line template approach is almost uniformly better than the Bézier model. We need next to determine if this is just a matter of model order. The LSP parameter trajectories proved to be of a complex nature and so it was not possible to fit all transitions adequately using a Bézier model with only one segment.

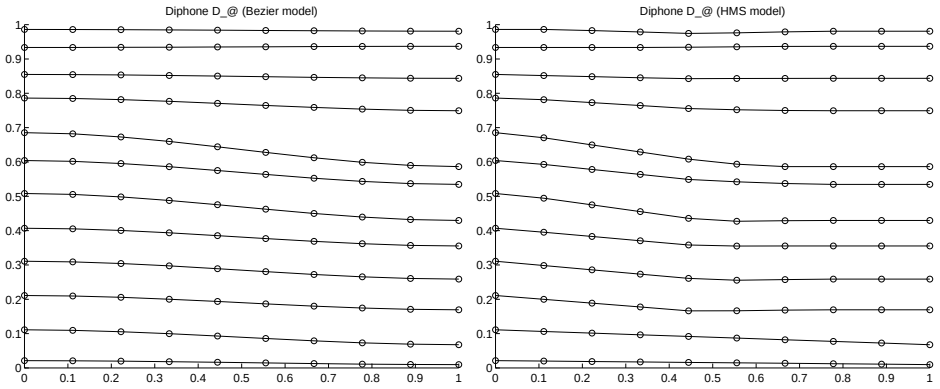

Figure 6.7: The Bézier model and the Holmes-Mattingly-Shearme model of diphone "D@".

6.4 Blending Adjacent Diphones

Since diphones abut during the steady state conditions in the middle of each phoneme, it seems sensible to blend the Bézier models of each diphone so as to ensure a smooth transition between each diphone, as described in the methodology chapter. Figure 6.9 shows the trajectory of the first line spectral parameter for the word "look" ("lUk" according to the SAMPA phonetic alphabet) generated from Bézier curve segments using this approach. The defining polygons are also shown. It appears that a single Bézier segment is able to model the transition between sonorants adequately; however, where abrupt transitions are required, principally during plosive sounds, a more complex model is needed.

Plosive sounds are modelled in many synthesis-by-rule systems as a sequence of smaller speech sounds representing the closure, release and post-release phases so that the existing interpolation method can accommodate a more complex trajectory. This would also seem to be a sensible approach to adopt in this work.

It is possible to locate cues that define the closure and release phases in plosives, as shown in Figure 6.10. Large differences between adjacent LSP parameter values and maximum values in the energy function denote a new

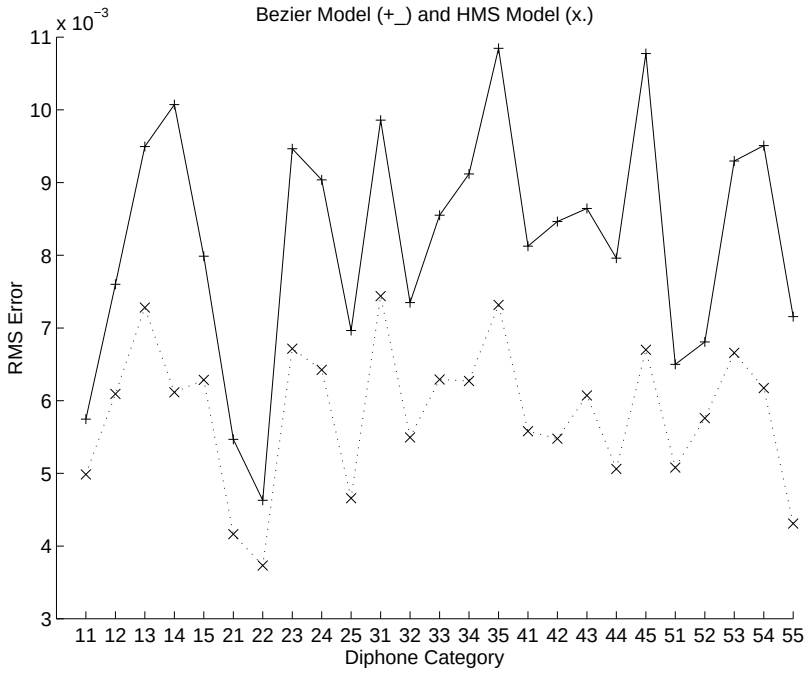

Figure 6.8: The root mean square error for the Bézier models and the Holmes-Mattingly-Shearme like models of each diphone category (all examples in diphone inventory).

phase.

6.5 Synthesis

First attempts to synthesize each word in the sentence "look out of the window and see if it's raining" ("lUk aUt @v D@ wInd@U =n si If Its reInIN") produced unsatisfactory results. The utterances "look", "see" and "it's" sounded barely intelligible. The words "window" and "raining" caused LSP speech synthesis filter instability.

The LSP parameters *start frame* and *parameter 1* to *parameter 12* were fitted to cubic Bézier segments and new data files were produced restoring the analysed speech data residuals *frame length*, *energy* and *gain* (see Appendix

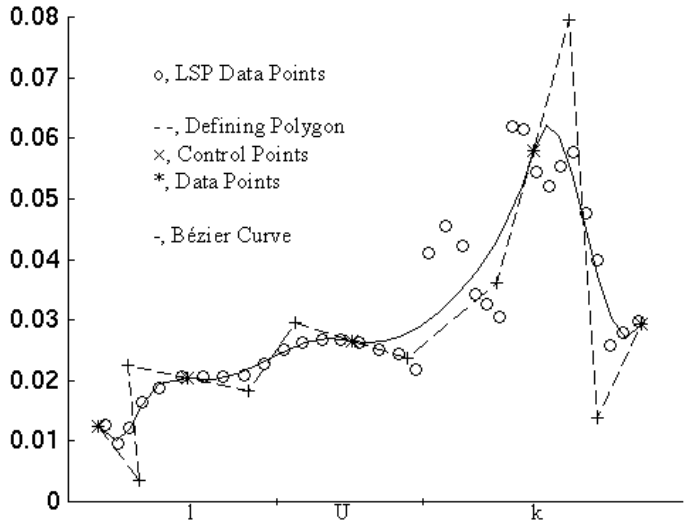

Figure 6.9: Trajectory of the first line spectral pair parameter for the word "look" modelled by Bézier segments, also showing the defining polygons.

B). It was possible to obtain a considerable amount of data compression as shown in Table 6.1. Figure 6.11 shows the average RMS error for each word.

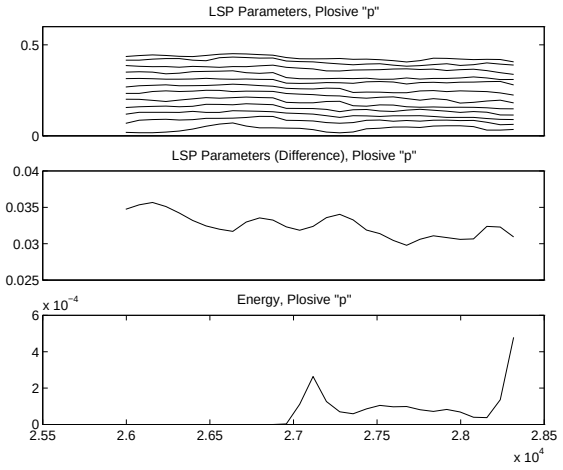

Figure 6.10: Displaying LSP parameters, average difference between adjacent LSP parameters and energy in plosive "p".

Table 6.1: Number of speech parameters (data points) and number of parameters controlling the model (polygon points).

Word (SAMPA)	Data points	Polygon points
look (lUk)	35	16
out (aUt)	36	12
of (@v)	23	12
the (D@)	11	8
window (wInd@U)	47	16
and (=n)	6	4
see (si)	34	8
if (If)	27	12
it's (Its)	25	8
raining (reInIN)	85	24

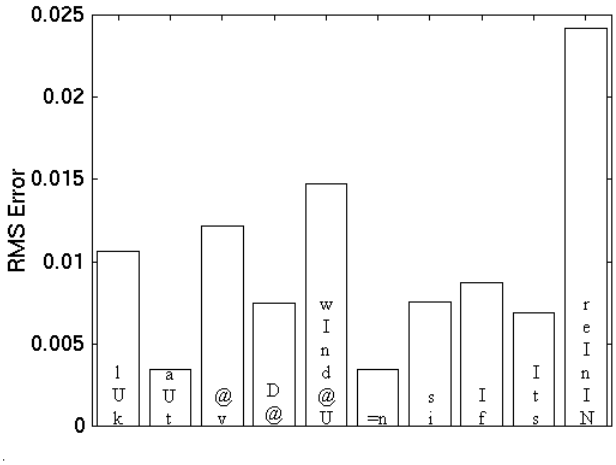

Figure 6.11: Bar chart showing average RMS error for each word in the sentence "look out of the window and see if it's raining" ("lUk aUt @v D@ wInd@U =n si If Its reInIN").

Chapter 7

Conclusions

This research has been directed at investigating the use of Bézier curves to form models of co-articulation in human speech. The objective of this work was to find the numerical accuracy of the Bézier model. A 12^{th} order, pitch synchronous line spectral pair (LSP) analysis is performed on a corpus of 239 phonetically balanced sentences of English speech. The data is used to form an inventory of diphones and the trajectory of their LSP parameters is modelled by a single cubic Bézier curve segment found using the Levenberg-Marquardt fitting method. It appears that the Bézier curve parameters form a good model of co-articulation and that they can be easily merged to produce smooth transitions between synthesis units. Speech coding using the Bézier curve defining polygon points as parameters provides a considerable amount of data compression.

The principle results obtained indicate the following:

- The parameters for unvoiced sounds exhibit considerably more variation than those of voiced sounds. The variability in unvoiced speech parameters based on linear predictive analysis of a fixed order, is largely due to the analysis of an almost random signal, through a window of limited duration. However the exact spectral envelope of unvoiced speech sounds are unimportant, provided that the coarse spectral shape is correct and there are no well defined spectral features.
- The most rapid transitions in speech parameters occur during plosive and nasal sounds. The resulting model is naturally less accurate.

- A single cubic Bézier curve can be used as an effective model of the trajectory for line spectral pair parameters during the transitions between a majority of speech sounds.

- Where abrupt transitions are required, principally during plosive sounds, a more complex model is needed. A sensible approach to adopt would be to model plosives as a sequence of smaller speech sounds representing the closure, release and post-release phase so that a more complex trajectory is modelled by three separate Bézier segments.

- The straight line template approach performs appreciably better than the Bézier model. It is thought that this is a matter of model order (the Bézier model has 4 degrees of freedom and the Holmes-Mattingly-Shearme based model has 6 degrees of freedom). The LSP parameter trajectories proved to be of a complex nature not possible to model with only one segment.

The research work presented in this thesis provides a detailed analysis of speech parameter transitions. The phoneme and diphone inventories, together with the data pre-processing techniques presented in Chapter 4, can be used for future research involving the same speech corpus. The accuracy of the fitting procedure provides useful information on the nature of the transitions. The Bézier model of co-articulation is a flexible parametric representation of speech with possible applications ranging from speech coding to speech synthesis.

7.1 Further Work

Further ideas to improve and extend the work presented in this thesis include the following:

- Further research is needed to determine suitable strategies for blending Bézier segments to produce a natural sounding utterance. The existing strategy produces a smooth transition between Bézier segments by ensuring that the end point of each segment coincides with the start point of the next and that the necessary first derivative continuity conditions apply.

- Other parametric curves could be used to model parameter transitions, providing alternative solutions for some classes of speech sounds.
- Further experimentation with the Bézier model order should be encouraged, in order to find the best compromise between data compression and speech quality.
- An effective resynthesis procedure should be implemented in the future. The restored LSP parameters and analysed speech data residuals could be used by a VOCODER.

This work has provided a deeper insight into speech data and a set of paths for further developments.

Appendix A

Displaying Speech Data

A.1 Introduction

This appendix documents the use, during a preliminary phase of research, of a set of tools developed to visualize sentences, phonemes and diphones. By this we intended to identify the variability amongst phonemes in the speech corpus and characterize the diphone transitions. The results disclosed relevant cues when conceiving the experimental procedure.

Figure A.1 shows the LSP parameters of the first sentence in the speech corpus. This was the first set of data used to test the curve fitting technique.

A.2 Gaussian Kernel Interpolation

A Gaussian kernel interpolation method was developed in order to identify the regions in space where each one of the 12 LSP parameters are located.

Speech data is normalized and a grid dividing the LSP parameter space distribution into equally sized squares is generated as shown in Figure A.2.

A matrix of n=128 equally spaced points is created

$$grid(i) = \begin{bmatrix} 0 & \frac{1}{n-1} & \ldots & \frac{n-2}{n-1} & 1 \end{bmatrix}.$$

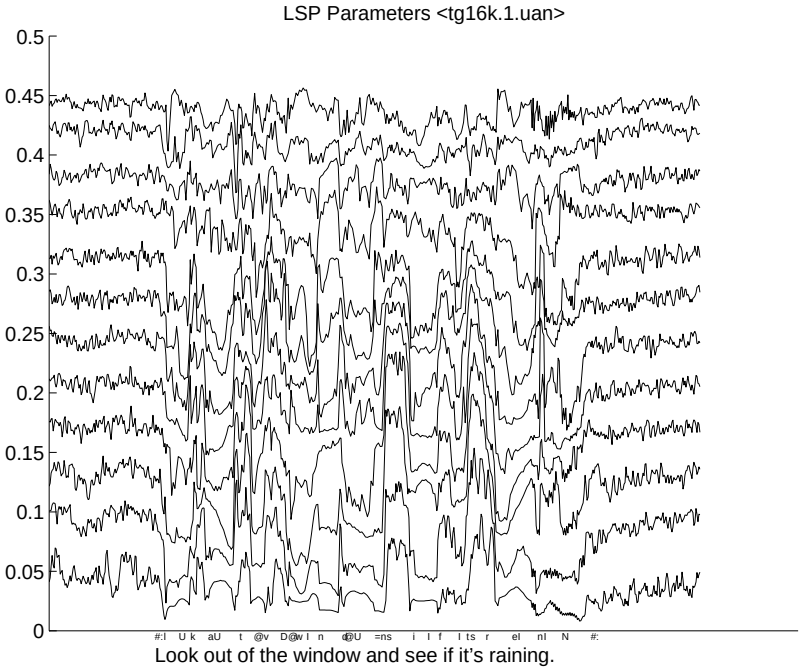

Figure A.1: Displaying LSP parameters of sentence "look out of the window and see if it's raining" ("lUk aUt @v D@ wInd@U =n si If Its reInIN").

In order to determine the number of times the LSP parameter trajectories pass through a certain region the following algorithm was developed

```
pattern=zeros(n-1,n-1)

for i=1:number of rows in xdata

  for j=2:n
    if grid(j-1) ≤ xdata(i) ≤ grid(j)
      patx=j-1
      exit for
    end
  end

  for j=2:n
    if grid(j-1) ≤ ydata(i) ≤ grid(j)
      paty=j-1
      exit for
    end
  end

  pattern(paty,patx)=pattern(paty,patx)+1


end
```

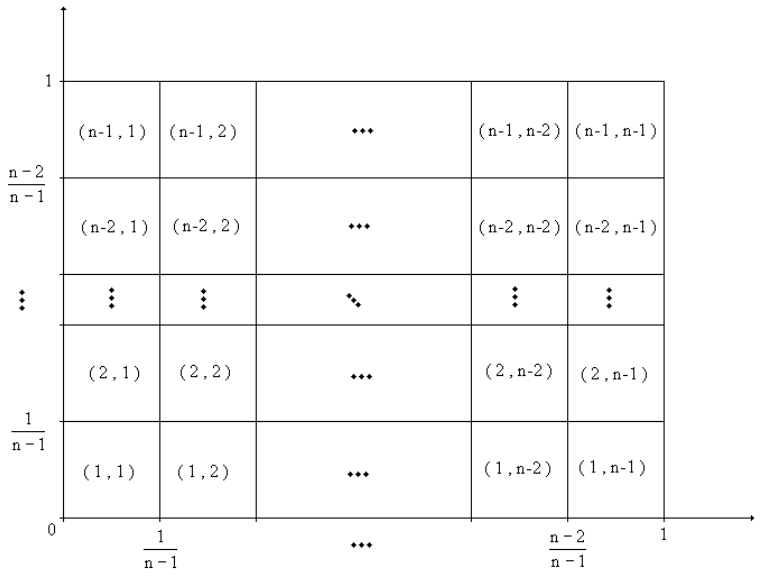

Figure A.2: Interpolation grid.

The resulting vector *pattern* is filtered using a rotational symmetric $n \times n$ Gaussian lowpass filter with main lobe width of 4 pixels [58]. The image data is then displayed as shown in Figure A.5 and Figure A.7.

A.3 Displaying Phonemes

The phoneme annotation files include information identifying the sentence from which it has been extracted and the original starting and ending frames. This is shown when displaying the individual examples as in Figure A.3.

It can be clearly seen in Figure A.4 and Figure A.5 that phoneme "3" presents high variability. The data has been normalized and all the examples are shown.

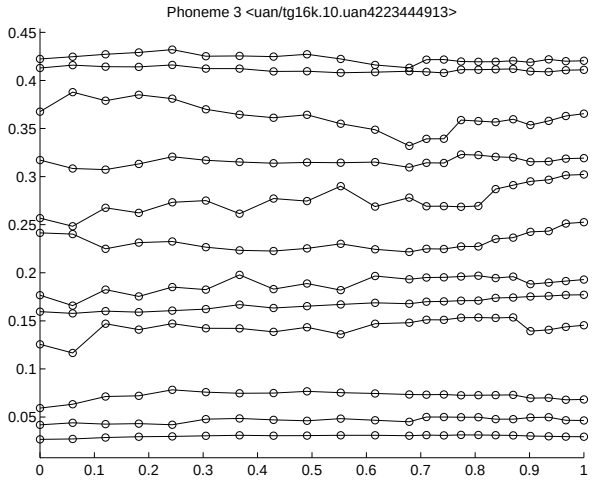

Figure A.3: Displaying LSP parameters of phoneme "3" occurring in the sentence "thank you for giving me the power to tour the world" ("T{Nk ju f@ gIvIN mi D@ paU@ t@ tO D@ w3ld")

A.4 Displaying Diphones

Figure A.6 shows all the examples of diphone "En". The LSP parameters are displayed after normalization and resampling. This is the result of processing the data before the actual curve fitting method is applied. It can be seen in Figure A.7 that the 12 LSP parameters are dispersed, whereas the parameters in Figure A.5 concentrate in relatively distinct regions of space.

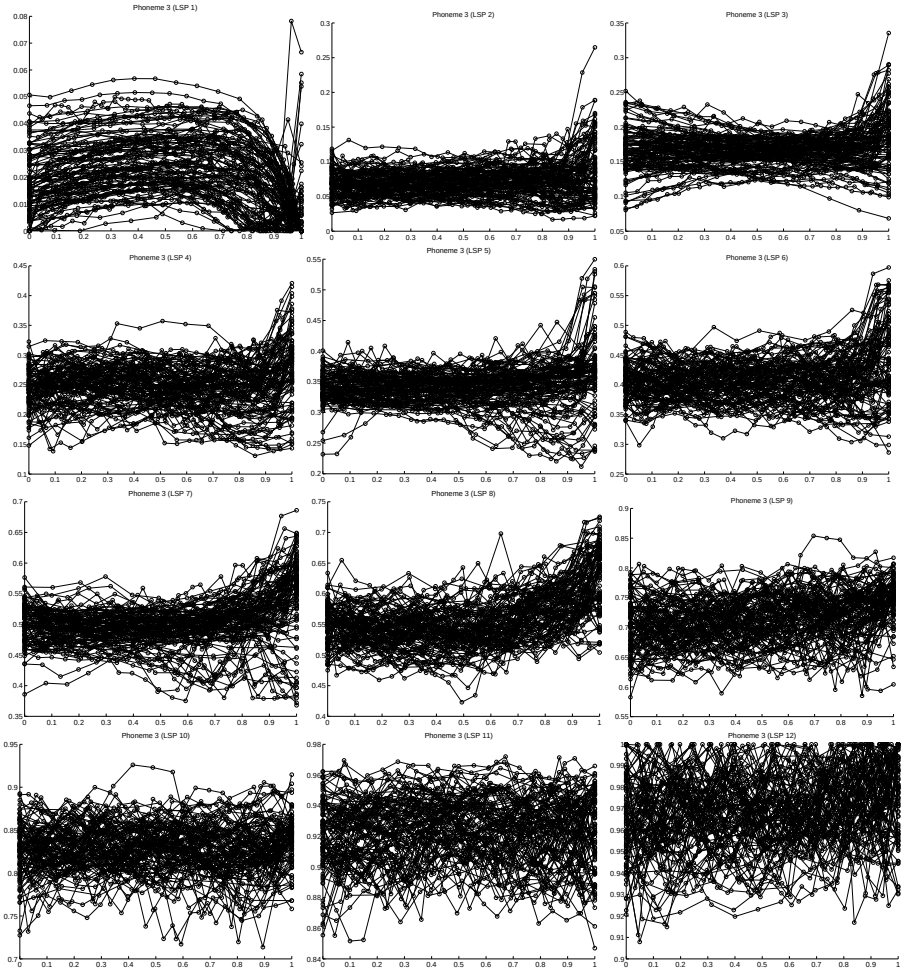

Figure A.4: Displaying LSP parameters of phoneme "3" (all examples in speech corpus).

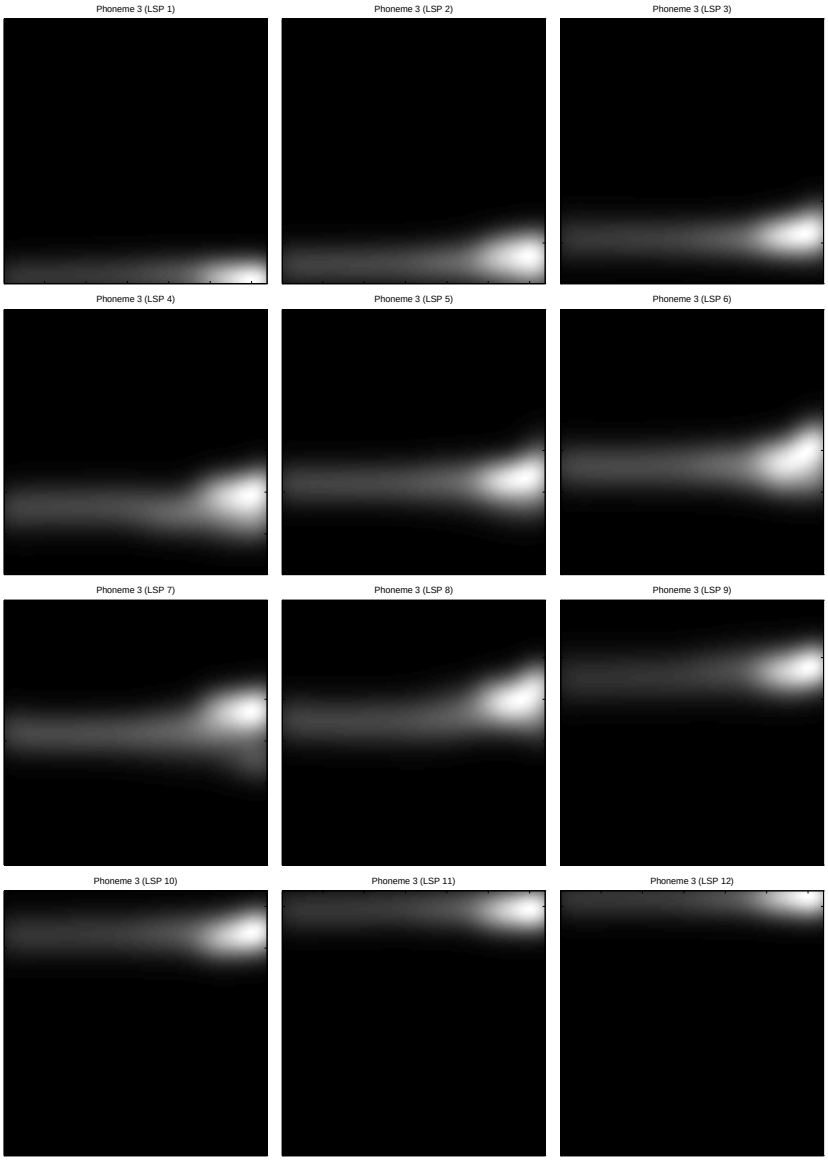

Figure A.5: Displaying Gaussian kernel interpolation results for LSP parameters of phoneme "3" (all examples in speech corpus).

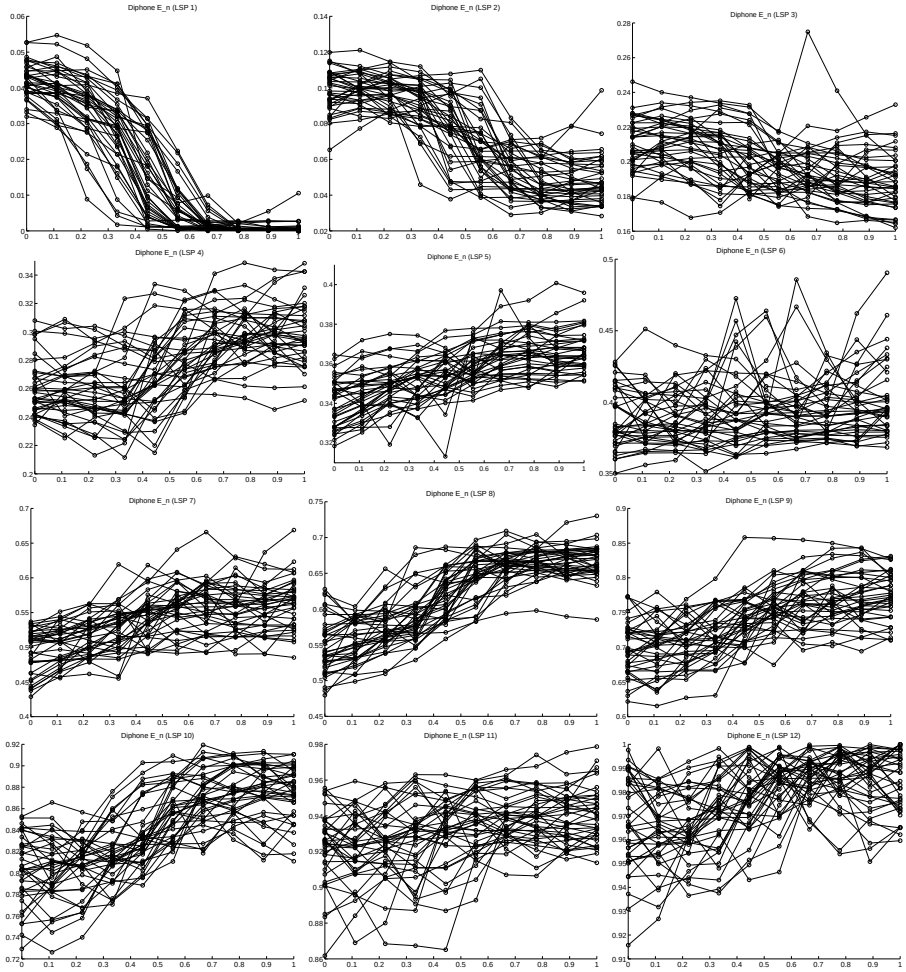

Figure A.6: Displaying LSP parameters of diphone "En" (all examples in speech corpus).

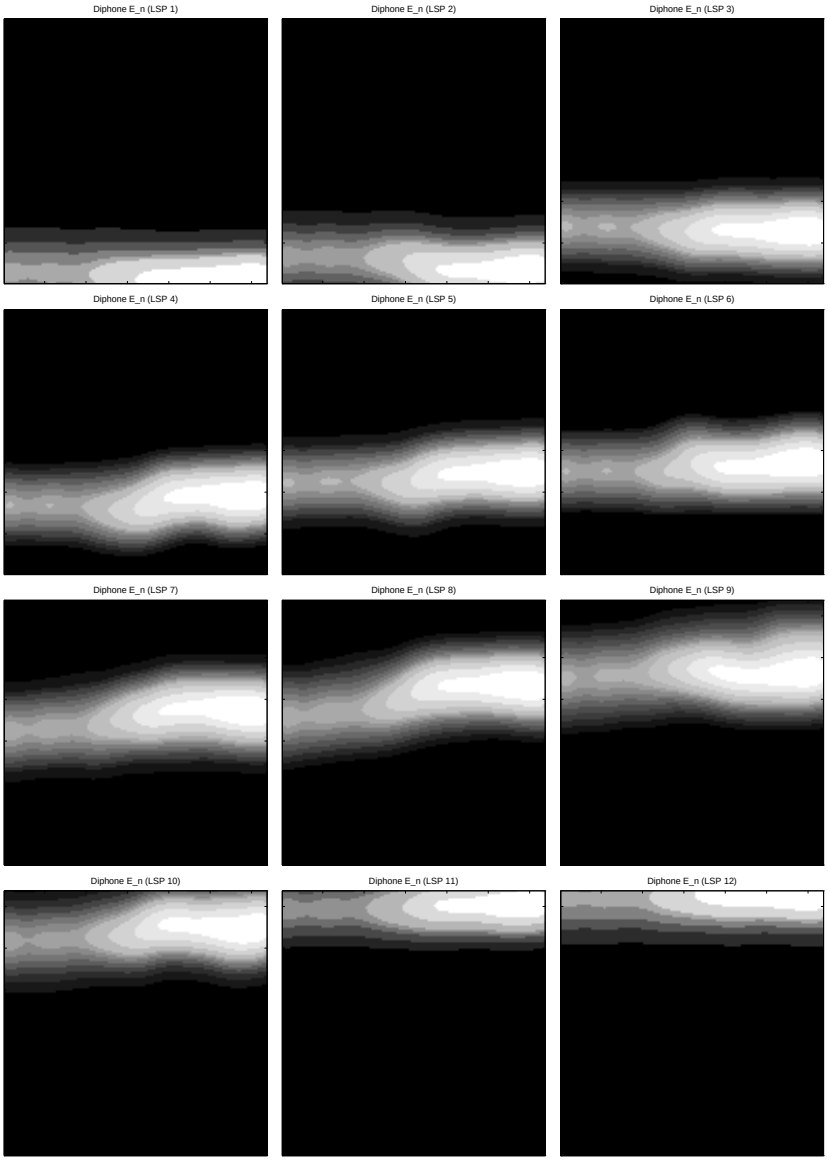

Figure A.7: Displaying Gaussian kernel interpolation results for LSP parameters of diphone "En" (all examples in speech corpus).

Appendix B

File Formats and Phonetic Notation

B.1 Introduction

This appendix presents detailed information about the formats of files in the speech corpus and produced by the LSP analysis. The speech file format is 16 bit linear unsigned samples at a rate of 16KHz.

B.2 Speech Corpus Annotation Files

Table B.1: Annotation file format.

Phonetic symbol	Stress tone and syllable markers	Start phoneme	End phoneme
...	...	...	...

The database includes annotation files with the format shown in Table B.1. The time aligned phonetic transcription is based on the SAMPA phonetic alphabet described in Tables B.2 to B.8 and the modifiers listed in Table B.9. The order of diacritic information is not relevant so $_ is equivalent to _$.

Table B.2: SAMPA and IPA notation for the approximants of English speech.

SAMPA	IPA	Example	Description
=l	**l̩**	bottle	syllabic voiced alveolar lateral liquid
l	**l**	leer	voiced alveolar lateral liquid
r	**r**	rear	voiced unrounded palato-alveolar liquid
w	**w**	wear	voiced rounded labio-velar glide
j	**j**	year	voiced palatal central glide

Table B.3: SAMPA and IPA notation for the nasals of English speech.

SAMPA	IPA	Example	Description
m	**m**	men	voiced bilabial
=m	**m̩**	prism	syllabic voiced bilabial
n	**n**	near	voiced alveolar
=n	**n̩**	button	syllabic voiced alveolar
N	**ŋ**	wing	voiced velar

Table B.10 illustrates a typical .uan annotation file. The symbol #: is the start and end marker.

B.3 LSP Data Files

The LSP data files have the format depicted in Table B.11. Default values for frame length and interval are used during unvoiced periods in pitch synchronous analysis and in all pitch asynchronous analysis. The interval between frames varies during voiced periods.

The first lines in the data file of the sentence "look out of the window and see if it's raining" ("lUk aUt @v D@ wInd@U =n si If Its reInIN") have the format depicted in Table B.12.

It is clearly illustrated in Table B.13 that neither start and nor end of the phoneme marked in the annotation file coincide with the start of a frame.

Table B.4: SAMPA and IPA notation for the plosives of English speech.

SAMPA	IPA	Example	Description
b	**b**	bear	voiced bilabial
p	**p**	pear	voiceless bilabial
d	**d**	dear	voiced alveolar
t	**t**	tear	voiceless alveolar
g	**g**	gear	voiced velar
k	**k**	king	voiceless velar

Table B.5: SAMPA and IPA notation for the affricates of English speech.

SAMPA	IPA	Example	Description
tS	**ʧ**	cheer	voiceless palato-alveolar
dZ	**ʤ**	jeer	voiced palato-alveolar

A fixed frame length is used during the unvoiced sound "k" and there is an unequivocal gain increase. The interval between frames varies while there's voicing, i.e., for the duration of vowel "U".

Table B.6: SAMPA and IPA notation for the fricatives of English speech.

SAMPA	IPA	Example	Description
f	**f**	fear	voiceless labio-dental
v	**v**	very	voiced labio-dental
D	**θ**	this	voiceless dental
T	**ð**	thing	voiced dental
s	**s**	sing	voiceless alveolar
z	**z**	zing	voiced alveolar
S	**ʃ**	sheer	voiceless palato-alveolar
Z	**ʒ**	treasure	voiced palato-alveolar
h	**h**	hear	voiceless glottal

Table B.7: SAMPA and IPA notation for the monophthongs of English speech.

SAMPA	IPA	Example	Description
i	**i**	bead	front close unrounded (cardinal 1)
I	**I**	bit	front close unrounded (between cardinal 1 and cardinal 2)
@	**ə**	ago	central mid unrounded (schwa)
E	**ɛ**	bet	front open-mid unrounded (cardinal 3)
3	**ɜ**	bird	central open-mid unrounded
{	**æ**	bat	front open unrounded (between cardinal 3 and cardinal 4)
A	**ɑ**	bard	back open unrounded (cardinal 5)
Q	**ɒ**	cod	back open rounded (cardinal 13)
V	**ʌ**	bud	back open-mid unrounded (cardinal 6)
O	**ɔ**	bore	back open-mid rounded (cardinal 14)
U	**ʊ**	good	back close rounded (between cardinal 7 and cardinal 8)
u	**u**	boot	back close rounded (cardinal 16)

Table B.8: SAMPA and IPA notation for the diphthongs of English speech.

SAMPA	IPA	Example	Description
I@	**Iə**	peer	centring
E@	**eə**	hair	centring
U@	**ʊə**	poor	centring
eI	**eɪ**	pay	front closing
aI	**aɪ**	pie	front closing
OI	**ɔɪ**	boy	front closing
@U	**əʊ**	zero	back closing
aU	**aʊ**	cow	back closing

Table B.9: Notation for the Modifiers.

Modifier	Description
"	primary stress
'	secondary stress
$	syllable boundary
_	word boundary
EMPTY	

Table B.10: Annotation file of the sentence "look out of the window and see if it's raining" ("lUk aUt @v D@ wInd@U =n si If Its reInIN").

#:	_	8466	9189
l	"$	9190	10369
U	"	10370	11319
k	"_	11320	12733
aU	$	12734	15259
t	_	15260	16381
@	$	16382	17199
v	_	17200	18519
D	$	18520	19151
@	_	19152	19731
w	"$	19732	20619
I	"	20620	21545
n	"	21546	23471
d	$	23472	23625
@U	_	23626	26099
=n	_$	26100	27051
s	"$	27052	29117
i	"_	29118	30303
I	$	30304	31211
f	_	31212	32755
I	$	32756	33431
t	EMPTY	33432	33753
s	_	33754	34977
r	"$	34978	37091
eI	"	37092	39111
n	$	39112	39631
I	EMPTY	39632	41071
N	_	41072	43379
#:	_	43380	52488

Table B.11: Data file format.

Frames - total number of analysis frames.
Order - order of the analysis (default 12).
Frame length - frame duration in milli-seconds (default 30ms).
Interval - interval between frames in milli-seconds (default 10ms).
Sample frequency - (default 16KHz).
Encoding - analysis type (default 'lsp' encoding).

Start frame	Frame length	Energy	Gain	Parameter 1	...	Parameter 12
...	...	...	...	...	...	...

Table B.12: First lines in the data file of the sentence "look out of the window and see if it's raining" ("lUk aUt @v D@ wInd@U =n si If Its reInIN").

frames : 862 order : 12 frame length : 320 interval : 80 sample frequency : 16000 encoding : 'lsp'

0	320	3.175339e-04	4.791240e-02	0.075133 ...
80	320	4.033906e-08	7.624114e-06	0.057463 ...
160	320	6.525871e-08	1.291537e-05	0.049438 ...
240	320	7.166697e-08	1.496974e-05	0.036431 ...
320	320	8.799476e-08	1.741512e-05	0.038862 ...
...	...	...	...	...

Table B.13: Fragment from the data file of the sentence "look out of the window and see if it's raining" ("lUk aUt @v D@ wInd@U =n si If Its reInIN"), comprising phonemes "U" and "k".

...	...	...	...	...
10341	141	3.922142e-04	6.307191e-03	0.016897 ...
10482	139	2.831446e-04	3.897202e-03	0.016710 ...
10621	137	3.114320e-04	3.761641e-03	0.016603 ...
10758	132	3.931908e-04	4.308158e-03	0.016898 ...
10890	129	4.834232e-04	4.397213e-03	0.017314 ...
11019	123	9.799378e-04	5.719008e-03	0.017458 ...
11142	120	1.066473e-03	4.735089e-03	0.017134 ...
11262	117	1.833939e-03	6.001561e-03	0.017220 ...
11379	115	1.476675e-03	4.316185e-03	0.016933 ...
11494	112	1.240349e-03	4.495319e-03	0.016952 ...
11606	113	1.286068e-03	5.661201e-03	0.017152 ...
11719	112	1.274983e-03	7.347956e-03	0.017356 ...
11831	115	8.967329e-04	5.917798e-03	0.017502 ...
11946	123	5.846880e-04	5.206842e-03	0.017303 ...
12069	320	1.728033e-04	2.687579e-02	0.022139 ...
12149	320	6.420578e-04	5.276157e-02	0.032455 ...
12229	320	1.540061e-03	1.050520e-01	0.058675 ...
12309	320	2.047852e-03	1.385504e-01	0.068067 ...
12389	320	2.852130e-03	1.578715e-01	0.074945 ...
12469	320	3.575248e-03	2.075236e-01	0.087772 ...
12549	320	4.253044e-03	2.067872e-01	0.096153 ...
12629	320	4.512618e-03	2.201567e-01	0.094986 ...
12709	320	3.724911e-03	2.438814e-01	0.079071 ...
12789	320	4.445788e-03	2.529410e-01	0.073197 ...
...	...	...	...	...

Bibliography

[1] Jonathan Allen, M. Sharon Hunnicutt, and Dennis H. Klatt. *From Text to Speech: The MITalk System.* Cambridge University Press, 1987.

[2] Richard H. Bartels, John C. Beatty, and Brian A. Barsky. *An Introduction to Splines For Use in Computer Graphics and Geometric Modelling.* Morgan Kaufmann, 1987.

[3] Pierre E. Bézier. How Renault uses numerical control for car body design and tooling. In *Society Of Automotive Engineers Paper 680010, Automotive Engineering Congress*, pages 1–7, Detroit, U.S.A., January 1968.

[4] Pierre E. Bézier. *The Mathematical Basis of the UNISURF CAD System.* Butterworths, 1986.

[5] Gloria J. Borden, Katherine S. Harris, and Lawrence J. Raphael. *Speech Science Primer: Physiology, Acoustics, and Perception of Speech.* Williams & Wilkins, third edition, 1994.

[6] Gavin C. Cawley. *The Application of Neural Networks to Phonetic Modelling.* Ph.D. Thesis, Department of Electronic Systems Engineering, University of Essex, Essex, U.K., March 1996.

[7] Gavin C. Cawley and A. D. P. Green. The application of neural networks to cognitive phonetic modelling. In *Proceedings of the I.E.E. International Conference on Artificial Neural Networks*, pages 280–284, Bournemouth, U.K., 1991.

[8] Gavin C. Cawley and P. D. Noakes. Diphone synthesis using a neural network. In *Proceedings of the I.E.E. International Conference on Artificial Neural Networks*, pages 795–798, Brighton, U.K., 1992.

[9] Francis J. Charpentier and M. G. Stella. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *Proceedings of the I.E.E.E. International Conference on Acoustics, Speech and Signal Processing*, pages 2015–2018, Tokyo, Japan, April 1986.

[10] John Ellery Clark and Colin Yallop. *An Introduction to Phonetics & Phonology.* Basil Blackwell, 1990.

[11] Elaine Cohen and Richard F. Riesenfeld. General matrix representations for Bézier and b-spline curves. *Computers in Industry*, 3:9–15, 1982.

[12] John R. Deller, John G. Proakis, and John H. L. Hansen. *Discrete-Time Processing of Speech Signals.* Macmillan, 1993.

[13] N. Rex Dixon and H. David Maxey. Terminal analog synthesis of continous speech using the diphone method of segment assembly. *I.E.E.E. Transactions on Audio and Electroacoustics*, AU-16(1):40–50, March 1968.

[14] Mike Edgington, Andrew Lowry, Peter Jackson, Andrew Breen, and Steve Minnis. Overview of current text-to-speech techniques: Part II - prosody and speech generation. *BT Technology Journal*, 14(1):84–99, January 1996.

[15] Mike Edgington, Andrew Lowry, Peter Jackson, Andrew Breen, and Steve Minnis. Overview of current text-to-speech techniques: Part I - text and linguistic analysis. *BT Technology Journal*, 14(1):68–83, January 1996.

[16] C. Gunnar M. Fant. *Acoustic Theory of Speech Production.* Mouton, second edition, 1970.

[17] C. Gunnar M. Fant. What can basic research contribute to speech synthesis? *Journal of Phonetics*, pages 75–90, 1991.

[18] A. Robin Forrest. Interactive interpolation and approximation by Bézier polynomials. *CAD Computer-Aided Design, Special Issue: Bézier Techniques*, 22(9):527–537, November 1990.

[19] William J. Gordon and Richard F. Riesenfeld. Bernstein-Bézier methods for the computer-aided design of free-form curves and surfaces. *Journal of the Association for Computing Machinery*, 21(2):293–310, April 1974.

[20] Paulo Duarte Ferreira Gouveia. Codificação de fala por modelos variáveis no tempo. Dissertação de Mestrado, Universidade de Aveiro, Aveiro, Portugal, Abril 1996.

[21] Andrew Grace. *Optimization Toolbox User's Guide*. The MathWorks, 1994.

[22] R. M. Gray. Vector quantization. *I.E.E.E. Acoustics, Speech and Signal Processing Magazine*, pages 4–29, April 1984.

[23] John Nicholas Holmes. The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer. *I.E.E.E. Transactions on Audio and Electroacoustics*, AU21(3):298–305, June 1973.

[24] John Nicholas Holmes. Research report - formant synthesizers: Cascade or parallel? *Speech Communication*, 2(4):251–273, 1983.

[25] John Nicholas Holmes. *Speech Synthesis and Recognition*. Van Nostrand Reinhold (UK), 1988.

[26] John Nicholas Holmes, Ignatius G. Mattingly, and J. N. Shearme. Speech synthesis by rule. *Language and Speech*, 7:127–143, 1964.

[27] Wendy J. Holmes, John Nicholas Holmes, and Michael W. Judd. Extension of the bandwidth of the jsru parallel-formant synthesizer for high quality synthesis of male and female speech. In *Proceedings of the I.E.E.E. International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 313–316, Albuquerque, U.S.A., 1990.

[28] K. Ishizaka and J. L. Flanagan. Synthesis of voiced sounds from a two-mass model of the vocal chords. *Bell Systems Technology Journal*, (50):1233–1268, 1972.

[29] Fumitada Itakura. Line spectrum representation of linear predictor coefficients of speech signals. *Journal of the Acoustical Society of America*, 57(S35), 1975.

[30] Luis Miguel Teixeira de Jesus and Gavin C. Cawley. Speech coding and synthesis using parametric curves. In *Proceedings of the* 5^{th} *European Conference on Speech Communication and Technology (EuroSpeech'97)*, volume 2, pages 597–600, Rhodes, Greece, September 1997.

[31] Dennis H. Klatt. Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, 67(3):971–995, March 1980.

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

[33] M. Kohata. Interpolation of lsp coefficients using recurrent neural networks. *I.E.E. Electronics Letters*, 32(16):1441–1442, August 1996.

[34] Gernot Kubin. Nonlinear processing of speech. In W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*, chapter 16, pages 557–610. Elsevier, 1995.

[35] Peter Ladefoged. *A Course in Phonetics.* Harcourt Brace, third edition, 1993.

[36] Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly Of Applied Mathematics*, II(2):164–168, July 1944.

[37] Eric Lewis. *A 'C' Implementation of the JSRU Text-to-Speech System.* Computer Science Department, University of Bristol, August 1989.

[38] R. Linggard. *Electronic Synthesis of Speech.* Cambridge University Press, 1985.

[39] John D. Markel and Augustine H. Gray Jr. *Linear Prediction of Speech.* Springer-Verlag, 1976.

[40] Donald W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal Soc. Indust. Applied Mathematics*, 11(2):431–441, June 1963.

[41] M. V. Mathews, Joan E. Miller, and E. E. David Jr. Pitch synchronous analysis of voiced sounds. *The Journal of the Acoustical Society of America*, 33(2):179–186, February 1961.

[42] The MathWorks. *NAG Foundation Toolbox User's Guide*, 1995.

[43] P. Mermelstein. Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America*, 53(4):1070–1082, 1973.

[44] Eric Moulines and Francis J. Charpentier. Diphone synthesis using a multipulse LPC technique. In *Proceedings of Speech'88 (7th FASE Symposium)*, pages 47–53, Edinburgh, U.K., August 1988.

[45] Eric Moulines and Francis J. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5,6):453–467, December 1990.

[46] Joseph P. Olive. Rule synthesis of speech from dyadic units. In *Proceedings of the I.E.E.E. International Conference on Acoustics, Speech and Signal Processing*, pages 568–570, 1977.

[47] Joseph P. Olive and N. Spickenagel. Speech resynthesis from phoneme-related parameters. *The Journal of the Acoustical Society of America*, 59(4):993–996, April 1976.

[48] Julian H. Page and Andrew Breen. The Laureate text-to-speech system - arquitecture and applications. *BT Technology Journal*, 14(1):57–67, January 1996.

[49] Gordon E. Peterson, William S-Y. Wang, and Eva Sivertsen. Segmentation techniques in speech synthesis. *The Journal of the Acoustical Society of America*, 30(8):739–742, August 1958.

[50] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C, the Art of Scientific Computing.* Cambridge University Press, second edition, 1992.

[51] Lawrence R. Rabiner and Ronald W. Schafer. *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[52] David F. Rogers and J. Alan Adams. *Mathematical Elements For Computer Graphics*. MacGraw-Hill, second edition, 1990.

[53] J. M. Rye and John Nicholas Holmes. A versatile software parallel-formant speech synthesizer. JSRU Research Report 1016, Joint Speech Research Unit, Cheltenham, U.K., November 1982.

[54] M. G. Stella and Francis J. Charpentier. Diphone synthesis using multipulse coding and a phase vocoder. In *Proceedings of the I.E.E.E. International Conference on Acoustics, Speech and Signal Processing*, pages 740–743, Tampa, U.S.A., March 1985.

[55] John Q. Stewart. An electrical analogue of the vocal organs. *Nature*, 110(2757):311–312, 1922.

[56] Noboru Sugamura and Fumitada Itakura. Speech analysis and synthesis methods developed at ECL in NTT -from LPC to LSP-. *Speech Communication*, 5:199–215, 1986.

[57] Don X. Sun. Statistical modelling of co-articulation in continuous speech based on data driven interpolation. In *Proceedings of the I.E.E.E. International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1751–1754, Munich, Germany, April 1997.

[58] Clay M. Thompson and Loren Shure. *Image Processing Toolbox User's Guide*. The MathWorks, 1993.

[59] William S-Y. Wang and Gordon E. Peterson. Segment inventory for speech synthesis. *The Journal of the Acoustical Society of America*, 30(8):743–746, August 1958.

[60] Willard R. Zemlin. *Speech and Hearing Science. Anatomy and Physiology*. Prentice Hall, third edition, 1988.