

Speech Synthesis Based on a Harmonic Model

A thesis submitted for the degree of Doctor of Philosophy
in Computer Applications

Darragh P. O'Brien B.Sc.
School of Computer Applications
Dublin City University

February 2000

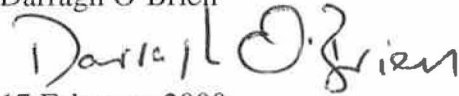
Supervisor: Dr. A. I. C. Monaghan

*This thesis is based on the candidate's own work and has not previously been
submitted for a degree at any academic institution.*

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of a Doctor of Philosophy degree in Computer Applications, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Darragh O'Brien

A handwritten signature in black ink, appearing to read 'Darragh O'Brien', written in a cursive style.

17 February 2000

For Niall

ACKNOWLEDGEMENTS

To my supervisor, Dr. Alex Monaghan, I extend my sincerest thanks. His patience, encouragement and expert guidance were both much appreciated and essential to the completion of this work.

I wish to thank my family - my parents, Noel and Sheila, and my sister, Eimear. The support of each played a fundamental role in ensuring a successful conclusion to my research.

I am grateful to my friends for making my stay in DCU a happy and humorous one. I particularly want to thank Fiona Forrestal, Lejla Rovčanin, Noel O'Hara & Rachel Mannion, John Kelleher, Tom Doris, the musical Jer Hayes and Pochacco. I hope to collaborate with all of you again on future research projects.

A final and special thanks to Marina, who often had to endure my talking about sine waves. Not only did she seem genuinely interested but often offered many insightful comments.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	ii
LIST OF FIGURES	v
ABSTRACT	vii
Chapter	
1. INTRODUCTION	1
1.1 Analysis	3
1.2 Synthesis	5
1.2.1 Formant Synthesis	5
1.2.2 Articulatory Synthesis	7
1.2.3 Concatenative Synthesis	8
1.3 Thesis Outline	14
2. SINUSOIDAL MODELLING OF SPEECH	16
2.1 The Magnitude-Only Model	16
2.2 Incorporating Phase	19
2.2.1 Phase Interpolation	19
2.3 Hybrid Models	22
2.4 The Analysis-by-Synthesis/Overlap-Add Model	25
2.5 Summary	26
3. TRANSFORMATIONS USING A SINUSOIDAL MODEL	29
3.1 Early Approach	30
3.1.1 Time-Scale Modification	32
3.1.2 Frequency-Scale and Pitch Modification	34
3.1.3 Summary	35
3.2 Shape Invariant Transformations	35
3.2.1 Shape Invariant Time-Scale Modification	36
3.2.2 Shape Invariant Pitch Modification	40
3.2.3 Summary	42
3.3 The Harmonic Plus Noise Model	43
3.4 The Analysis-by-Synthesis/Overlap-Add Model	46
3.5 Discussion	48

4.	TRANSFORMATIONS USING A HARMONIC MODEL . . .	52
4.1	Analysis	53
4.2	Time-Scale Modification	53
4.2.1	Voiced Speech	54
4.2.2	Voiceless Speech	59
4.3	Pitch Modification	61
4.4	Joint Pitch and Time-Scale Modification	67
4.5	Discussion	71
5.	SYNTHESIS	74
5.1	AT&T's Next Generation TTS System	75
5.1.1	Synthesis	75
5.1.2	Experimental Results	78
5.2	The ABS/OLA TTS System	78
5.2.1	Synthesis	79
5.2.2	Experimental Results	80
5.3	The SHMDCU TTS System	80
5.3.1	Synthesis	81
5.3.2	Experimental Results	84
5.4	Summary	88
6.	RESULTS	89
6.1	The COST 258 Coder Evaluation Server	89
6.2	Prosodic Transplantation	90
6.3	Experimental Procedure	91
6.4	Results	96
6.5	Discussion	97
7.	CONCLUSIONS AND FUTURE WORK	98
7.1	References	101

LIST OF FIGURES

Figure		Page
1.	Text-to-speech synthesis (from [2])	3
2.	Nearest-neighbour matching (from [46])	18
3.	Phase-unwrapping (from [36])	20
4.	Sinusoidal analysis/synthesis model (from [36])	23
5.	Refined sinusoidal analysis/synthesis model (from [55])	33
6.	Pitch pulse onset time estimation with $\rho = 2$ (from [37])	37
7.	Shape-invariant time-scale modification (from [37])	39
8.	Pitch pulse onset time estimation with $\beta = 2$ (from [37])	41
9.	PSOLA algorithm (from [31])	45
10.	Original speech, $\rho = 1$	56
11.	Time-scaled speech, $\rho = 0.6$	57
12.	Time-scaled speech, $\rho = 1.3$	57
13.	Time-scaling algorithm	58
14.	Bandwidth expansion in voiceless speech	60
15.	Iterative adaptive inverse filtering (IAIF) algorithm (from [65])	63
16.	Original speech, $\lambda = 1$	66
17.	Pitch-scaled speech, $\lambda = 0.7$	66
18.	Pitch-scaled speech, $\lambda = 1.6$	67
19.	Pitch-scaling algorithm	68

20.	Original speech, $\rho = 1$, $\lambda = 1$	69
21.	Pitch- and time-scaled speech, $\rho = 0.7$, $\lambda = 0.7$	70
22.	Pitch- and time-scaled speech, $\rho = 1.6$, $\lambda = 1.6$	70
23.	Diphone boundary waveform with phase mismatch	85
24.	Diphone boundary waveform with phase mismatch correction . .	85
25.	Diphone boundary waveform with phase mismatch	86
26.	Diphone boundary waveform with phase mismatch correction . .	86
27.	Diphone boundary spectrum with phase mismatch	87
28.	Diphone boundary spectrum with phase mismatch correction . .	87
29.	Original waveform	91
30.	Original pitch contour	92
31.	Pitch- and time-scaled waveform	92
32.	Target and actual pitch contours	93
33.	Original waveform	93
34.	Original pitch contour	94
35.	Pitch- and time-scaled speech	94
36.	Target and actual pitch contours	95

ABSTRACT

The wide range of potential commercial applications for a computer system capable of automatically converting text to speech (TTS) has stimulated decades of research.

One of the currently most successful approaches to synthesising speech, concatenative TTS synthesis, combines prerecorded speech units to build full utterances. However, the prosody of the stored units is often not consistent with that of the target utterance and must be altered. Furthermore, several types of mismatch can occur at unit boundaries and must be smoothed. Thus, pitch and time-scale modification techniques as well as smoothing algorithms play a critical role in all concatenative-based systems.

This thesis presents the development of a concatenative TTS system based on a harmonic model and incorporating new pitch and time-scaling as well as smoothing algorithms.

Experiment has shown our system capable of both very high quality prosodic modification and synthesis. Results compare very favourably with those of existing state-of-the-art systems.

CHAPTER 1

INTRODUCTION

Speech is the most natural form of communication between humans. It follows that for many years much effort has been directed towards enabling computers to convey information to users through the same medium. Decades of research, stimulated by the potentially large number of applications for a computer system capable of producing speech of truly human quality, has led to the development of text-to-speech (TTS) synthesis systems. Given an input text the aim of such systems is to automatically transform it into speech.

Some commercial applications of TTS systems include remote access to e-mail, stock prices, weather forecasts, bank details etc. Indeed, TTS is of potential use in all areas where textual data must be accessed remotely. On a social level TTS technology can be used to provide speaking aids for the vocally impaired and reading aids for the visually impaired as well as benefiting those with reading disabilities such as dyslexia.

At first the task of a TTS might seem most easily accomplished by simply concatenating prerecorded words to construct any target utterance. This approach has, however, been found to be unsuitable for all but a small number of applications. Some of the problems associated with word based concatenative TTS are outlined below.

- A natural utterance is very different to a set of isolated words uttered sequentially. In natural speech coarticulation serves to fuse word boundaries, an effect which is important for the perception of natural-

ness [1].

- Natural speech has intonation (e.g. rhythm, stress, timing and pitch) which serves to “bring an utterance to life”. Simply stringing isolated words together therefore results in a stilted, lifeless and monotone synthetic speech quality [1].
- Recording and storing all the words in the English language is not a practical option. Furthermore, new words are created every day and there exists an unlimited number of proper names [1].

Word based concatenative TTS systems are only of use in a well defined and restricted area of application where the prosodic characteristics and content of all possibly required utterances can be specified in advance. This is however not the case in many of TTS systems’ most potentially useful applications. For large or unrestricted vocabulary applications a more sophisticated model is required and modern approaches draw from such fields as linguistics, phonology and digital signal processing.

In most modern TTS systems the process of getting from raw textual input to spoken output is broken into at least two stages - analysis, which produces an abstract linguistic representation of the text, and synthesis, which takes the output of the analysis phase and produces a corresponding speech waveform. This process is depicted in Figure 1 (taken from [2]).

The remainder of this chapter serves two purposes - firstly, to briefly describe the steps in Figure 1 and give an overview of TTS technology and secondly, to situate the research undertaken in this thesis within that overview. In Section 1.1 the analysis phase is outlined. The output of this stage tells the synthesiser *what* to synthesise (i.e. supplies it a string of phonemes) and *how* to synthesise it (i.e. supplies it with target pitch, duration and intensity

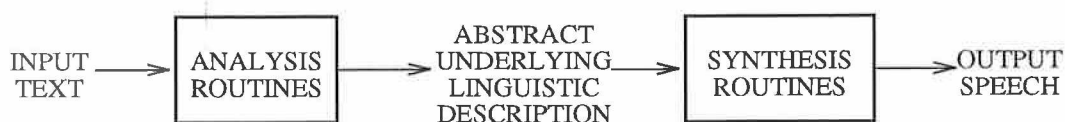


Figure 1 Text-to-speech synthesis (from [2])

values for each phoneme). Section 1.2 describes some of the approaches that have been adopted for implementing the synthesis backend of a TTS system. The chapter concludes with a brief outline of the rest of this thesis in Section 1.3.

1.1 Analysis

The purpose of the analysis phase is to deliver to the synthesis backend an abstract linguistic representation of the input text. Typically this representation consists of a string of phonemes annotated with target pitch, duration and intensity values. The first step in analysis is to convert the raw text to phonemes and the second is to ascribe, to each one, appropriate prosodic target values.

Converting the text to a string of phonemes is not as simple as might be thought. The text may contain numbers, dates, abbreviations and symbols which must all be expanded to word form during a process termed text normalisation. Once this has been done a dictionary look-up procedure is used to retrieve each word's phonetic transcription. If a word is not in the dictionary, (it may be an unrecognised proper name, for example), fallback grapheme-to-phoneme rules are applied to generate a phonetic transcription. Once a sequence of phonetic symbols has been generated, a phonological component is applied and makes allophonic substitutions dependent on the

phonetic context. This completes the phonetic specification of the target utterance which must subsequently be annotated to reflect the prosody of a natural utterance.

A natural prosodic contour is a product of many factors including syntax, semantics and pragmatics. Because much pragmatic information is not available to the synthesiser [3] an attempt is often made to assign a context neutral prosodic contour to each utterance. Syntactic and semantic analyses along with heuristic strategies [4] [5] all contribute to this end.

A syntactic analysis is first carried during which part of speech tags are assigned to each word (thus resolving certain category ambiguities) and a parse of the utterance is generated. Statistical based or phrase structure grammars are commonly used for this purpose. Phrasal and clausal boundaries are located as they have a marked effect on prosody. A dictionary look-up procedure is then used to access lexical stress information which has been stored for each word. Such information is important since depending on its lexical stress pattern a word such as "object" can have two distinct meanings. Semantic analysis may then be applied to resolve remaining ambiguities, distinguish function from content words and assign emphasis.

Typical duration and intensity values for each phoneme are retrieved from a database and adjusted based on the now available syntactic and semantic information (while also taking phonetic context into account) [6]. A model of intonation [7] [8] [9] is then applied to produce a pitch contour using which a target F0 value can be assigned to each phoneme. This completes our brief discussion of the analysis phase, a more detailed account of the issues involved is given in [9]. After analysis the derived abstract representation is passed to the synthesis backend for conversion into speech.

1.2 Synthesis

Over the years three main approaches to synthesising speech have emerged and each is outlined in this section. Section 1.2.1 presents formant synthesis, one of the earliest approaches to modelling speech production. Incorporating parametric models of the excitation signal and vocal tract, this method uses rules to describe coarticulatory effects. Articulatory synthesis, discussed in Section 1.2.2 is based on a more fine-grained modelling of the speech production process - vocal folds, articulators and sections of the vocal tract are all separately simulated. Rules are again used, this time to describe the behaviour of and interaction between the model's various components. Finally concatenative, currently the most popular approach and successful approach to TTS is presented in Section 1.2.3. Concatenative systems rely on a database of prerecorded speech units which are combined to build full utterances. Such an approach necessitates the implementation of pitch and time-scale modification as well as smoothing algorithms, some of which are also discussed in this section.

1.2.1 *Formant Synthesis*

Formant synthesis, also called synthesis-by-rule, is based on Fant's [10] source-filter theory of speech production. According to this theory the speech signal may be thought of as that produced when a linear filter, representing the vocal tract, is excited by one or more sources [1]. The source may be the glottal excitation signal produced by the vocal cords, or noise produced at some point of constriction along the vocal tract, or a mixture of the two. Clearly, such a model offers an extra degree of flexibility as both source and filter parameters can be modified separately. Assuming the independence of

source and vocal tract contributions to the speech production process, while this is not strictly the case, greatly simplifies the implementation of pitch and time-scaling algorithms. The ability to synthesise speech with specified pitch and duration is essential in all TTS systems as the prosody of an utterance contributes significantly to both its naturalness and its meaning.

In its simplest form, a formant synthesiser consists of a linear filter and a parametric source model. The linear filter is of an all-pole type, each complex conjugate pole pair accounting for a single formant or resonance frequency in the vocal tract. Numerous models of varying complexity have been suggested for the glottal excitation signal produced during voiced speech [11] [12] [13]. White noise is used to drive the filter during voiceless speech. Other sounds such as voiced fricatives are synthesised using a mixture of white noise and glottal pulses.

Before speech can be synthesised the information required to drive a formant synthesiser must be available. As described in Section 1.1 the analysis component delivers an abstract representation, consisting of phonetic strings annotated for pitch, duration and intensity, to the synthesis backend. In a formant synthesiser a table look-up is performed and returns a set of typical parameter values for each phoneme which include, for example, formant amplitude, frequency and bandwidth information. The acoustic realisation of a particular phoneme is, however, heavily influenced by context and therefore coarticulation rules need to be applied to adjust parameters to reflect a given phonetic context. Other, often complex, rules are then used to describe the evolution in time of each parameter track from one target to the next. Once this set of time-dependent interpolation functions has been derived (there are 19 such functions in the Klattalk system [12]) they are used to drive the formant synthesiser and synthetic speech is produced.

Although formant synthesis can be used to produce a perceptually indistinguishable copy of a natural utterance, the process is far from automatic. Target parameter values and interpolation rules must often be supplied by hand after a trial and error process that requires considerable expertise. Until the acoustic phenomena at work during speech production, particularly coarticulatory effects, are better understood, general rules for controlling formant synthesisers to produce high quality speech remain out of reach [14].

1.2.2 Articulatory Synthesis

Articulatory models take formant synthesis a step further and attempt to provide a more detailed modelling of the physics of speech production. Instead of using a single filter to model the vocal tract, the tract is divided into several smaller sections and each is individually modelled [1]. In this method a phoneme is represented by a typical configuration of the vocal tract specifying cross sectional areas, airflow volume velocity and any other relevant acoustic characteristics. Several such models have been proposed [16] [18] [17] [15]. Mathematical models of vocal fold behaviour have also been proposed [19] [20]. As with formant synthesis rules must be incorporated describing the behaviour of and interaction between each of the model's components. Given the paucity of scientific data pertaining to the dynamics of speech production such models must be approximations and contain many simplifications. As a consequence, speech quality suffers and does not, as yet, compare favourably with that produced by other methods (see Section 1.2.3). Furthermore, the computational costs involved in implementing an articulatory model are currently extremely high, making it unsuitable for real-time synthesis applications [1].

Articulatory synthesis may in the long-run deliver truly human-sounding synthetic speech [1] but until more is known about the complex physical processes involved in speech production other approaches will continue to produce speech which is of both higher quality and naturalness.

1.2.3 Concatenative Synthesis

Given the problems of rule definition associated with formant and articulatory synthesis many researchers have tried to circumvent the problem of attempting to model the physics of speech production by storing units of natural speech and using them as building blocks which are concatenated to create synthetic speech.

For the reasons outlined at the beginning of this chapter words are unsuitable as the basic unit. The phoneme, of which there are about 40 in English, would seem a more suitable unit to use. However, because of coarticulatory effects, the acoustic realisation of any particular phoneme can undergo extensive modification depending on its context [21]. As a result, a concatenative TTS system based on phonemes must incorporate coarticulatory rules and we return to the problems associated with both formant synthesis and articulatory synthesis.

As a solution and in order to avoid the problem of explicitly modelling coarticulation, the diphone has been proposed as a concatenative unit suitable for use in TTS applications [22]. A diphone consists of that speech segment from the centre (and most stable part) of one phoneme to the centre of the following one. There are about 1600 (40×40) diphones in English. There is some evidence however that a diphone database containing only a single copy of each diphone can be shown by listening tests to be detectably

deficient [23]. For this reason many diphone databases are extended to include stressed and unstressed vowels as well as allophonic diphone variations.

As the storage and memory capacity of computers continue to rise, so too do the size and complexity of speech unit repositories continue to grow. AT&T's Next-Generation TTS System [23] reflects this trend and uses a speech database consisting of 1.5 hours of recorded speech from a female speaker. Since the database contains several instances of any particular diphone, a unit selection algorithm is used to select, at run-time, an optimal sequence of diphones for synthesis [24]. This system is discussed in more detail in Chapter 5.

Although diphones in theory eliminate the problem of modelling at least local coarticulatory effects, their use poses several problems. In order to effect a desired prosodic contour the pitch, duration and intensity of each diphone may need to be altered. Furthermore, as they are extracted from disjoint phonetic contexts discontinuities in spectral shape, energy, pitch and phase often occur at unit boundaries and must be smoothed. Consequently sophisticated techniques for prosodically modifying and smoothing acoustic units are an integral part of any concatenative synthesiser. Output speech quality is critically dependent on the performance of these algorithms. That work remains to be done in this area has been indicated by listening tests which show that listeners consistently prefer speech synthesised from units which have not undergone prosodic changes (and which therefore contain slight discontinuities) to speech where even minor prosodic modifications have been applied to the concatenative units [24].

Nonetheless, concatenative systems have resulted in clear gains in the quality and naturalness of synthetic speech and as unit databases continue to grow in size and unit selection algorithms continue to become more sophis-

ticated these improvements are set to continue. Presented below are some of the techniques which have been used to implement the prosodic modification and smoothing techniques essential to concatenative TTS systems.

LPC

LPC (Linear Predictive Coding) [25] [26] [27] [28], again based on Fant's source-filter theory [10], stems from the idea that a speech sample, in the absence of any excitation, can be approximated as a weighted sum of a number of previous samples. Typically LPC analysis is applied on a frame-by-frame basis to the speech waveform, generating for each frame an LPC filter which simultaneously models the glottal excitation pulse shape, vocal tract and lip radiation effects. A simple pulse train is used to excite the predictor during voiced speech. During voiceless speech the predictor is driven by white noise. A mixture of the two excitation modes is used to synthesise voiced fricatives.

LPC has long been used in concatenative TTS systems to re-synthesise, prosodically transform and smooth acoustic units [29]. During synthesis, predictor coefficients are smoothed across concatenation points and the nature of the excitation sources makes pitch and duration modification trivial. However, it is this simplicity of excitation signal together with the assumptions inherent in traditional LPC analysis which results in poor quality re-synthesis. Speech produced by coders of this type suffers from an unnatural "buzzy" quality.

Analysis-by-synthesis techniques have been proposed to provide LPC coders with a more realistic excitation source and thus increase speech quality. One such approach is multipulse LPC (MPLPC) [30] where instead of using one excitation pulse per pitch period several are used. The location and amplitude of each pulse is chosen such that a perceptually weighted error

criterion is minimised. Although capable of reproducing speech perceptually indistinguishable from the original with a relatively small number of pulses the rigidity of the excitation pulse sequence makes pitch and duration modifications cumbersome and difficult to implement.

PSOLA

PSOLA (pitch-synchronous overlap add) [31] [32] [33] is a widely used “model” for performing pitch and time-scale modification. Using a non-parametric representation of the speech waveform, it is a simple and efficient algorithm capable of high quality results.

During “analysis” the speech is pitch marked i.e. points of glottal closure are calculated. Short-time analysis signals are extracted by scaling the speech by a window (e.g. Hanning), typically of a length twice the local pitch period, repeatedly centred on each pitch mark. It is these short-time signals which serve as building blocks during pitch and time-scale modification.

From the target prosody a set of synthesis pitch mark locations can be calculated along the synthesis time-axis. A mapping is established from each synthesis pitch mark to an analysis signal. Analysis signals are then synchronised on their corresponding synthesis pitch mark and are overlapped and added to produce speech conforming with the target prosodic contour.

In other words time-scale modification amounts to duplicating or deleting analysis signals while pitch modification involves increasing or decreasing the time interval between analysis signals. Both operations are carried out along the synthesis time-axis.

PSOLA, although capable of high quality modifications over stationary speech segments, does not perform so well on voiceless speech. The repetition of voiceless frames during time-scale expansion can introduce an unnatural

tonality into the modified speech as periodicities become perceptible over extended voiceless intervals. Furthermore, given its non-parametric nature, PSOLA's smoothing capabilities are of limited scope. Lastly, accurate pitch marking of the speech database is required and although automatic pitch marking detection algorithms exist, their output must often be hand corrected making it a laborious and time-consuming process. PSOLA's main advantage is its speed - synthesis can be implemented with only 7 operations per sample [34] making real-time TTS feasible [35].

Sinusoidal Models

Traditional LPC coders employ different excitation models depending on whether voiced or voiceless speech is being synthesised. This requires accurate \pm voice decisions during analysis, a non-trivial task. Partially voiced frames pose further problems for classification. A solution, developed by McAulay and Quatieri, is the sinusoidal coder [36]: by modelling the individual frequency components of the speech signal they found that speech which was perceptually indistinguishable from the original could be re-synthesised. Importantly, the sinusoidal approach requires no distinction between voiced and voiceless frames. Quatieri and McAulay [37] extended the original sinusoidal model to permit pitch and time-scale modification of speech. These techniques combined with the sinusoidal model's parametric speech representation (allowing for both more sophisticated control and smoothing of signal properties than non-parametric PSOLA, for example) make it a strong candidate for use in concatenative TTS.

Building on the work of McAulay and Quatieri, George and Smith developed an analysis-by-synthesis/overlap-add (ABS/OLA) speech model [38] [39] [40]. Capable of high quality prosodic modification and using an efficient

inverse-FFT procedure for re-synthesis, the ABS/OLA model is well suited to TTS. Such an application, based on an improved ABS/OLA model, has been proposed by Macon and Clements [41] [42].

Sinusoidal modelling however, like PSOLA, has problems implementing time-scale expansion of voiceless speech. The aperiodic speech component, in theory consisting of an infinite number of sinusoids, in the sinusoidal model is approximated by the sum of a discrete number of frequency components. When stretched over an expanded time-interval periodicities become perceptible in noise modelled in this way resulting in an unwanted tonal character being imparted to the noise. Macon [41] has proposed phase and frequency dithering techniques to maintain perceptual randomness. Other researchers have taken the more “drastic” approach of proposing hybrid models to separately model the deterministic and stochastic components of the speech signal [43] [44].

In one such hybrid model, HNM (harmonic plus noise model) [44], the deterministic component is modelled as a sum of harmonics while an LPC filter describes the spectral density of the stochastic component. The separate modelling of these two components adds an extra degree of flexibility as each can be independently modified in different ways. In HNM, for example, a PSOLA type algorithm is applied to the harmonic part and as the stochastic part is modelled by LPC-filtered white noise the problem of tonality does not arise. Using PSOLA in this way, i.e. to do only that for which it is best suited and using a separate model for non-stationary speech components, gives the possibility of higher quality results than traditional PSOLA [45]. As mentioned, sinusoidal modelling, unlike PSOLA, also lends itself to simple smoothing techniques. For these reasons HNM was chosen by AT&T to serve as the backend in their Next Generation TTS System which is claimed

to produce extremely high quality synthetic speech.

1.3 Thesis Outline

Given the current popularity and high quality output of concatenative TTS systems based on the sinusoidal model, a programme of research was undertaken with the aim of developing a similarly based system while improving on existing algorithms for pitch modification, time-scaling and smoothing at concatenation points. This thesis presents the course of that development while contrasting our approach to that of other systems throughout. The structure of the remaining chapters is as follows:

- Chapter 2: The original sinusoidal model, as proposed by McAulay and Quatieri is presented in some detail along with two of its derivatives - the ABS/OLA and HNM models.
- Chapter 3: The various techniques employed by each of the models presented in Chapter 2 for implementing pitch and time-scale modification of speech are discussed.
- Chapter 4: New pitch and time-scaling techniques developed for use within our concatenative TTS synthesiser are presented in detail. A novel approach to frequency dithering to be applied when time-scale expanding voiceless regions is also put forward.
- Chapter 5: Existing TTS applications based on HNM and ABS/OLA along with our new synthesis system using those algorithms developed in Chapter 4 are discussed. Each system's approach to handling discontinuities at unit boundaries is analysed. A new phase mismatch correction algorithm is proposed and incorporated into our system.

- Chapter 6: Results of formal listening tests carried out to compare the performance of our pitch and time-scaling techniques against two other frequency domain based approaches are discussed.
- Chapter 7: Conclusions and suggestions for future work are presented.

CHAPTER 2

SINUSOIDAL MODELLING OF SPEECH

In Chapter 1 it was established that in order to qualify for use within a concatenative TTS system a speech coder must be capable of both high quality re-synthesis and high quality prosodic manipulation. In this chapter the sinusoidal model [46] [36] of speech is introduced and its fulfilment of the first of these requirements is demonstrated. (Chapter 3 deals with the model's prosodic modification capabilities.) The chapter is organised as follows. In Section 2.1 the early magnitude-only model [46] is presented. This approach is extended in Section 2.2 to incorporate a more sophisticated approach to phase interpolation [36]. A number of other models have been developed based on the sinusoidal approach. Sections 2.3 and 2.4 deal, respectively, with the harmonic plus noise model (HNM) [44] [47] and the analysis-by-synthesis/overlap-add model (ABS/OLA) [38] [39] [40]. Both are derivatives of the original sinusoidal model. The chapter concludes with a summary in Section 2.5

2.1 The Magnitude-Only Model

Hedelin proposed a model of speech wherein each frame was modelled as a sum of sinusoids according to

$$\hat{s}^l(n) = \sum_{k=1}^{K^l} A_k^l(n) \cos[\theta_k^l(n)] \quad (2.1)$$

where $\hat{s}^l(n)$ is the l^{th} speech frame, K^l is the number of sinusoids in the frame and $A_k^l(n)$ and $\theta_k^l(n)$ represent, respectively, the time-varying amplitude and

phase of the k^{th} sinusoid [48] [49]. In Hedelin's model the phase of each sinusoid is defined to be the integral of its instantaneous frequency and may thus be written

$$\theta_k^l(n) = \theta_k^l(n-1) + \omega_k^l(n) \quad (2.2)$$

where $\omega_k^l(n)$ is the time-varying frequency of the k^{th} sinusoid. Importantly, and as pointed out by McAulay and Quatieri [46], modelling phase in this way, i.e. in terms of the instantaneous frequency, ensures phase continuity and thus waveform continuity. In other words, as long as each component frequency is smoothly interpolated so too, by definition, is phase and the generation of a smooth discontinuity-free waveform is guaranteed.

McAulay and Quatieri developed a sinusoidal analysis/synthesis model [46] based on Hedelin's approach. During analysis each speech frame is windowed and its FFT computed. The FFT is then scanned for peaks whose amplitudes and frequencies are coded. McAulay and Quatieri's innovation on Hedelin's model lies in the way parameters are interpolated from one frame to the next. They proposed a nearest-neighbour matching algorithm whereby each frequency in frame l is matched with the closest available frequency in frame $l+1$ provided the difference in frequency between the two lies within a specified bandwidth. Over stable speech segments, e.g. sustained vowels, this matching is quite straightforward, as the frequency content of adjacent frames remains reasonably constant. Over transitions however, e.g. between voiced and voiceless speech, rapid frequency fluctuations are introduced into the speech signal. To model such variability the notions of "birth" and "death" of frequencies are added to the model. A frequency in frame l , if it cannot be matched with a frequency in frame $l+1$ is said to die in frame $l+1$ where its target amplitude is set to zero. Conversely, if a frequency in frame $l+1$ cannot be matched with one in frame l , it is said to have been born in

frame l where its start amplitude is set to zero. This matching algorithm is depicted in Figure 2.1.

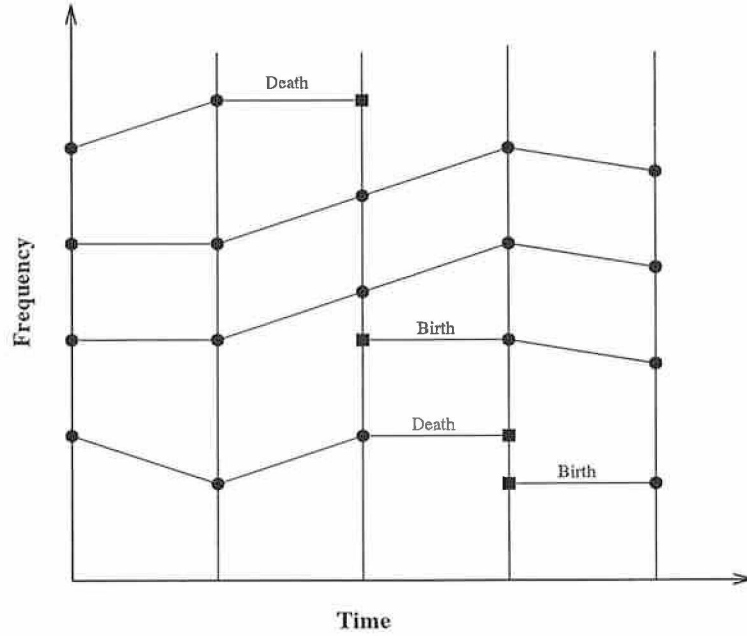


Figure 2 Nearest-neighbour matching (from [46])

During synthesis amplitudes and frequencies are linearly interpolated. Assuming, for simplicity, that the k^{th} frequency in frame l has been matched with the k^{th} frequency in frame $l+1$ this gives

$$\begin{aligned} A_k^l(n) &= A_k^l + \frac{A_k^{l+1} - A_k^l}{S}n \\ \omega_k^l(n) &= \omega_k^l + \frac{\omega_k^{l+1} - \omega_k^l}{S}n \end{aligned} \quad (2.3)$$

where S is the frame interval length. Speech may then be re-synthesised from Equation (2.1).

Obviously, given the simplicity of the treatment of phase, i.e. defining it to be the integral of instantaneous frequency, phase relations inherent in the

original are not retained in the reconstructed speech. The human ear is, however, “phase deaf” [50] and sensitive only to the short-time spectral magnitude of the speech signal, which is preserved. McAulay and Quatieri reported very high quality speech re-synthesis for both male and female speakers using this model [46].

2.2 Incorporating Phase

While it is generally agreed that the ear is normally insensitive to phase [50], experiment has shown that under certain circumstances phase does play a role in speech perception [51] [52]. Sinusoidal modelling appears to be a case in point for although the magnitude-only model described in Section 2.1 achieves high quality re-synthesis, listeners perceive a change in voice quality compared to the original speech. Specifically, an unnatural reverberant quality is present in the re-synthesised version. This effect can be attributed to a lack of phase coherence, i.e. because phase relations present in the original are discarded, the re-synthesised speech does not share the same time-domain shape. As phases are allowed to “wander about”, unconstrained, waveform dispersion occurs and a reverberant artifact is introduced.

2.2.1 *Phase Interpolation*

To avoid the reverberation just described and to preserve the original waveform shape a more sophisticated handling of phase information must be incorporated into the sinusoidal model. The matter is complicated, however, by two factors. Firstly, phase and frequency are bound together - phase is defined as the integral of instantaneous frequency. Secondly, as phase is evaluated modulo 2π , “phase-unwrapping” must be carried out, a process

depicted in Figure 2.2.1. Adding any multiple of 2π to a measured target phase value does not affect a sinusoid's final phase angle but does alter its average frequency across the synthesis interval. Phase-unwrapping thus consists of choosing a suitable multiple of 2π to add to the measured target phase. McAulay and Quatieri's solution to these problems is presented in [36] and outlined below.

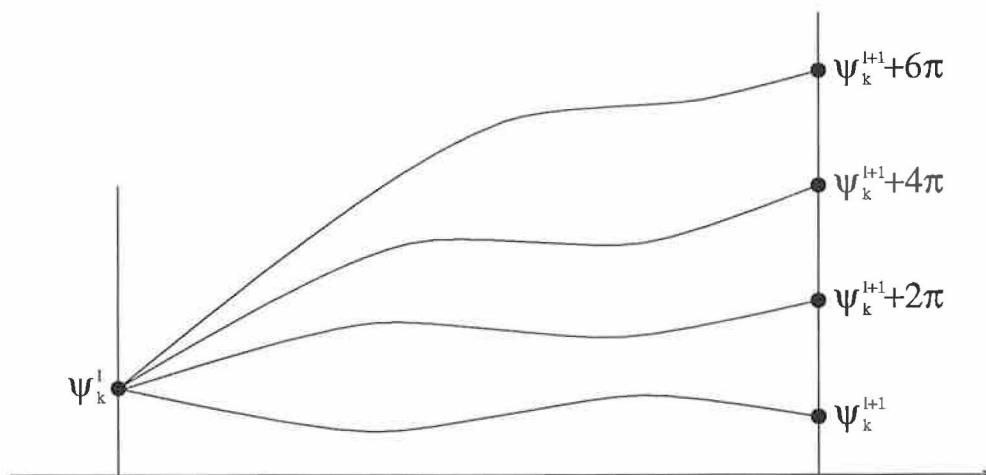


Figure 3 Phase-unwrapping (from [36])

After peak-picking, let $\{A_k^l, \omega_k^l, \psi_k^l\}$ and $\{A_k^{l+1}, \omega_k^{l+1}, \psi_k^{l+1}\}$ denote the instantaneous amplitude, frequency and phase of the k^{th} sinusoid at the centre of frames l and $l+1$ respectively. Amplitude is interpolated linearly as in Section 2.1. McAulay and Quatieri postulate a cubic function, given in Equation (2.4), to model phase interpolation. Given that instantaneous frequency is defined as the derivative of phase, the phase and frequency of each sine wave at any time n are given by Equations (2.4) and (2.5) respectively.

$$\tilde{\theta}(n) = \zeta + \gamma n + \alpha n^2 + \beta n^3 \quad (2.4)$$

$$\dot{\tilde{\theta}}(n) = \gamma + 2\alpha n + 3\beta n^2 \quad (2.5)$$

Setting $n = 0$ and substituting the known phase and frequency values obtained from the FFT analysis into (2.4) and (2.5) gives

$$\begin{aligned}\tilde{\theta}(0) &= \zeta = \psi_l^k \\ \dot{\tilde{\theta}}(0) &= \gamma = \omega_l^k\end{aligned}\quad (2.6)$$

Similarly, substituting the known phase and frequency values, when $n = S$ gives

$$\begin{aligned}\tilde{\theta}(S) &= \zeta + \gamma S + \alpha S^2 + \beta S^3 = \psi_l^{k+1} + 2\pi M \\ \dot{\tilde{\theta}}(S) &= \gamma + 2\alpha S + 3\beta S^2 = \omega_l^{k+1}\end{aligned}\quad (2.7)$$

As mentioned, the target phase ψ_k^{l+1} is measured modulo 2π so phase unwrapping must be performed and the $2\pi M$ term is included in Equation (2.7) where M is an integer. It remains to solve for the three unknowns α , β and M . For any M , α and β can be calculated from

$$\begin{bmatrix} \alpha(M) \\ \beta(M) \end{bmatrix} = \begin{bmatrix} 3/S^2 & -1/S \\ -2/S^3 & 1/S^2 \end{bmatrix} \begin{bmatrix} \psi_k^{l+1} - \psi_k^l - \omega_k^l S + 2\pi M \\ \omega_k^{l+1} - \omega_k^l \end{bmatrix}\quad (2.8)$$

In [36] the value of M is chosen such that a maximally smooth frequency track is obtained. This is achieved by minimising $f(x)$ in Equation (2.9) with respect to the continuous variable x . $f(x)$ can be regarded as a measure of the amount of variation in frequency of each sinusoid across a frame interval.

$$f(x) = \int_0^S [\ddot{\theta}(n; x)]^2 dn\quad (2.9)$$

The minimising value of x can be shown to be that given by Equation (2.10). Rounding to the closest integer gives M^* , as shown in Equation (2.11), where $[[\]]$ denotes the “nearest integer” operator.

$$x^* = \frac{1}{2\pi} \left[(\psi_k^l + \omega_k^l S - \psi_k^{l+1}) + (\omega_k^{l+1} - \omega_k^l) \frac{S}{2} \right]\quad (2.10)$$

$$M^* = [[x^*]]\quad (2.11)$$

Once M^* has been determined, $\alpha(M^*)$ and $\beta(M^*)$ are computed from Equation (2.8) completing the model. Speech may then be re-synthesised from Equation (2.12).

$$\hat{s}^l(n) = \sum_{k=1}^{K^l} A_k^l(n) \cos[\tilde{\theta}_k^l(n)] \quad (2.12)$$

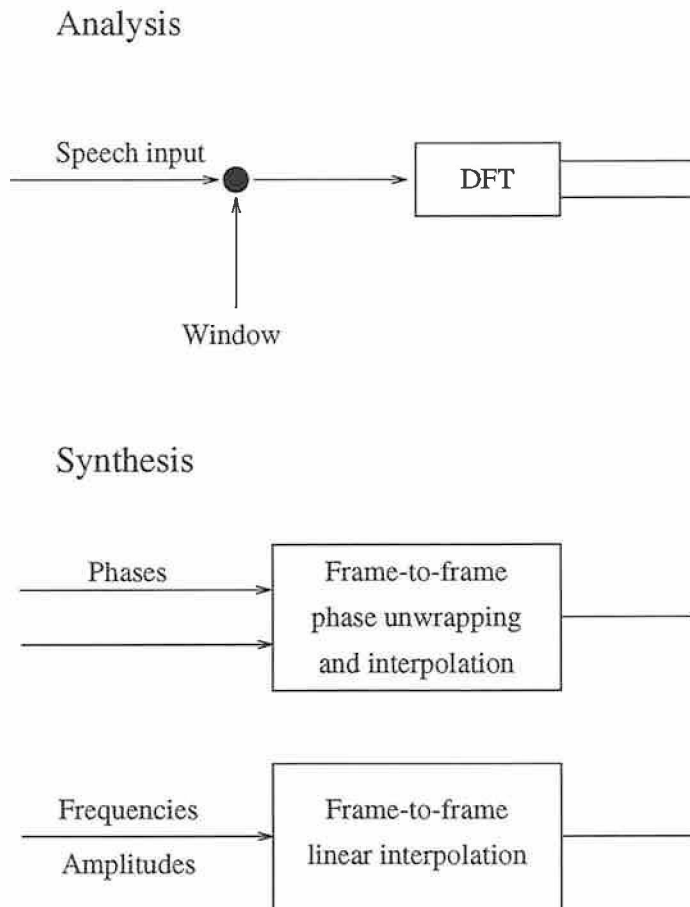
McAulay and Quatieri report that speech re-synthesised using this method was found by listeners to be free of reverberation and, furthermore, perceptually indistinguishable from the original [36]. The shape of the original speech was also well preserved. The complete analysis/synthesis system is depicted in Figure 4.

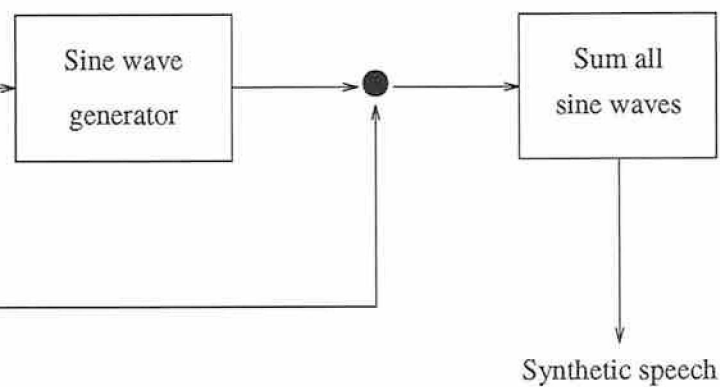
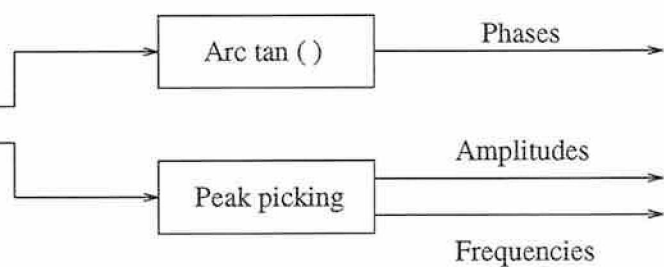
2.3 Hybrid Models

In theory an infinite number of sinusoids are required to accurately model noise. However, McAulay and Quatieri maintain that the aperiodic component of the speech signal can be adequately modelled by a finite number of sinusoids provided they are rapidly varying in frequency and “close enough” together to produce a relatively smooth spectrum (i.e. one free of harmonic structure) [36]. While this approach works well for straightforward re-synthesis where the random nature of noise is well preserved, difficulties arise when time-scale expansion is applied to voiceless regions coded in this way. These problems are dealt with in more detail in Chapter 3.

For these reasons some researchers have proposed the separate modelling of deterministic and stochastic components in the speech signal. By modelling the stochastic component in a more “suitable” and economic way, the problems mentioned above can be eliminated. Such models have been proposed by Serra [53] and Griffin and Lim [54]. Given its current popularity we concentrate here on the harmonic plus noise model (HNM) [44] [47].

Figure 4 Sinusoidal analysis/synthesis model (from [36])





In HNM the approach adopted is to model the deterministic component as a set of harmonically related sinusoids while the stochastic component is modelled with an LPC filter along with a time-domain parametric envelope. For the l^{th} frame this gives

$$\hat{s}^l(n) = \sum_{k=1}^{K^l} A_k^l(n) \cos[k\theta^l(n) + \phi_k^l(n)] + e(n) \quad (2.13)$$

where

$$\theta^l(n) = \int_{-\infty}^n \omega_0^l(t) dt \quad (2.14)$$

and where $\hat{s}^l(n)$ is the input speech signal, K^l is the number of harmonics, $A_k^l(n)$ and $\phi_k^l(n)$ are the time-varying amplitude and phase of the k^{th} harmonic. ω_0^l is the fundamental frequency and $e(n)$ is the stochastic component's contribution.

During analysis (which is pitch synchronous) an initial pitch estimate is assigned to each frame and a voicing decision made by comparing the actual speech spectrum to a synthetic one based on the estimated pitch. If a frame is deemed voiced (i.e. there is a reasonably good match between the two spectra) a peak-picking algorithm is used to separate the actual speech spectrum into two bands. The lower band is considered deterministic and the upper band stochastic. The boundary between the two is defined by F_{max} , the maximum voicing frequency. A new and more accurate pitch estimate is then calculated by finding the F0 value whose harmonics best fit the deterministic band. Analysis frames are then set at a pitch-synchronous rate across voiced regions and at a fixed rate over voiceless regions. For the l^{th} voiced frame, where n_i is the frame centre, the harmonic component can be written

$$h^l(n) = \sum_{k=1}^{K^l} A_k^l(n_i) \cos[kn\omega_0^l + \phi_k^l(n_i)] \quad (2.15)$$

The parameters $A_k^l(n_i)$ and $\phi_k^l(n_i)$ are then calculated by minimising a time-domain error criterion. This approach importantly allows shorter analysis frame intervals and therefore greater resolution than conventional FFT peak-picking methods.

Analysis is completed by computing an LPC filter to model the spectrum of each frame along with a parametric envelope to describe the time-domain behaviour of the noise within the frame.

During synthesis the harmonic part is reconstructed by applying

$$h^l(n) = \sum_{k=1}^{K_l} A_k^l(n) \cos[k\theta^l(n) + \phi_k^l(n)] \quad (2.16)$$

Amplitudes and phases are linearly interpolated (where phase is unwrapped by predicting the phase at the centre of the current frame from that at the centre of the last one). In order to reconstruct the stochastic component white noise is first passed through the LPC filter and the output is then high-pass filtered with a cut-off frequency equal to F_{max} . The noise is scaled using the envelope calculated during analysis and added to the harmonic part to complete the synthesis process.

Speech re-synthesised using this procedure is perceptually almost indistinguishable from the original [44] [47].

2.4 The Analysis-by-Synthesis/Overlap-Add Model

Computational complexity in the sinusoidal model is centred on the synthesis stage where the outputs of a number of oscillators are summed to generate speech. In the analysis-by-synthesis/overlap-add model (ABS/OLA) [38] [39] [40] complexity is transferred to the analysis phase where instead of peak-picking, an analysis-by-synthesis approach is used to choose model

parameters. Furthermore, synthesis is efficiently implemented as an Inverse-FFT followed by OLA procedure making the model particularly suitable for real-time TTS synthesis.

Within the model each speech frame is modelled as a sum of constant frequency sinusoids where for the l^{th} frame we have [40]

$$\hat{s}^l(n) = \sum_{k=1}^{K^l} A_k^l(n) \cos[\omega_k^l n + \phi_k^l] \quad (2.17)$$

As each frame is modelled as a standalone entity the parameter tracks of the original sinusoidal model which link one frame to the next are not available here. The re-synthesised speech waveform can be expressed as

$$\hat{s}(n) = \sigma(n) \sum_{l=0}^{L-1} w_s(n - lN_s) \hat{s}_l(n - lN_s) \quad (2.18)$$

where $\hat{s}(n)$ is the re-synthesised speech, L is the number of synthesis frames, N_s is the synthesis frame length, $w_s(n)$ a window over the interval $[-N_s, N_s]$ and $\hat{s}_l(n)$ the l^{th} synthetic speech frame [40]. Lastly, $\sigma(n)$ is a time-domain envelope used for improved modelling of transitions.

Speech, perceptually indistinguishable from the original signal, can be re-synthesised using the ABS/OLA model [40],

2.5 Summary

In this chapter the sinusoidal model and two of its derivatives have been presented in some detail. By parametrically modelling individual frequency tracks, the magnitude-only model of Section 2.1 accurately reproduced the short-time magnitude spectrum of the speech signal. While the re-synthesised speech was highly intelligible, neglect of phase information in the original signal led to waveform dispersion and an unnatural reverberant

quality was introduced. This problem was overcome in Section 2.2 by the addition to the model of a cubic polynomial to simultaneously model phase and frequency tracks within the speech. The result was high quality re-synthesis, perceptually indistinguishable from the original speech.

The currently prevalent harmonic plus noise model was presented in Section 2.3. By modelling deterministic and stochastic components separately, another level of flexibility is added to the sinusoidal model and problems associated with the time-scale expansion of purely sinusoidally coded voiceless regions (discussed in detail in Chapter 3) are alleviated. In informal listening tests Laroche et al. [44] report that speech perceptually almost indistinguishable from the original can be re-synthesised using this model. Similar findings are reported by Stylianou et al. [47].

Lastly, the analysis-by-synthesis/overlap-add model was briefly outlined in Section 2.4. By using an ABS approach to choose model parameters the number of sinusoids required to code a frame is reduced. Furthermore an Inverse-FFT followed by OLA scheme allows for efficient high quality re-synthesis.

Although the sinusoidal model affords excellent re-synthesis quality, a high computational cost is incurred. Each frequency track is individually modelled and during synthesis the contribution of each to the speech signal must be summed. However, it is precisely this fine-grained modelling which provides speech transformation algorithms with a greater control over signal characteristics than other non-parametric approaches e.g. PSOLA [33]. The abilities of a speech coder to, firstly, accurately reproduce speech and, secondly, allow for high quality prosodic manipulation were established as prerequisites for use in a concatenative TTS system in Chapter 1. The sinusoidal model has been shown in this chapter to satisfy the former requirement.

In Chapter 3 its prosodic modification capabilities are explored.

CHAPTER 3

TRANSFORMATIONS USING A SINUSOIDAL MODEL

Speech transformations have a number of important applications. Degraded or noisy speech may be made more intelligible by slowing it down using time-scaling techniques. Time-scale modification may also be used to increase the speech rate in TTS systems thus enabling visually impaired users to rapidly scan text. Finally, as mentioned in Chapter 1 and most importantly for our purposes, both pitch and time-scale modification functions are required in concatenative TTS synthesis where the prosodic characteristics of stored synthesis units often require adjustment prior to concatenation in order to effect the target prosody.

An important characteristic of a speech coder is then the ease with which it facilitates the implementation of the above transformations. Given its explicit and time-dependent functional modelling of individual frequency components within the speech signal the sinusoidal model is well suited to such tasks and indeed offers a greater degree of control over signal characteristics than other simpler non-parametric models like PSOLA [33].

This chapter is organised as follows. Section 3.1 presents Quatieri and McAulay's first attempt at implementing pitch and time-scale modification using a sinusoidal model [55] [56]. Speech transformed using this approach, however, suffers from the same reverberant quality ascribed to speech re-synthesised using the magnitude-only model described in Section 2.1. Shape invariant techniques are introduced in Section 3.2 to solve this problem through the incorporation of "pitch pulse onset times" [37]. In Sections 3.3

and 3.4, respectively, algorithms for performing transformations using HNM [44] [47] and the ABS/OLA model [39] [40] are described. A discussion, in Section 3.5, of the problems associated with each of the models concludes the chapter.

3.1 Early Approach

In order to gain the flexibility required to implement the above transformations Quatieri and McAulay refined their sinusoidal model [36] (discussed in Section 2.2) to incorporate separate models of the glottal excitation and vocal tract [55] [56]. The role of each in the speech production process can be expressed as

$$s(t) = \int_0^t h(t - \tau) e(\tau) d\tau \quad (3.1)$$

where $s(t)$ is the speech waveform, $h(t - \tau)$ is a time-varying filter representing the vocal tract and $e(\tau)$ is the excitation signal [55]. Writing each frame of the excitation signal as a sum of sinusoids gives *

$$e(n) = \sum_{k=1}^K a_k(n) \cos[\Omega_k(n)] \quad (3.2)$$

where K is the number of frequency components in the frame and $a_k(n)$ and $\Omega_k(n)$ denote, respectively, the time-varying amplitude and phase of the k^{th} frequency component. As outlined in Section 2.2 excitation phase is modelled as the integral of the instantaneous frequency plus a starting phase offset ϕ_k .

$$\Omega_k(t) = \int_0^t \omega_k(\sigma) d\sigma + \phi_k \quad (3.3)$$

The behaviour of the time-varying vocal tract filter $H(\omega; n)$ can be expressed in terms of its system amplitude and phase as

$$H(\omega; n) = M(\omega; n) \exp[j\varphi(\omega; n)] \quad (3.4)$$

*Frame superscripts are henceforth omitted for convenience.

since the excitation signal is modelled as the sum of a number of discrete frequency tracks, the system amplitude and phase along each one can be written as

$$\begin{aligned} M_k(n) &= M(\omega_k(n); n) \\ \varphi_k(n) &= \varphi(\omega_k(n); n) \end{aligned} \quad (3.5)$$

Combining the excitation and vocal tract models gives the composite amplitude and phase along each frequency track and speech can be re-synthesised from

$$\hat{s}(n) = \sum_{k=1}^K A_k(n) \cos[\Theta_k(n)] \quad (3.6)$$

where

$$\begin{aligned} A_k(n) &= a_k(n)M_k(n) \\ \Theta_k(n) &= \Omega_k(n) + \varphi_k(n) \end{aligned} \quad (3.7)$$

In Equation (3.7) the composite amplitude along the k^{th} frequency track is the product of the excitation amplitude and the vocal tract system amplitude and the corresponding composite phase is the sum of the excitation and vocal tract system phase.

This approach, while adding to the flexibility of the model, obviously necessitates the separation of excitation and vocal tract contributions to speech production. In [55] Quatieri and McAulay use a technique known as homomorphic deconvolution to estimate vocal tract system amplitude and phase spectra. Excitation amplitude and phase are then easily calculated from Equation (3.7) where composite amplitude and phase values are taken at peaks in the FFT calculated over each speech frame.

During synthesis nearest-neighbour frequency matching is carried out. Excitation amplitudes along with vocal tract system amplitudes and phases

are slowly varying and may be linearly interpolated. Excitation phase must however be unwrapped and interpolated using the cubic polynomial method described in Section 2.2. Speech, re-synthesised from Equation (3.6), is described by Quatieri and McAulay as being nearly indistinguishable from the original [55]. The entire analysis/synthesis scheme is depicted in Figure 5.

3.1.1 Time-Scale Modification

The aim of time-scale modification is to speed up/slow down the rate of articulation while preserving the quality and naturalness of the original speech. Using the model outlined in the last section this can be accomplished by compressing/expanding excitation frequency tracks across the scaled frame interval and updating other model parameters at a faster/slower rate. For a scaling factor ρ (where $\rho > 1$ and $\rho < 1$ correspond respectively to time-scale expansion and compression) Quatieri and McAulay give the following transformation [55]

$$\hat{s}'(n) = \sum_{k=1}^K A'_k(n) \cos[\Theta'_k(n)] \quad (3.8)$$

where

$$A'_k(n) = a_k(\rho^{-1}n)M_k(\rho^{-1}n) \quad (3.9)$$

$$\Theta'_k(n) = \frac{\Omega_k(\rho^{-1}n)}{\rho^{-1}} + \varphi_k(\rho^{-1}n) \quad (3.10)$$

and

$$\Omega_k(\rho^{-1}n) = \int_0^n \omega_k(\rho^{-1}\tau) d\tau + \phi_k \quad (3.11)$$

Amplitudes are interpolated linearly (Equation (3.10)) as is the vocal tract system phase ($\varphi_k(n)$ in Equation (3.10)). The integral of the instantaneous

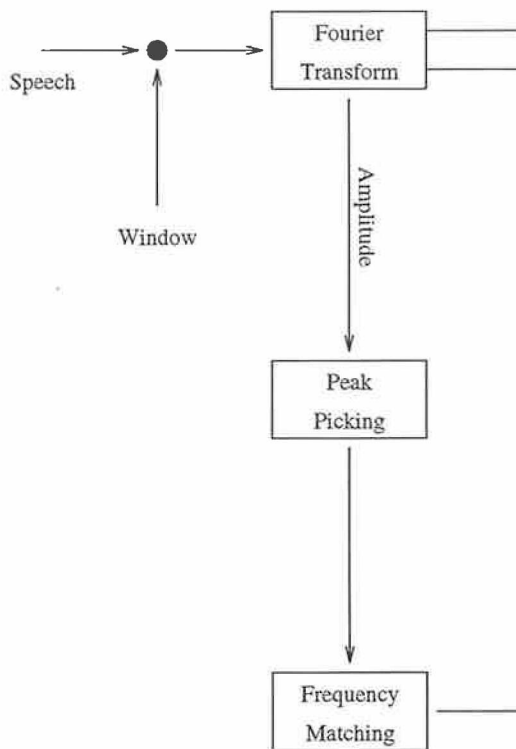
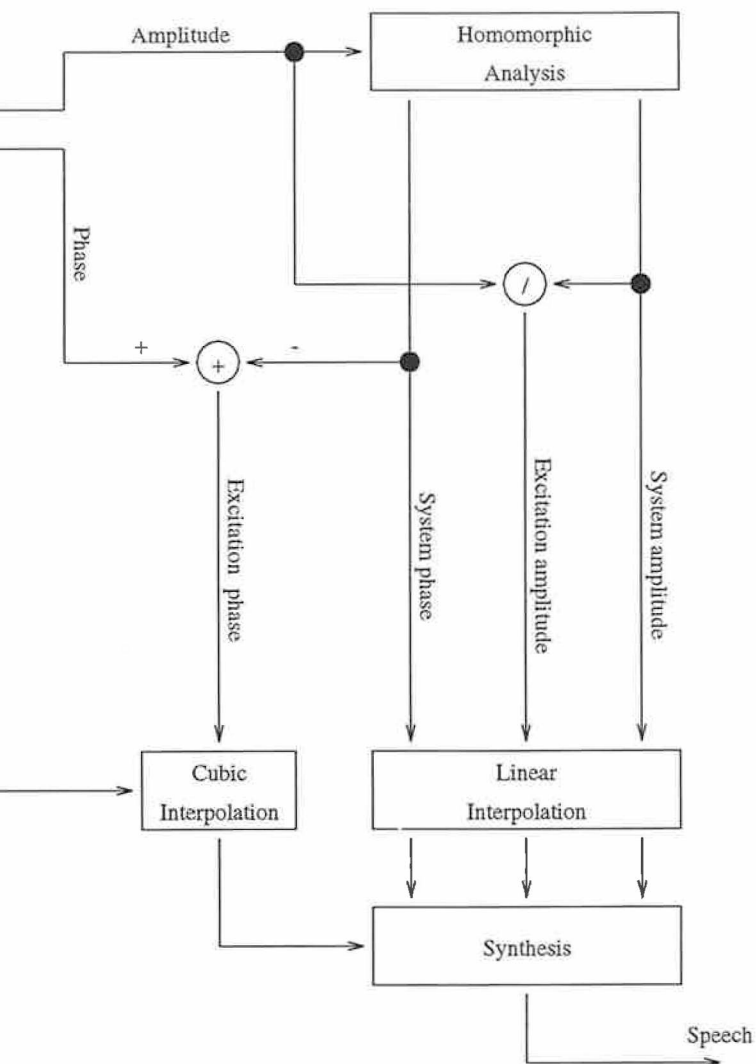


Figure 5 Refined sinusoidal analysis/synthesis model (from [55])



frequency is also interpolated but scaled by $\frac{1}{\rho-1}$ in order to maintain the original frequency ($\Omega_k(n)$ in Equation (3.10)). Quatieri and McAulay tested this approach using time-scaling factors varying from $\rho = 0.5$ to $\rho = 4.0$ and reported generally high quality results [55].

3.1.2 Frequency-Scale and Pitch Modification

Implementing spectral warping is a trivial matter using this composite excitation-vocal tract model. By scaling the excitation phase function by β (where $\beta > 1$ and $\beta < 1$ correspond to frequency scale expansion and compression respectively) and thus reducing or increasing the amount of phase “consumed” by each sinusoid the effect is to scale each frequency (and the overall pitch) by β .

$$\hat{s}'(n) = \sum_{k=1}^K A_k(n) \cos[\beta\Omega_k(n) + \varphi_k(n)] \quad (3.12)$$

As frequency scaling is not accompanied by re-sampling of the vocal tract system amplitude and phase spectra at the new sinusoidal frequencies, the original spectral envelope is stretched/compressed over the modified bandwidth, formants move to new locations, and the resultant speech is distorted. Quatieri and McAulay reported “successful” frequency scaling experiments using factors of $\beta = 0.8$ and $\beta = 1.2$ [55].

In order to change only the pitch of the speech signal and avoid such distortion the vocal tract system amplitude and phase spectra must be re-sampled at each scaled frequency. Excitation amplitudes remain unaltered. Pitch modification can then be achieved using

$$\hat{s}'(n) = \sum_{k=1}^K a_k(n) M'_k(n) \cos[\beta\Omega_k(n) + \varphi'_k(n)] \quad (3.13)$$

where

$$\begin{aligned} M'_k(n) &= M(\beta\omega_k; n) \\ \varphi'_k(n) &= \varphi(\beta\omega_k; n) \end{aligned} \tag{3.14}$$

Quatieri and McAulay reported that smooth and artifact-free speech was obtained when pitch-scaling was carried out using this method [55].

3.1.3 *Summary*

Using the refined approach presented at the beginning of Section 3.1 Quatieri and McAulay succeeded in extending their sinusoidal model to handle both pitch and time-scale modification. Although good results were reported for a range of scaling factors, an unnatural reverberant quality was apparent in the modified speech [37]. As was the case with the magnitude-only model of Section 2.1 this effect is due to a breakdown in phase coherence. During scaling, no attempt is made to retain the original waveform shape in the transformed version. In the following section we present how Quatieri and McAulay refined further the method described here and introduce their “shape invariant” approach to speech transformations using a sinusoidal model [37].

3.2 Shape Invariant Transformations

As mentioned, during speech modification the time- and pitch-scaling algorithms presented in Sections 3.1.1 and 3.1.2 do not take into account phase relations existing between the frequency components of the excitation signal. In order to avoid waveform dispersion and the consequent reverberation, excitation phases must be “locked” together in the modified speech in the same

way as they are in the original. To impose such a structure on excitation phases Quatieri and McAulay added the notion of a “pitch pulse onset time” to their model [37]. At each onset time all excitation frequencies are assumed to be in phase i.e. the phase of each is assumed to be some integer multiple of 2π . Their locations are used as a means of anchoring phases correctly in the new pitch- and time-scales and maintaining the original waveform shape.

3.2.1 Shape Invariant Time-Scale Modification

A pitch pulse occurs where all waves in the excitation signal add coherently. Incorporating this notion into the excitation model gives (in the vicinity of a pitch pulse)

$$e(n) = \sum_{k=1}^K a_k(n) \cos[(n - n_o)\omega_k] \quad (3.15)$$

where n_o is the onset time closest to the centre of the analysis frame. From Equation (3.15) it can be seen that when $n = n_o$ all waves sum coherently and a peak occurs. Excitation phase is thus bound to pitch pulse onset time location. During analysis a set of relative pitch pulse onset times can be generated from the pitch period contour $P(n)$. Actual onset times cannot be reliably calculated and errors in their estimation give rise to objectionable artifacts in the speech [57] [58] [59]. Given the high precision of the pitch detection algorithm [59] used by Quatieri and McAulay relative onset times can be accurately calculated. The human ear is deaf to the linear offset added to phase values estimated in this way. Once a set of onset time locations has been calculated, excitation and vocal tract system phase can be estimated at each analysis interval using the closest onset time, n_o , from

$$\begin{aligned} \Omega_k(n) &= (n - n_o)\omega_k \\ \varphi_k(n) &= \theta_k(n) - \Omega_k(n) \end{aligned} \quad (3.16)$$

where once again composite phase, $\theta_k(n)$, is taken directly from the FFT.

In order to implement time-scale modification it remains to estimate new excitation phases at synthesis frame intervals. A set of onset times along the new time-scale is estimated by pitch period accumulation using the time-scaled pitch period contour $P'(n)$. This process is depicted in Figure 6. Once calculated, the onset time closest to the centre of each synthesis frame, n'_o , is used to give an excitation phase estimate using

$$\Omega'_k(n) = (n - n'_o)\omega_k \quad (3.17)$$

Excitation phases are thus synchronised on pitch pulse onset times consis-

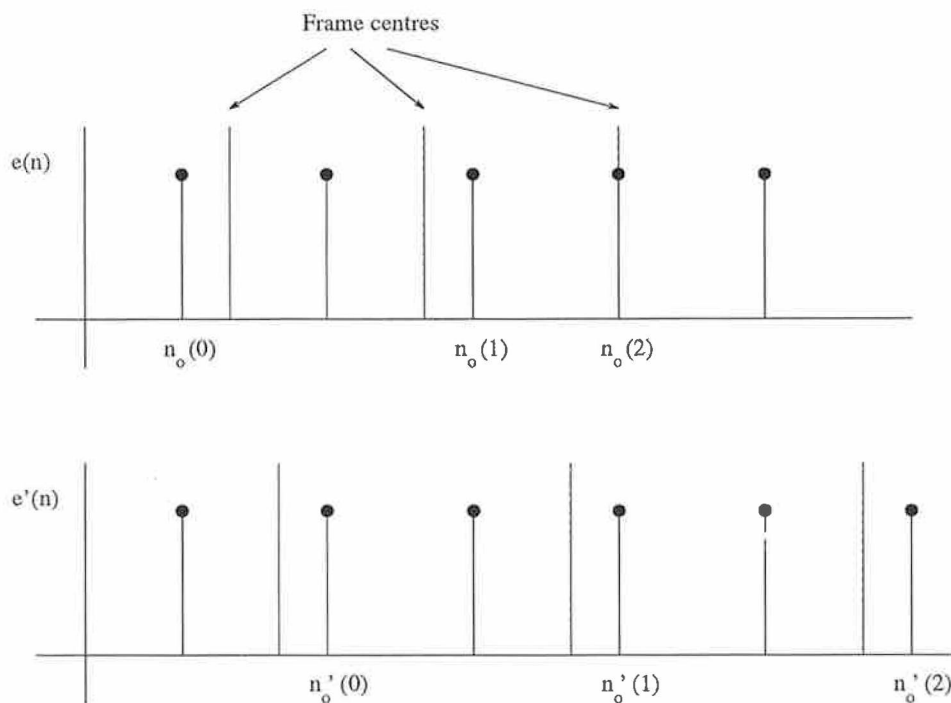


Figure 6 Pitch pulse onset time estimation with $\rho = 2$ (from [37])

tent with the new time-scale. Using Equation (3.18) synthesis is straightforward. Frequency track amplitudes are interpolated linearly as is vocal

tract system phase. Excitation phase and system phase are summed and interpolated using the now familiar cubic polynomial method.

$$\hat{s}'(n) = \sum_{k=1}^K A'_k(n) \cos[(\Omega'_k(n) + \varphi'_k(n))] \quad (3.18)$$

where

$$\begin{aligned} A'_k(n) &= A_k(\rho^{-1}n) \\ \varphi'_k(n) &= \varphi_k(\rho^{-1}n) \end{aligned} \quad (3.19)$$

A fixed rate of time-scale modification has been assumed in the analysis presented above. A time-varying rate can be implemented if a fixed scaling factor is assigned to each frame while allowing factors to vary from one frame to the next. By making the analysis frame interval small enough a continuously varying time-scaling rate can be approximated [37]. Quatieri and McAulay applied a range of both fixed and time-varying scaling factors to a large speech database containing both male and female speech samples and reported generally good quality, natural-sounding results [37]. Moreover, the modified speech retained the shape of the original and was free of reverberation. The complete time-scale modification algorithm is depicted in Figure 7.

Both voiced and voiceless speech are treated in the same way by the sinusoidal model - i.e. both are coded as a set of sinusoids of time-varying amplitude and frequency. As mentioned in Section 2.3, McAulay and Quatieri maintain that noise (in theory consisting of an infinite set of sinusoids) can be adequately modelled in this way provided that the sinusoids used produce a relatively smooth spectrum [36]. This is guaranteed if the amplitudes and frequencies of each component are rapidly varying in a random fashion.

During time-scale modification however, specifically time-scale expansion, the distance between start and target parameters is increased thus lessening

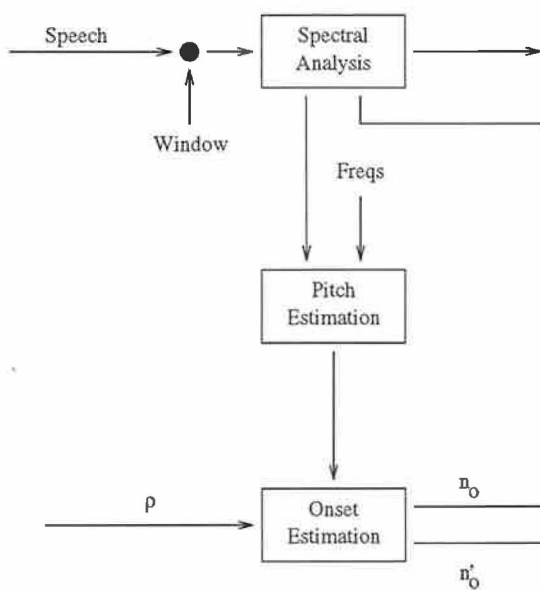
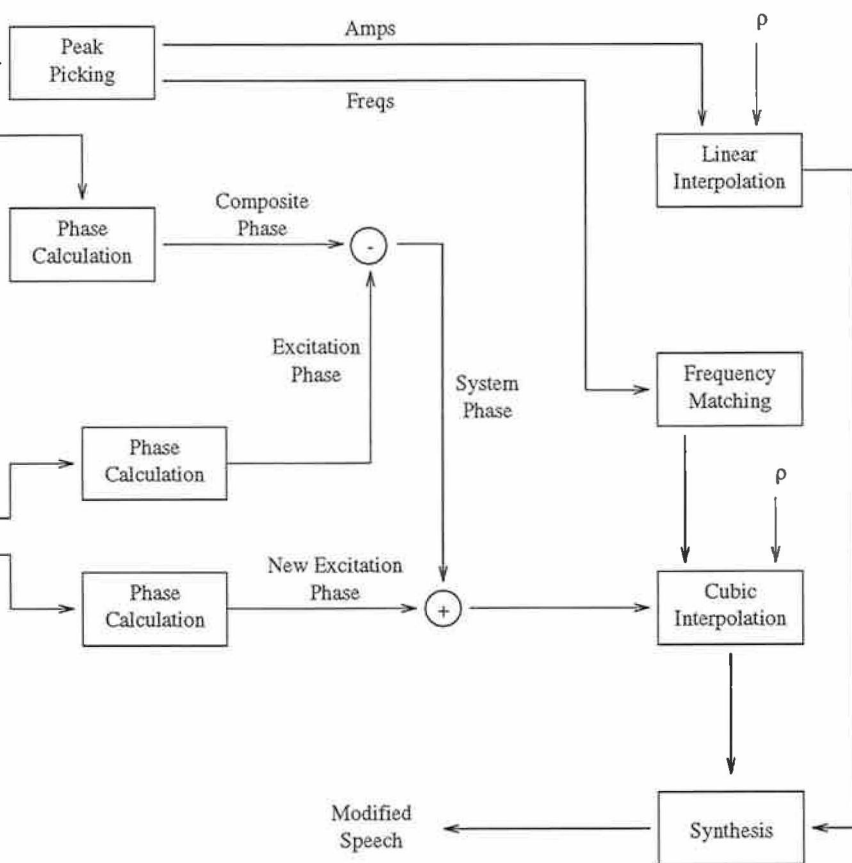


Figure 7 Shape-invariant time-scale modification (from [37])



the rate of variation of voiceless parameter tracks which consequently become smoother. As a result, peaks at individual frequencies become discernible in the spectrum. These peaks add a periodic component to the signal to which the human ear is sensitive with the result that an unnatural “tonal” quality is imparted to the speech.

A simple way of reducing such effects is to use a smaller analysis frame interval e.g. in their experiments Quatieri and McAulay used a 5ms analysis frame interval thus ensuring that, for a time-scale modification factor of 2, no synthesis interval would exceed 10ms which is short enough to guarantee sufficient variation of voiceless parameters [37]. If larger time-scale modification factors are required, shorter frame intervals and phase dithering techniques can be employed to help ensure de-correlation of voiceless frequencies [37].

3.2.2 *Shape Invariant Pitch Modification*

During pitch modification a new set of pitch pulse onset times, relative to the synthesis axis, must again be computed using the pitch-scaled pitch period function $P'(n)$. This procedure is illustrated in Figure 8. Using the onset time closest to the centre of each synthesis frame, n'_o the new excitation phase is calculated from

$$\Omega'_k(n) = (n - n'_o)\beta\omega_k \quad (3.20)$$

where β is the pitch-scale modification factor. In order to maintain the overall spectral shape of the original speech, vocal tract amplitude and phase responses must be estimated at each of the new frequencies $\beta\omega_k$ i.e.

$$\begin{aligned} M'_k(n) &= M(\beta\omega_k; n) \\ \varphi'_k(n) &= \varphi(\beta\omega_k; n) \end{aligned} \quad (3.21)$$

The SEEVOC algorithm [60] is used to generate a smooth spectral envelope, $\hat{M}(\omega; n)$, from vocal tract system amplitudes estimated by homomorphic deconvolution. $\hat{M}(\omega; n)$ is then re-sampled at the new frequency values to give $M'_k(n)$. Vocal tract system phases (estimated from the original onset times using the procedure described in Section 3.2.1) are interpolated to give $\varphi'_k(n)$. Synthesis is carried out using Equation (3.22) where vocal tract and excitation amplitudes are multiplied and linearly interpolated. Vocal tract system phase is added to the excitation phase and interpolated using the cubic polynomial method.

$$s'(n) = \sum_{k=1}^L a_k(n) M'_k(n) \cos[\Omega'_k(n) + \varphi'_k(n)] \quad (3.22)$$

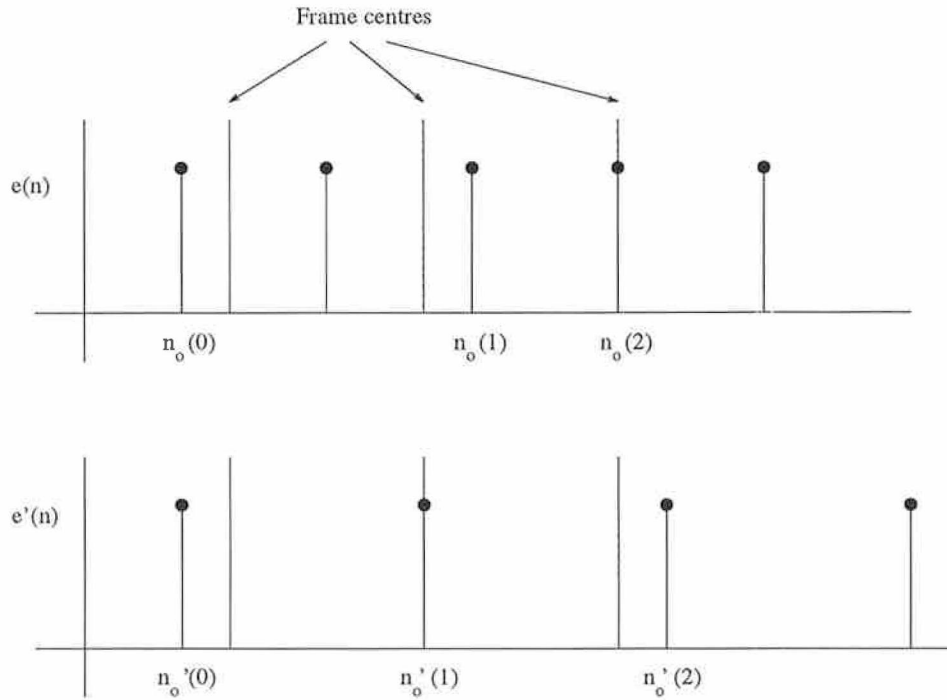


Figure 8 Pitch pulse onset time estimation with $\beta = 2$ (from [37])

Time-varying pitch-scaling is again implemented by assigning frame varying scaling factors. Quatieri and McAulay applied their pitch modification algorithm to a large speech database containing both male and female speech samples. For pitch modification factors in the range 0.5-2.0 they reported results being generally smooth and free of artifacts [37]. The shape of the original speech was also well preserved. However, for pitch modification factors greater than 20% a hoarseness was present in the reconstructed speech. This effect Quatieri and McAulay attributed to inaccurate system phase estimation and scaling of voiceless frequencies [37].

3.2.3 *Summary*

By incorporating pitch pulse onset times into their model, Quatieri and McAulay removed the reverberant quality associated with pitch- and time-scaled speech using the approach in Section 3.1. Onset times are used in the modified speech to impose coherence on excitation phases thus guaranteeing waveform shape retention.

Finally, Pollard et al. [61] have pointed out that synchronising excitation phases on a single pitch pulse onset time per synthesis frame, while adequate for small degrees of modification, breaks down when larger modification factors are used. As the distance between start and target phases is increased (as is the case during time-scale expansion) the constraint on intervening phase values is weakened and waveform dispersion can occur, reintroducing reverberation. To counter these effects Pollard et al. synchronise excitation frequencies on every pitch pulse onset time in a frame thus imposing more rigidity on phase structure and limiting the ability of phase values to “wander about” [61]. Doing so necessarily entails a much increased synthesis complexity but Pollard et al. report successful time-scaling experiments using factors

as high as 6 [61].

3.3 The Harmonic Plus Noise Model

The harmonic plus noise model (HNM) was presented in detail in Section 2.3. In HNM the speech signal is viewed as containing both a deterministic and a stochastic component. Each is modelled separately. A set of harmonics describes the deterministic part while an LPC filter along with a time-domain amplitude envelope account for the noise.

$$\hat{s}(n) = \sum_{k=1}^{K(n)} A_k(n) \cos(k\theta(n) + \phi_k(n)) + e(n) \quad (3.23)$$

where

$$\theta^l(n) = \int_{-\infty}^n \omega_0^l(t) dt \quad (3.24)$$

As the analysis is pitch-synchronous, i.e. analysis windows, each of which is two pitch periods long, are placed at pitch periodic intervals across the speech waveform, the model lends itself to a PSOLA-type [33] pitch and time-scale modification scheme [44] [47]. Although the analysis is pitch-synchronous, analysis windows are not centred on pitch marks. Instead, a centre of gravity shifting technique is used to move (by adding a linear term to the phase of each harmonic) the point of glottal closure to the centre of each frame [62]. This process is described in more detail in Chapter 5.

During reconstruction, a set of synthesis time instants (pitch marks) is first calculated from the pitch and time-scaled pitch contour $P'(n)$. A mapping is then established from synthesis to analysis pitch marks. An example mapping, implementing time-scale expansion, is given in Figure 9. The upper part of the figure illustrates the time-warping function, i.e. it indicates by how much each segment of the original speech waveform is to be expanded,

while the lower part shows which short-time analysis signals are to be used in the OLA procedure in order to implement the desired time-scaling. In this example, two of the analysis signals are used twice during synthesis.

During pitch-scaling, amplitudes and phases are re-sampled at the new harmonic frequencies. Quatieri and McAulay's approach to pitch modification leads to a loss of high frequency information when pitch-lowering is involved - all frequencies above $\frac{\beta F_s}{2}$ are lost where β is the pitch-modification factor and F_s is the sampling frequency are lost. This does not occur in HNM where the entire frequency band is re-sampled during pitch-scaling.

The scaled speech is then synthesised by either interpolating the parameters of Equation (3.23) from one synthesis instant to the next or individually synthesising each frame and applying an OLA procedure. Results using either approach, according to Stylianou (personal communication), are of similar quality.

Time-scaling of voiceless sections does not pose the same problems (outlined in Section 3.2.1) as it did for the original sinusoidal model. In HNM, white noise is filtered through the LPC filter derived during analysis and then high-pass filtered with a cut-off frequency equal to F_{max} . The output is finally scaled with an amplitude envelope to account for the particular time-domain behaviour of the noise. This procedure maintains randomness over any synthesis interval length and avoids tonal artifacts. However, some researchers have reported that using such models can lead to a lack of "perceptual fusion": noisy and deterministic components are perceived separately, as if generated by different sources.

Laroche et al. [44] successfully time-scaled a voiced fricative by a factor of 2 using this method. The reconstructed speech was free of the buzziness associated with PSOLA [33] methods. Stylianou et al. [47], on the basis of

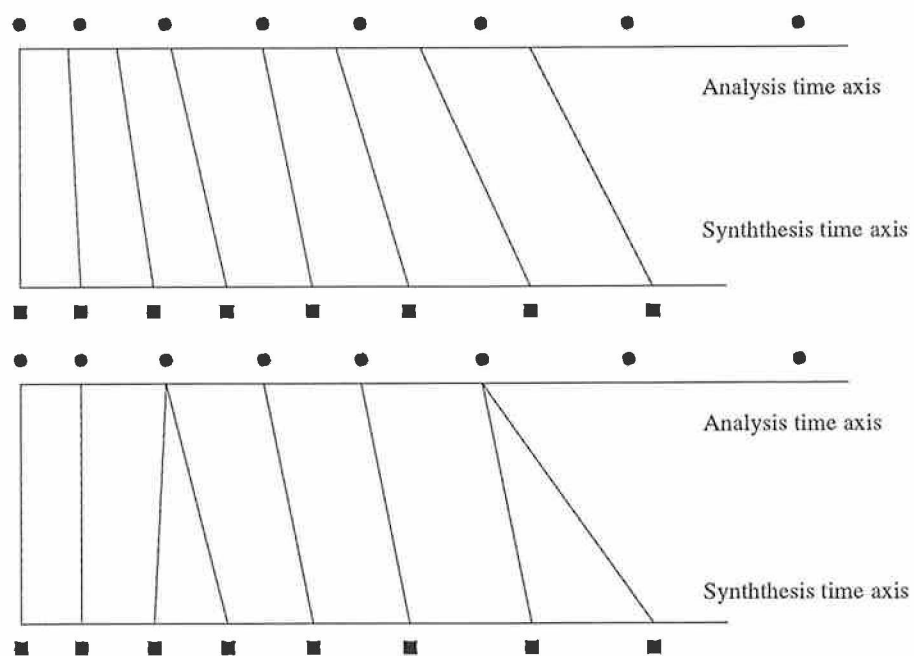


Figure 9 PSOLA algorithm (from [31])

informal listening tests, report high quality results for even large time-scaling and moderate pitch-scaling factors (0.5-1.7).

3.4 The Analysis-by-Synthesis/Overlap-Add Model

Using the ABS/OLA approach [38] [39] [40], each speech frame is modelled as a sum of sinusoids of constant amplitude and frequency where

$$\hat{s}(n) = \sum_{k=1}^K A_k \cos(\omega_k n + \phi_k) \quad (3.25)$$

Writing Equation (3.25) as a quasi-harmonic model (i.e. where each frequency is represented as an underlying harmonic plus a frequency differential component) gives [40]

$$\hat{s}(n) = \sum_{k=1}^K A_k \cos((k\omega_0 + \Delta_k)n + \phi_k) \quad (3.26)$$

Straightforward time-scaling of Equation (3.26) (i.e. simply stretching each frequency track over the new synthesis interval) leads to phase coherence breakdown and waveform dispersion as the model is not a purely harmonic but a quasi-harmonic one and phase offsets do not repeat periodically [40]. George [40] overcomes this problem by treating the frequency differential component as a phase modulating term which serves to modify speech waveform shape across the synthesis interval. This idea is expressed in Equation (3.27).

$$\hat{s}(n) = \Re \sum_{k=1}^K e^{j(\Delta_k n)} (A_k e^{j(k\omega_0 n + \phi_k)}) \quad (3.27)$$

By scaling this term, $\Delta_k n$, by ρ where ρ is the time-scale modification factor, its effect can be kept consistent with the new synthesis interval length. Time-scale modification can then be carried out using

$$\hat{s}'(n) = \sum_{k=1}^K A_k \cos(k\omega_0 n + \frac{\Delta_k n}{\rho} + \phi_k) \quad (3.28)$$

Each frequency is modified during time-scaling according to Equation (3.29) from which it can be seen that as ρ increases the quasi-harmonic set of frequencies tends towards a harmonic series as $\frac{\Delta_k}{\rho}$ eventually disappears [40].

$$\omega_k = k\omega_0 + \frac{\Delta_k}{\rho} \quad (3.29)$$

The problems outlined in Section 3.2.1 relating to time-scale expansion of voiceless regions are also pertinent here. Rather than adopt a stochastic/deterministic approach, Macon retains the purely sinusoidal ABS/OLA model but introduces a phase dithering technique in order to prevent the perception of periodicities in the signal when expanding voiceless sections [41]. One reason for adopting this approach is, as mentioned in Section 3.3, the lack of “perceptual fusion” often associated with synthetic speech produced by hybrid models. Macon’s approach to preserving perceptual randomness, within the OLA framework, is to split each synthesis frame up into a number of sub-frames and randomise sub-frame phase offsets [41]. The phase offset of the k^{th} sinusoidal component for sub-frame m is given by

$$\phi_{k,m} = \phi_k + V_k \psi_{k,m} \quad (3.30)$$

where ϕ_k is the original phase offset, V_k is a frequency dependent random number in the range $[0, 1]$ and $\psi_{k,m}$ is a random phase variable in the range $[-\pi, \pi]$. If $V_k = 0$ for all k then the original phase offset is used. If, however, $V_k = 1$ for all k then phase offsets will vary randomly from one sub-frame to the next. This procedure can be shown to be equivalent to widening the bandwidth of each sinusoidal component thus smoothing the spectrum [41].

The ABS/OLA model has been used to time-scale speech (both male and female) and the modified speech was reported to be of high quality and free of artifacts [40].

As mentioned, pitch-lowering leads to a loss of high frequency components using Quatieri and McAulay's approach. Furthermore, as each frequency is scaled by β , noisy sinusoidal components may be transported to formant frequency locations giving rise to artifacts in the modified speech. To avoid such problems George [39] [40] introduces the notion of "phasor interpolation" whereby the excitation's amplitude and phase spectra are interpolated and re-sampled across the entire frequency band.

Pitch-scaling using the ABS/OLA was tested on a number of utterances (produced by both male and female speakers) and was found to produce very satisfactory results [40]. Macon [41], however, reports a "pulse" structure being imparted to the speech signal when pitch-lowering occurs using this method, a result of amplitude modulation due to windowing. This effect is addressed and compensated for by Macon [41] but at the expense of adding a reverberant quality to the pitch-modified speech.

Furthermore, inaccurate pitch pulse onset time estimation (onset time locations are required in order to remove linear phase trend during phasor interpolation) leads to artifacts in the modified speech. A robust algorithm for their calculation is presented by Macon in [41].

3.5 Discussion

In Section 3.1 Quatieri and McAulay's first attempt at using the sinusoidal model to implement pitch and time-scale modification was presented [55] [56]. Due to its simplistic handling of phase, speech transformed using this approach suffered from the same reverberant quality as that produced by the magnitude-only model of Section 2.1. This problem was solved, however, in Section 3.2 through the introduction of "pitch pulse onset times" [37]

which are used to lock excitation frequencies together in a manner consistent with the local pitch period. This “shape-invariant” approach eliminates reverberation from the transformed speech.

Time-scaling of voiceless segments using this approach can be problematic. Specifically, during its time-scale expansion, if the random nature of noise is not preserved it takes on a tonal character. Quatieri and McAulay suggest a number of solutions, namely, the use of shorter analysis frame intervals and/or phase and frequency dithering techniques [37].

The pitch and time-scaling techniques described above are quite complex with the inclusion of pitch pulse onset times seriously detracting from the simplicity of the original sinusoidal model. Furthermore, as pointed out by Pollard et al. [61] using only a single onset time per frame to synchronise phases may not be adequate for accurate shape preservation. As the distance between start and target parameters is increased, more freedom is allowed to phases which can lead to waveform dispersion. Pollard et al. ’s solution [61], which comes with a high computational cost, is to impose more control on phases by synchronising excitation phase on all pitch pulse onset times rather than a single one per synthesis frame. However, both approaches rely on accurate pitch pulse onset time estimation, by no means a trivial task. Erroneous onset time estimates introduce distortion into the modified speech. A final problem is the “hoarseness” Quatieri and McAulay have reported in pitch-modified speech when large scaling factors are used. This they have partly attributed to inaccurate system phase estimation [37]

In Section 3.3 the HNM’s [44] [47] pitch-synchronous analysis was shown to lend itself to a PSOLA-type synthesis scheme. Each short-time analysis signal is coded as a sum of deterministic and stochastic contributions. A mapping is generated from analysis to synthesis signals and an OLA tech-

nique applied to reconstruct the speech. This “parametric PSOLA” approach removes many of the problems associated with the original PSOLA and the separate modelling of the stochastic component eliminates the risk of tonal noise. Laroche et al. [44] and Stylianou et al. [47] report excellent quality results using this approach.

However, modelling the stochastic and deterministic speech components separately, while eliminating the problem of tonal noise can result in a lack of perceptual fusion in the reconstructed speech. This occurs when the stochastic component is not accurately synchronised with the point of glottal closure. Stylianou [62] claims to have solved this problem through the use of a centre of gravity shifting technique which ensures that the point of glottal closure is always at the centre of each frame (thus rendering synchronisation a simple task). Nonetheless, the duplication and deletion of frames, inherent in all PSOLA-based methods, is intuitively unappealing.

In Section 3.4 we saw how the ABS/OLA model [39] [40] was applied to implementing speech transformations. Time-scaling was successfully handled by regarding the model as a quasi-harmonic one where each harmonic was accompanied by a phase modulating term. Harmonics are allowed to evolve in time while scaling the modulating term to keep its effect consistent with the new synthesis interval. A phase dithering technique is applied during voiceless time-scaling to preserve randomness. During pitch-modification “phasor interpolation” is used to re-sample the excitation spectrum across the entire frequency band. Thus, no information is lost during pitch-lowering.

In order to produce high quality results phasor interpolation requires highly accurate onset time estimation. Errors introduce artifacts into the modified speech. Furthermore, on pitch lowering the resulting speech however exhibits a “pulse” appearance, an effect compensated for by Macon [41] at

the expense of introducing some reverberance.

In the next chapter a new and conceptually simple approach to high quality pitch and time-scaling of speech is presented. Based on a harmonic model the method requires no pitch pulse onset time estimation nor is the analysis pitch-synchronous. The harmonic coding of each frame is exploited to ensure phase coherence in a simple and efficient way. furthermore, to each frame is assigned a pitch and time-scale modification factor. There is no duplication or deletion of frames: every frame is used exactly once during reconstruction. A frequency dithering technique is used to maintain randomness over time-scale expanded voiceless regions. Experimental results (presented in detail in Chapter 6) have shown our approach to be capable of very high quality speech transformation.

CHAPTER 4

TRANSFORMATIONS USING A HARMONIC MODEL

In this chapter a new and conceptually simple approach to pitch and time-scale modification of speech is presented. When implementing speech transformations pitch pulse onset times play a crucial role in both Quatieri and McAulay's shape invariant model [37] (Section 3.2) and the ABS/OLA model [39] [40] (Section 3.4). The harmonic based approach put forward here, however, does not rely on onset times to maintain phase coherence. Instead, waveform shape is preserved during transformations by exploiting the harmonic relationship existing between the sinusoids used to code each frame to cause them to be in phase at synthesis frame intervals. Furthermore, our modification algorithms are not PSOLA based and therefore unlike in HNM [44] [47] (Section 3.3) analysis need not be pitch synchronous and the duplication/deletion of frames during scaling is avoided. Lastly, time-scale expansion of voiceless regions is handled not through the use of a hybrid model but by widening the bandwidth of "noisy" sinusoids, thus smoothing the spectrum and alleviating the problem of tonal artifacts. Importantly, our approach allows for a straightforward implementation of joint pitch and time-scale modification.

A formal evaluation of the results produced using the method presented in this chapter, which were found to compare favourably with those generated using other frequency-domain approaches including HNM, is presented in Chapter 6.

The chapter is organised as follows. Section 4.1 briefly outlines the anal-

ysis phase of our model which is similar to that used in the sinusoidal model [36] (Section 2.1) except that harmonics, as opposed to peaks, are used to code voiced frames. Section 4.2 details how, using this new approach, firstly voiced and secondly voiceless speech can be time-scaled. The algorithm is extended to handle pitch modification in Section 4.3. Joint pitch and time-scale modifications are implemented in Section 4.4 and the chapter concludes with a discussion in Section 4.5.

4.1 Analysis

Pitch analysis is carried out on the speech signal using Entropic's pitch detection software*. The resulting pitch contour, after smoothing, is used to assign an F0 estimate to each frame (zero if voiceless). Over voiced regions the length of each frame is typically three times the local pitch period. Over voiceless regions frames are 20ms long. A constant frame interval of 10ms is used throughout the analysis phase. A Hanning window is applied to each frame and its FFT calculated. Over voiced frames the amplitudes and phases of sinusoids at harmonic frequencies are coded. Peak picking applies over voiceless frames.

4.2 Time-Scale Modification

Due to the differences in the transformation techniques employed, time-scaling of voiced and voiceless speech are treated separately. Time-scale modification of voiced speech is first presented.

*get.f0 Copyright Entropic Research Laboratory, Inc. 5/24/93.

4.2.1 Voiced Speech

If their frequency is kept constant, the phases of the harmonics used to code each voiced frame repeat periodically every $\frac{2\pi}{\omega_0}$ s where ω_0 is the fundamental frequency expressed in $rad\ s^{-1}$. Each parameter set (i.e. the amplitudes, phases and frequencies at the centre of each analysis frame) can therefore be viewed as defining a periodic waveform. For any adjustment factor δ a new set of “valid” (where valid means being in phase) phases can be calculated from

$$\psi'_k = \psi_k + \omega_k \delta \quad (4.1)$$

where ψ'_k is the new and ψ_k the original phase of the k^{th} sinusoid with frequency ω_k . After time-scale modification, harmonics should be in phase at each synthesis frame interval i.e. their new and original phases should be related by Equation (4.1). Thus, the task during time-scaling is to estimate the factor δ for each frame, from which a new set of phases at each synthesis frame interval can be calculated. Equipped with phase information consistent with the new time-scale, synthesis is straightforward and is carried out as in Section 2.2. A procedure for estimating δ is presented below.

After nearest neighbour matching (over voiced frames this simplifies to matching corresponding harmonics), has been carried out the frequency track connecting the fundamental of frame l with that of frame $l + 1$ is computed as in Section 2.2 and may be written as

$$\dot{\tilde{\theta}}_0(n) = \gamma + 2\alpha n + 3\beta n^2 \quad (4.2)$$

Time-scaling Equation (4.2) is straightforward. For a given time-scaling factor, ρ , a new target phase, $\psi_0^{l+1'}$, must be determined. Let the new time-

scaled frequency function be

$$\dot{\tilde{\theta}}'_0(n) = \dot{\tilde{\theta}}'_0\left(\frac{n}{\rho}\right) \quad (4.3)$$

The new (unwrapped) target phase, $\psi_0^{l+1'}$, is found by integrating Equation (4.3) over the time interval ρS (where S is the analysis frame interval) and adding back the start phase ψ_0^l ,

$$\int_0^{\rho S} \dot{\tilde{\theta}}'_0(n) dn + \psi_0^l = \rho S (\gamma + \alpha S + \beta S^2) + \psi_0^l \quad (4.4)$$

By evaluating Equation (4.4) modulo 2π , $\psi_0^{l+1'}$ is determined. The model (for F0) is completed by solving for α and β , again, as outlined in Section 2.2.

Applying the same procedure to each remaining matched pair of harmonics will, however, lead to a breakdown in phase coherence after several frames as waves gradually move out of phase. To overcome this, and to keep waves in phase, δ is calculated from Equation (4.1) as

$$\delta = \frac{\psi_0^{l+1'} - \psi_0^{l+1}}{\omega_0^{l+1}} \quad (4.5)$$

δ simply represents the linear phase shift from the fundamental's old to its new target phase value. Once δ has been determined, all new target phases, $\psi_k^{l+1'}$, are calculated from Equation (4.1). Cubic phase interpolation functions may then be calculated for each sinusoid using the method outlined in Section 2.2. Re-synthesis of time-scaled speech is then carried out using Equation 4.6.

$$\hat{s}^l(n) = \sum_{k=1}^{K^l} A_k^l(n) \cos[\tilde{\theta}_k^l(n)] \quad (4.6)$$

It is necessary to keep track of previous phase adjustments when moving from one frame to the next. This is handled by Δ which must be applied,

along with δ , to target phases thus compensating for phase adjustments in previous frames. The complete time-scaling algorithm is presented in Figure 13. It should be noted that this approach is slightly different to that presented in [63] where the difference between the time-scaled and original frequency tracks was minimised (see Section 4.2.2 for an explanation of why this approach was adopted). Here, in the interests of efficiency, the original frequency track is not computed.

Some example waveforms, extracted from speech time-scaled using this method, are given in Figures 10, 11 and 12. Results were found to be of high quality and as can be seen in the figures the shape of the original is well preserved in the modified speech. As mentioned earlier, results are evaluated formally in Chapter 6.

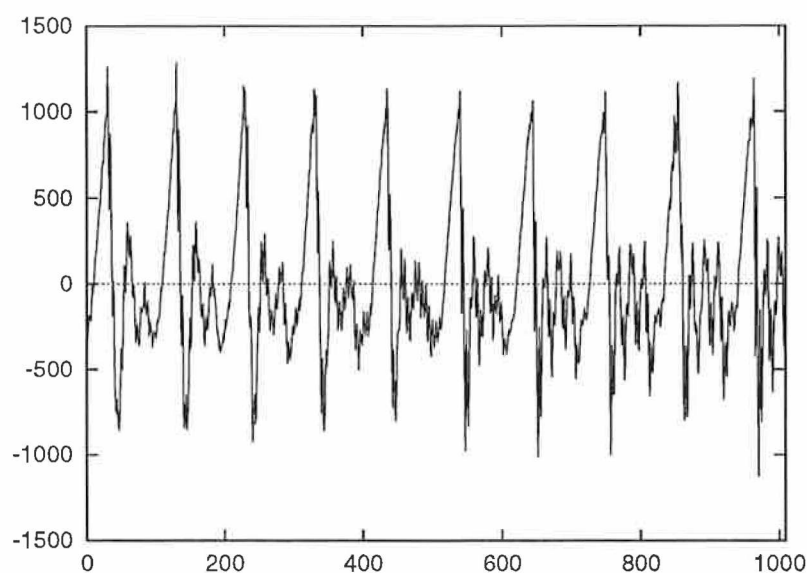


Figure 10 Original speech, $\rho = 1$

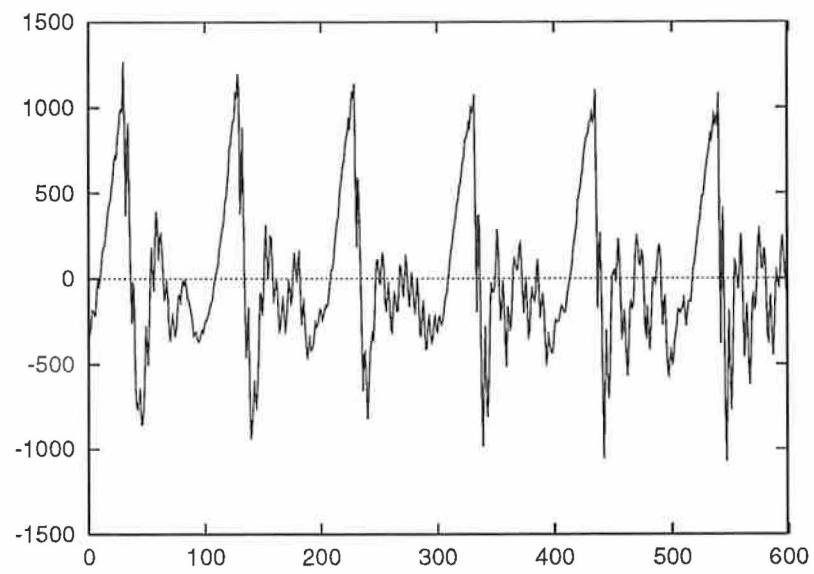


Figure 11 Time-scaled speech, $\rho = 0.6$

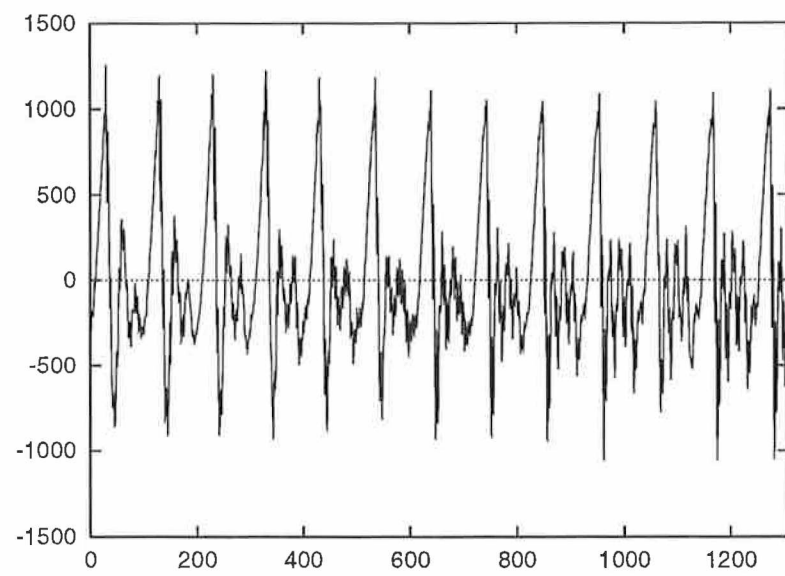


Figure 12 Time-scaled speech, $\rho = 1.3$

```

 $\Delta = 0$ 
 $\delta = 0$ 
For each Frame  $l$ 
  Begin
     $\Delta = \Delta + \delta$ 
    For  $\omega_0$ 
      Begin
        Adjust  $\psi_0^{l+1}$  by  $\Delta$ 
        Compute frequency track  $\tilde{\theta}_0(n)$ 
        Compute new frequency track  $\tilde{\theta}'_0(n)$ 
        Solve for  $\psi_0^{l+1'}$ 
        Solve for  $\delta$ 
        Compute phase function  $\theta_0^l(n)$ 
      End
    End
    For  $\omega_k$  where  $k \neq 0$ 
      Begin
        Adjust  $\psi_k^{l+1}$  by  $\delta + \Delta$ 
        Compute phase function  $\theta_k^l(n)$ 
      End
    End
  End
End

```

Figure 13 Time-scaling algorithm

4.2.2 *Voiceless Speech*

As mentioned in the preceding section an attempt was made in [63] to minimise the difference between the original and time-scaled frequency tracks. Such an approach, it was thought, would help to preserve the random nature of frequency tracks in voiceless regions thus avoiding the need for phase and frequency dithering or hybrid modelling and providing a unified treatment of voiced and voiceless speech during time-scale modification. Using this approach, as opposed to computing the smoothest frequency track, meant slightly larger scaling factors could be accommodated before tonal artifacts were introduced. The improvement, however, was deemed insufficient to outweigh the extra computational cost incurred.

For this reason phase and frequency dithering techniques, to be applied over voiceless speech during time-scale expansion, were implemented. Initially, two simple methods of increasing randomness in voiceless regions were incorporated into the model:

- Upon birth or death of a sinusoid in a voiceless frame a random start or target phase is assigned.
- Upon birth or death of a sinusoid in a voiceless frame a random (but within a specified bandwidth) start or target frequency is assigned.

These simple procedures can be combined, if necessary, with shorter analysis frame intervals to handle most time-scale expansion requirements. However, for larger time-scale expansion factors these measures may not be enough to prevent tonality. In such cases the bandwidth of “noisy” sinusoids is widened thereby smoothing the spectrum and helping to preserve perceptual randomness. This procedure is outlined below.

As each frequency track is modeled with a parabola its bandwidth is necessarily constrained to lie either above or below the line L (see Figure 14) connecting the start and target frequencies. In order to increase, in this case double, the bandwidth of each frequency track, it is simply reflected through the line L to give an auxiliary track. Amplitude interpolation must be adapted to take the existence of this new track into account. These ideas are illustrated in Figure 14.

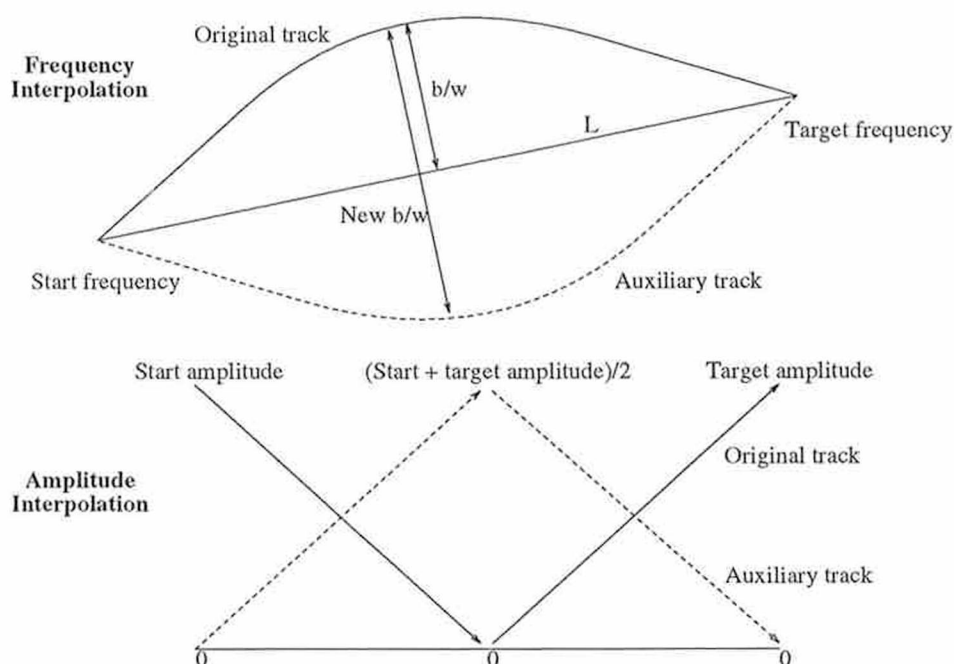


Figure 14 Bandwidth expansion in voiceless speech

During time-scale expansion the smoothest track connecting the start and target phases of each sinusoid, across the time-scaled analysis frame interval, is calculated as in Section 2.2. Each function can be written as

$$\theta(n) = \zeta + \gamma n + \alpha n^2 + \beta n^3 \quad (4.7)$$

The bandwidth of the corresponding frequency track is given by

$$Bw = \sqrt{\frac{\alpha^2(-\alpha S - \omega^l + \omega^{l+1})^2}{9\beta^2(S^2 + (\omega^l - \omega^{l+1})^2)}} \quad (4.8)$$

The phase-unwrapping parameter M , chosen (again as outlined in Section 2.2) such that the smoothest frequency track is obtained, is incremented and α and β re-calculated until $Bw \geq 100\text{Hz}$. This ensures that after reflection the combined bandwidth of both sinusoids is at least 200Hz.

The phase interpolating function of the auxiliary frequency track, obtained by reflecting the original in L , can be shown to be given by

$$\theta(n) = \zeta + \gamma n + \alpha' n^2 + \beta' n^3 \quad (4.9)$$

where

$$\begin{aligned} \alpha' &= -5\alpha + 3d - 6\beta S \\ \beta' &= \frac{4\alpha - 2d + 5\beta S}{S} \end{aligned} \quad (4.10)$$

$$d = \frac{\omega^{l+1} - \omega^l}{S} \quad (4.11)$$

Each sinusoid is effectively split in two and both parts synthesised separately. Amplitude is interpolated linearly as depicted graphically in Figure 14. Using this approach the tonal quality associated with time-scale expanded voiceless speech is eliminated even for large scaling factors (see Chapter 6 for a more detailed analysis of results).

4.3 Pitch Modification

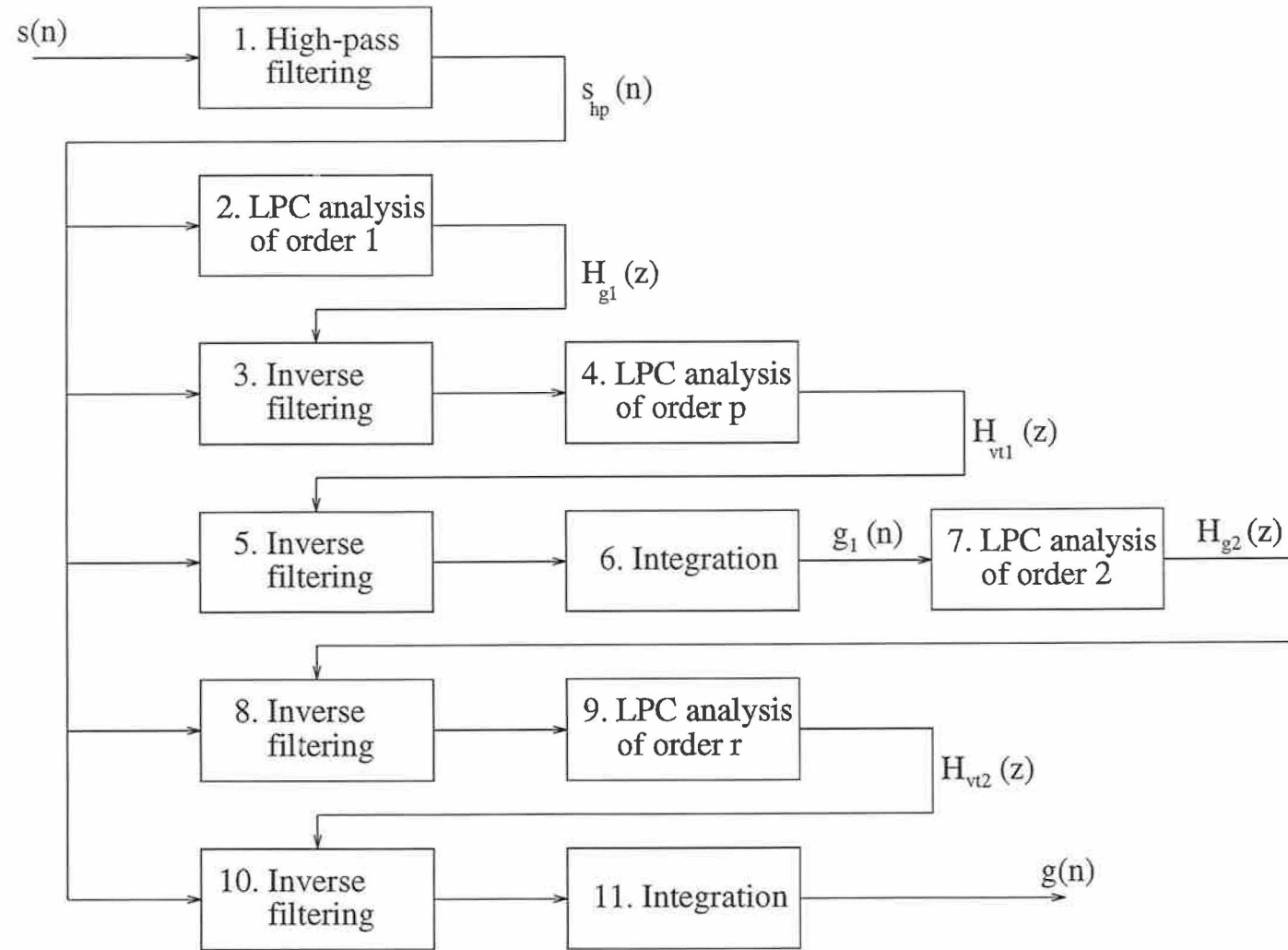
In order to perform pitch modification it is necessary to separate vocal tract and excitation contributions to the speech production process. Here,

a simple LPC-based inverse filtering technique (IAIF: Iterative Adaptive Inverse Filtering) [65] is applied to the speech signal to yield a glottal excitation estimate which is then sinusoidally coded using the approach outlined in Section 4.1. The inverse filtering algorithm is presented in Figure 15 and the various blocks within are explained below[†].

1. The speech is first high-pass filtered to remove “undesirable” fluctuations from the final glottal wave estimate. In [65] a linear phase FIR filter is used with a cut-off frequency of 20Hz.
2. The glottal wave contribution is firstly approximated by $H_{g1}(z)$, an LPC filter of order 1.
3. The initial glottal contribution estimate is eliminated by inverse filtering $s_{hp}(n)$ through $H_{g1}(z)$.
4. An initial vocal tract transfer function $H_{vt1}(z)$ is estimated by applying LPC analysis of order p to the output of block 3.
5. The latter vocal tract contribution is eliminated from $s_{hp}(n)$ by inverse filtering through $H_{vt1}(z)$.
6. A first glottal wave estimate $g_1(n)$ is generated by integrating the output of block 5.
7. A new glottal contribution model $H_{g2}(z)$ is calculated by applying LPC analysis of order 2 to $g_1(n)$.
8. The contribution of $H_{g2}(z)$ to $s_{hp}(n)$ is removed through inverse filtering.

[†]In our implementation of the IAIF algorithm, given that the speech under analysis was sampled at a rate of 16kHz, values of $p = 18$ and $r = 20$ were used.

Figure 15 Iterative adaptive inverse filtering (IAIF) algorithm (from [65])



9. The final vocal tract model $H_{vt2}(z)$ is obtained by applying LPC analysis of order r to the output of block 8.
10. Vocal tract effects are eliminated from $s_{hp}(n)$ by inverse filtering with $H_{vt2}(z)$.
11. The final glottal wave estimate $g(n)$ is obtained by integration to remove lip radiation effects.

Assuming sinusoidal analysis has been carried out on the glottal wave estimate, the frequency track connecting the fundamental of frame l with that of frame $l+1$ is given by

$$\ddot{\theta}_0(n) = \gamma + 2\alpha n + 3\beta n^2 \quad (4.12)$$

Pitch-scaling Equation (4.12) is quite simple. Let λ^l and λ^{l+1} be the pitch modification factors associated with frames l and $l+1$ respectively. Interpolating linearly, the modification factor across the frame is given by

$$\lambda(n) = \lambda^l + \frac{\lambda^{l+1} - \lambda^l}{S}n \quad (4.13)$$

where S is the analysis frame interval. The pitch-scaled fundamental can then be written as

$$\ddot{\theta}'_0(n) = \ddot{\theta}_0(n)\lambda(n) \quad (4.14)$$

The new (unwrapped) target phase, $\psi_0^{l+1'}$, is found by integrating Equation (4.14) over S and adding back the start phase, ψ_0^l .

$$\begin{aligned} \int_0^S \ddot{\theta}'_0(n)dn + \psi_0^l &= S[6\gamma(\lambda^l + \lambda^{l+1}) + 4\alpha S(\lambda^l + 2\lambda^{l+1}) + \\ &\quad + 3\beta S^2(\lambda^l + 3\lambda^{l+1})] / 12 + \psi_0^l \end{aligned} \quad (4.15)$$

Evaluating Equation (4.15) modulo 2π gives $\psi_0^{l+1'}$ from which δ can be calculated and a new set of target phases derived as outlined in Section 4.2.

Each start and target frequency is scaled by λ^l and λ^{l+1} respectively. Composite amplitude values are calculated by multiplying excitation amplitude values by the LPC system magnitude response at each of the scaled frequencies. (Note that the excitation magnitude spectrum is not re-sampled but frequency scaled.) Composite phase values are calculated by adding the new excitation phase values to LPC system phase response measured at each scaled frequency. Re-synthesis of pitch-scaled speech may then be carried out as in Section 2.2 by computing a phase interpolation function for each sinusoid and substituting into Equation (4.16).

$$\hat{s}^l(n) = \sum_{k=1}^{K^l} A_k^l(n) \cos[\tilde{\theta}_k^l(n)] \quad (4.16)$$

But for the way $\psi_0^{l+1'}$ is calculated, pitch-scaling is quite similar to the time-scaling technique presented in Figure 13. The pitch-scaling algorithm is given in Figure 19. This approach is slightly different to our previous approach, presented in [64], where pitch-scaling was, in effect, converted to a time-scaling problem. Both approaches produce similar quality results.

In Chapter 6 a formal evaluation of various speech samples pitch-scaled using the method put forward above is presented. Results were found to be of high quality and some example waveforms, taken from speech which was pitch-scaled using this method, are given in Figures 16, 17 and 18. Again, it should be noted that the original waveform shape has been generally well preserved.

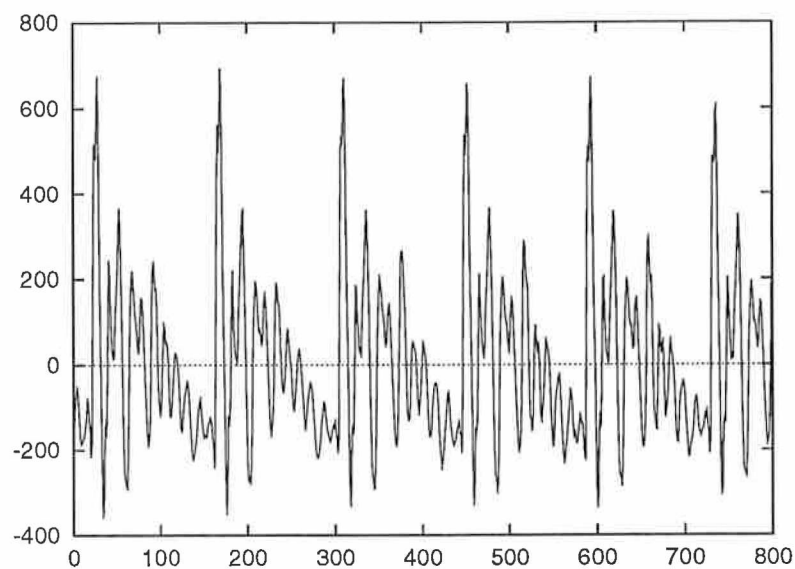


Figure 16 Original speech, $\lambda = 1$

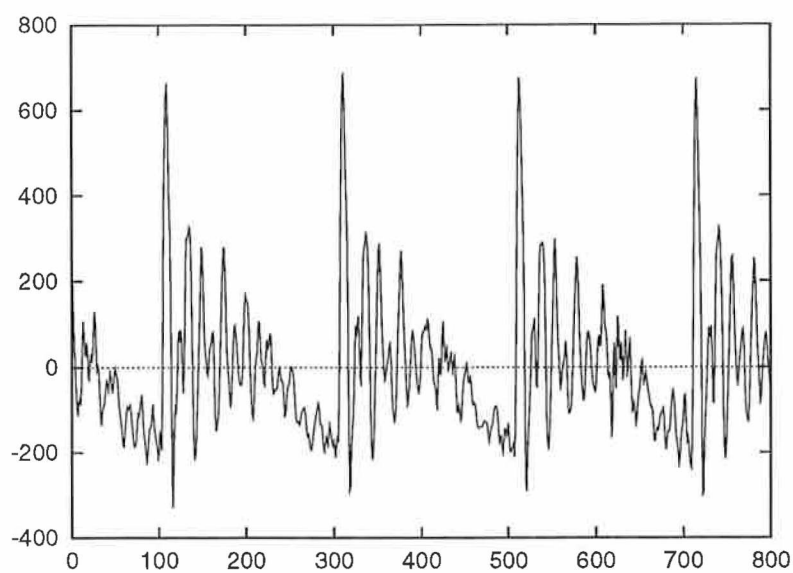


Figure 17 Pitch-scaled speech, $\lambda = 0.7$

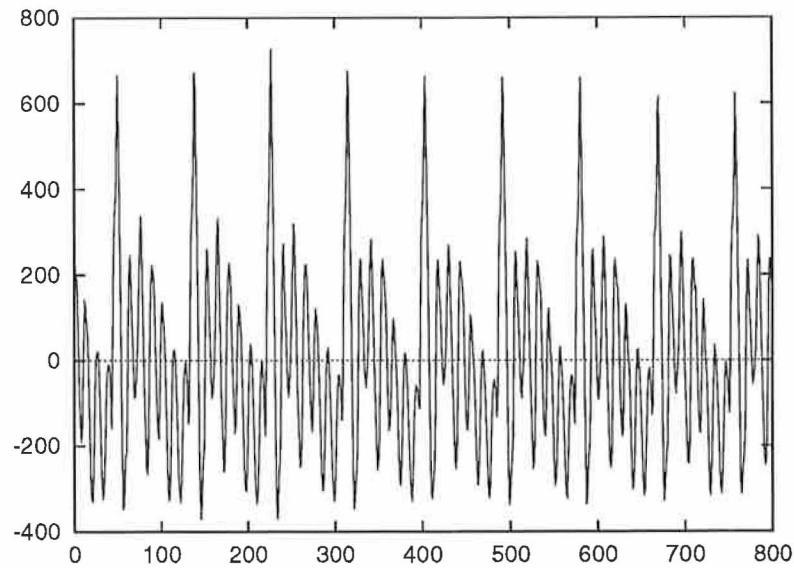


Figure 18 Pitch-scaled speech, $\lambda = 1.6$

4.4 Joint Pitch and Time-Scale Modification

The pitch and time-scale modification methods presented can be easily combined to perform joint modification. The frequency track linking the fundamental of frame l with that of frame $l+1$ can again be written as

$$\dot{\tilde{\theta}}_0(n) = \gamma + 2\alpha n + 3\beta n^2 \quad (4.17)$$

The pitch and time-scaled track, where ρ is the time-scale modification factor associated with frame l and λ^l and λ^{l+1} are the pitch modification factors associated with frames l and $l+1$ respectively, is given by

$$\dot{\tilde{\theta}}'(n) = \dot{\tilde{\theta}}\left(\frac{n}{\rho}\right) \lambda\left(\frac{n}{\rho}\right) \quad (4.18)$$

where $\lambda(n)$ is the linearly interpolated pitch modification factor given in Equation(4.13). Integrating Equation (4.18) over the interval ρS and adding

```

 $\Delta = 0$ 
 $\delta = 0$ 
For each Frame  $l$ 
  Begin
     $\Delta = \Delta + \delta$ 
    For  $\omega_0$ 
      Begin
        Adjust  $\psi_0^{l+1}$  by  $\Delta$ 
        Compute frequency track  $\tilde{\theta}_0(n)$ 
        Compute new frequency track  $\dot{\tilde{\theta}}_0(n)$ 
        Solve for  $\psi_0^{l+1'}$  and  $\delta$ 
        Compute composite amplitude and phase
        Compute phase function  $\theta_0^l(n)$ 
      End
    End
    For  $\omega_k$  where  $k \neq 0$ 
      Begin
        Adjust  $\psi_k^{l+1}$  by  $\delta + \Delta$ 
        Compute composite amplitude and phase
        Compute phase function  $\theta_k^l(n)$ 
      End
    End
  End
End

```

Figure 19 Pitch-scaling algorithm

back the start phase, ψ_0^l , gives

$$\int_0^{\rho S} \dot{\theta}_0'(n) dn + \psi_0^l = \rho S [6\gamma(\lambda^l + \lambda^{l+1}) + 4\alpha S(\lambda^l + 2\lambda^{l+1}) + 3\beta S^2(\lambda^l + 3\lambda^{l+1})] / 12 + \psi_0^l \quad (4.19)$$

Evaluating Equation (4.19) modulo 2π gives $\psi_0^{l+1'}$ from which δ can be calculated and a new set of target phases derived. Using the scaled harmonic frequencies and new composite amplitudes and phases, synthesis (again, as outlined in Section 2.2) is carried out to produce speech that is both pitch and time-scaled. Some example waveforms, taken from speech which has been simultaneously pitch and time-scaled using this method, are given in Figures 20, 21 and 22. In these examples the same pitch and time-scaling factors have been assigned to each frame although, obviously, this need not be the case as both factors are mutually independent. As with the previous examples, waveform shape can be seen to have been well preserved.

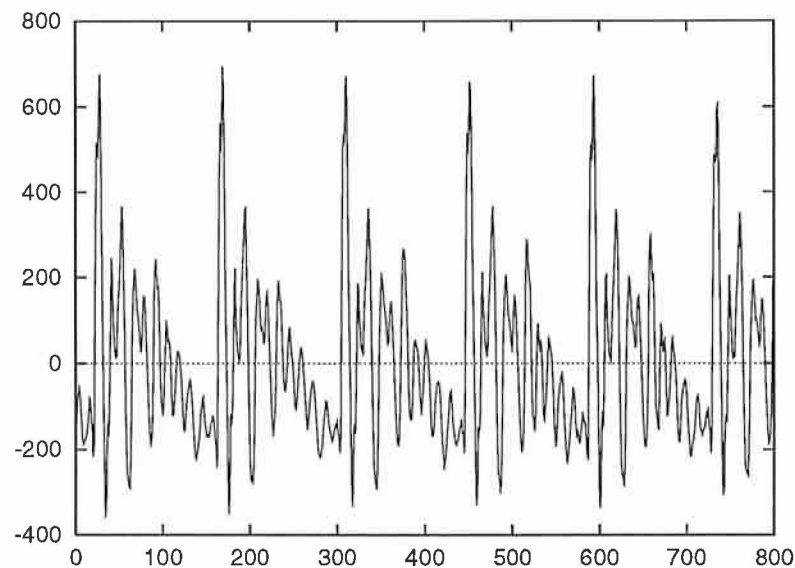


Figure 20 Original speech, $\rho = 1$, $\lambda = 1$

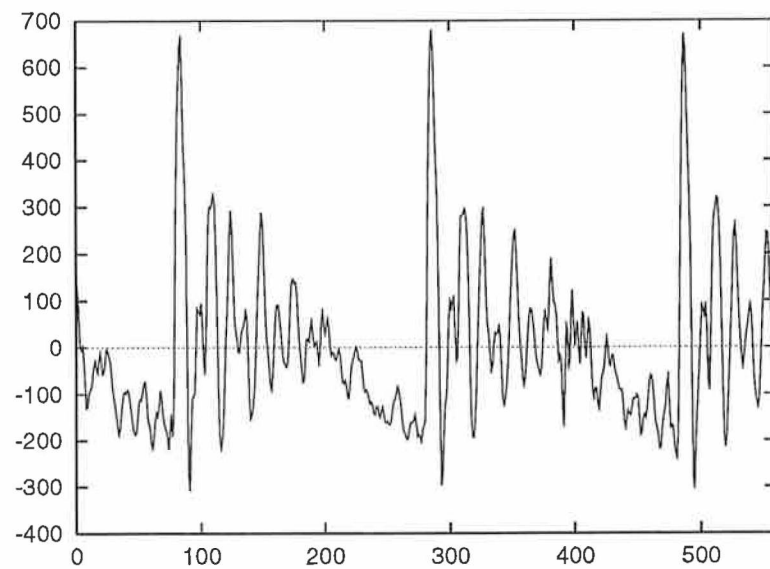


Figure 21 Pitch- and time-scaled speech, $\rho = 0.7$, $\lambda = 0.7$

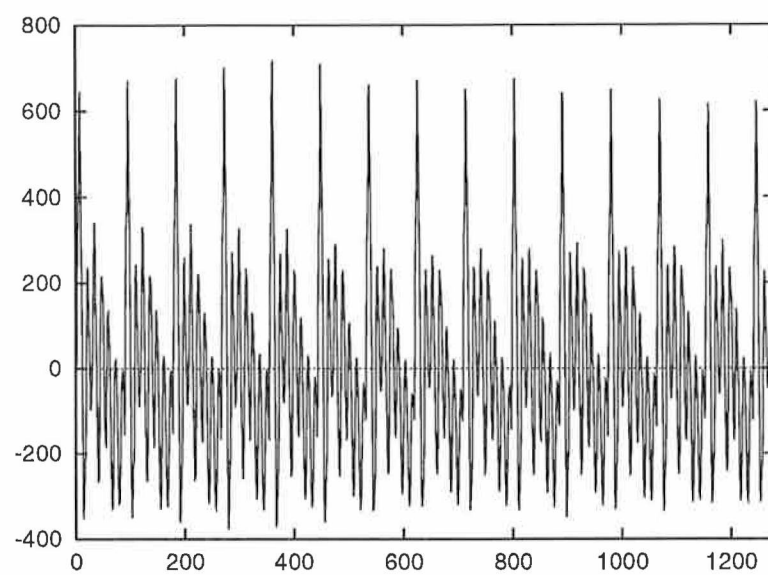


Figure 22 Pitch- and time-scaled speech, $\rho = 1.6$, $\lambda = 1.6$

4.5 Discussion

A high quality yet conceptually simple approach to pitch and time-scale modification of speech has been presented. Taking advantage only of the harmonic structure of the sinusoids used to code each (voiced) frame, phase coherence and waveform shape are well preserved after modification.

The simplicity of the approach stands in marked contrast to the shape invariant modification algorithms of Quatieri and McAulay [37] (Section 3.2). In [37], pitch pulse onset times, used to preserve waveform shape, must be estimated in both the original and target speech. In the approach presented here, onset times play no role and need not be calculated. Furthermore, in [37] onset times are used to impose a structure on phases and errors in their location lead to unnaturalness in the modified speech. In the approach described here, during modification phase relations inherent in the original speech are preserved. Phase coherence is thus guaranteed and waveform shape retained. Obviously, our approach holds a similar advantage over the ABS/OLA modification techniques [39] [40] which also make use of pitch pulse onset times (Section 3.4).

Unlike the PSOLA-inspired HNM [44] [47] approach to speech transformation (Section 3.3), using our technique no mapping need be generated from synthesis to analysis short-time signals. Furthermore, the duplication/deletion of information in the original speech, a characteristic of PSOLA-type techniques, is avoided. Every frame is used once and only once during re-synthesis.

The time-scale modification technique introduced here is somewhat similar to that employed in the ABS/OLA model [39] [40] (Section 3.4) in that both exploit the (quasi-)harmonic nature of the sinusoids used to code each frame. However, the frequency (and associated phase) tracks linking one

frame with the next and playing a crucial role in the sinusoidal model [36], while absent from the ABS/OLA model, are retained here. Furthermore, our new pitch modification algorithm is a direct extension of the new time-scaling approach and is considerably simpler than the “phasor interpolation” mechanism used in the ABS/OLA model.

The incorporation of transformation techniques specific to voiced and voiceless speech brings to light deficiencies in the analysis model presented in Section 4.1. Voicing errors can seriously lower the quality of the re-synthesised speech. For example, where voiced speech is deemed voiceless, phase dithering is wrongly applied, waveform dispersion occurs, and the speech is perceived as having an unnatural “rough” quality. Correspondingly, where voiceless speech is analysed as if voiced, its random nature is not preserved and the speech takes on a tonal character.

Apart from voicing errors other problem areas also exist. Voiced fricatives, by definition, consist of a deterministic and a stochastic component and, because our model applies a binary \pm voice distinction, cannot be accurately modelled. During testing, such sounds were modelled as a set of harmonics (i.e. as if purely voiced) and while this approach coped well with moderate time-scale expansion factors a tonal artifact was introduced for larger degrees of modification.

The model could be improved and the problems outlined above alleviated by incorporating several of the elements used in HNM analysis [47] which was described in Section 2.3. Firstly, leaving the rest of the model as it stands, a more refined pitch estimation procedure could be added to the analysis phase i.e. as in HNM the pitch could be chosen to be that whose harmonics best fit the spectrum. Secondly, the incorporation of a voicing cut-off frequency, used in HNM, would add the flexibility required to solve the problems mentioned

in the previous paragraph. Above the cut-off point, frequency dithering techniques could be employed to ensure noise retained its random character. Below the cut-off point the speech would be modelled as a set of harmonics. During testing of our system on voiced fricative time-scale expansion a cut-off frequency value of 2kHz was applied. Above 2kHz the frequency dithering technique presented in Section 4.2.2 was used to prevent tonality. Although obviously less sophisticated than HNM's approach where an optimal F_{max} value is calculated for each frame, our approach produced good quality results (see Chapter 6). Furthermore, given that our algorithm is intended for use in a concatenative TTS system, where the speech segment under analysis is known, cut-off frequencies such as that above can be imposed a priori.

The extra computational burden incurred in implementing pitch and time-scale modification, using the approach described here, centres on keeping frequencies in phase. The use of a cut-off frequency, above which phases can be considered random, would significantly improve the efficiency of the approach as only frequencies below the cut-off point would require explicit phase monitoring. Obviously, the same idea can also be applied in purely voiceless regions to reduce the total number operations. Another computationally expensive operation is bandwidth expansion during time-scale modification of voiceless regions where the number of frequency components per frame is doubled. This approach is justified on the basis that it is simple, produces high quality results and, using the test suite described in Chapter 6, is rarely required.

In the next chapter existing concatenative TTS systems based on the HNM and ABS/OLA models are described. Also the new pitch and time-scaling algorithms presented here are combined with smoothing techniques to implement our own concatenative TTS system.

CHAPTER 5

SYNTHESIS

In this chapter we describe how the sinusoidal model can be applied in concatenative TTS synthesis. As detailed in Chapter 1, in order to facilitate high quality concatenative synthesis a speech coder must represent speech units in a flexible and parametric form that lends itself to the simple implementation of prosodic modification and smoothing techniques. Sinusoidal modelling provides such a coding scheme. Its ability to perform high quality prosodic modification has been demonstrated in Chapters 3 and 4. This chapter focuses on the various types of discontinuity that arise during concatenation of speech units extracted from disjoint utterances and how the sinusoidal model can be applied to smoothing them. Such discontinuities include energy mismatches, spectral discontinuities (e.g. in formant location and bandwidth) and phase mismatches (see Chapter 1).

The rest of this chapter is organised as follows. Section 5.1 describes AT&T's Next Generation TTS system [24] [23] [66]. Using a HNM synthesis backend and a unit selection algorithm to choose an optimal diphone sequence from an extensive speech database, it is currently one of the best synthesisers available. Macon's TTS system [67] [42] [41], based on the ABS/OLA model, is described in Section 5.2. We have developed a TTS system based on the pitch and time-scaling algorithms presented in Chapter 4 and it is discussed in Section 5.3. In each of these sections emphasis is placed on each system's approach to eliminating the discontinuities mentioned above, particularly phase mismatch. The chapter concludes with a summary in Section 5.4.

5.1 AT&T's Next Generation TTS System

AT&T's Next Generation TTS system [24] [23] [66] draws its components from several existing systems. Implemented in the Festival framework [68], its text normalisation, letter-to-sound and prosody generation modules are taken from AT&T's Flextalk TTS system. Dynamic unit selection is carried out using an extended version of an algorithm employed by CHATR [69] [70] [71] (a speech synthesis system developed at ATR Interpreting Telecommunications Laboratories, Kyoto, Japan). A HNM (see Sections 2.3 and 3.3) synthesis backend is used to transform and smooth the prosody of selected speech units at synthesis time.

5.1.1 *Synthesis*

The speech unit database used in AT&T's system consists of 1.5 hours of labelled female speech which typically contains several instances of any given diphone [24]. (Listening tests have apparently indicated that a database containing only a single copy of each diphone, while leading to a trivial unit "selection" procedure, does not provide enough coverage for high quality, natural-sounding, speech synthesis [23].) At runtime a Viterbi search procedure is used to choose an optimal sequence of diphones, i.e. the one that minimises a defined "cost" function, for target speech synthesis. The smoother the transition between diphones and the closer a diphone's prosody is to its target, the lower the cost. It is worth noting that, provided a diphone's prosody does not differ greatly from its target, it is not modified during synthesis. Tests have shown that listeners consistently prefer the unmodified version [24] [23]. Current pitch modification techniques assume that the excitation signal and vocal tract contributions to the speech production process

are independent and can be separately modified and modelled. Beutnagel et al. [24] suggest that this may not be the case and changes in the excitation signal properties (e.g. F_0) may be accompanied by a change in spectral shape. Not taking such effects into account may cause the pitch-scaled speech to take on an unnatural character perceptible to listeners. Further research into the interaction between F_0 and spectral shape is required [24].

Once an optimal set of diphones has been retrieved from the database, it is passed to the HNM synthesis backend. If necessary, the PSOLA-inspired pitch and time-scale modification algorithms, described in Section 3.3, are applied. Over voiced concatenation points spectral smoothing is implemented by simple linear interpolation of harmonic amplitudes. (Discontinuities in the stochastic component are not smoothed as they are considered perceptually irrelevant [66].) Over purely voiceless concatenation points LPC gain and reflection coefficients are also linearly interpolated. There is no smoothing across voiced/voiceless concatenation points.

Phase mismatch is a problem for all concatenative speech synthesisers. In the case of HNM, because speech units are extracted from disjoint utterances and analysis frames are not centred on pitch marks, pitch period synchronisation during OLA is not guaranteed. Non-coherent OLA leads to serious synthetic speech quality degradation. Several approaches have been suggested for solving this problem. In [72], Stylianou et al. estimate phase offsets at frame boundaries using a cross-correlation technique and differences are adjusted in order to ensure coherent OLA. This method was subsequently deemed cumbersome and inefficient [62]. Also in [72], a minimum phase approach (i.e. one in which measured phase values were discarded) was tested but, while guaranteeing coherent OLA, led to poor quality output (a “buzziness” was apparent in the synthetic speech) and was considered unsuitable

for high quality synthesis.

A solution to the problem has been proposed by Stylianou [62] in which an off-line “centre of gravity” centring process is applied to each voiced speech frame. In this procedure analysis frames are placed at a pitch-synchronous rate across the speech waveform but are not centred on glottal closure instants or pitch-marks. Instead, by adding a linear component to the measured phase of each harmonic, the effect is to translate the speech in time such that its centre of gravity, corresponding to the instant of glottal closure, is moved to the centre of each frame. From [62], the new phase of the k^{th} harmonic, $\hat{\psi}(k\omega_0)$, is given by

$$\hat{\psi}(k\omega_0) = \psi(k\omega_0) - k\psi(\omega_0) \quad (5.1)$$

Crucially, the above time-shift can be applied during analysis and in no way complicates the synthesis procedure. This procedure has been applied to AT&T’s speech database and the resulting synthetic speech shows no evidence of phase mismatch [62]. Furthermore, in previous HNM versions [72], not knowing the location of the instant of glottal closure, with which the aperiodic speech component must be synchronised to ensure perceptual integration, was a serious problem. Implementing the above procedure, however, moves the instant of glottal closure to the centre of each frame and the stochastic can be easily synchronised with it by simply applying a triangular window [62] [66]. Lastly, Stylianou [62] points out that this procedure ought to greatly benefit TD-PSOLA based concatenative systems. In order to extract analysis signals TD-PSOLA relies on accurate pitch-marking and although automatic pitch marking algorithms exist, their output must often be hand checked and corrected. Stylianou reports no such errors using his method [62] thereby rendering the addition of new voices to TD-PSOLA systems, hitherto a labour intensive exercise, quite trivial.

5.1.2 *Experimental Results*

Listening tests have shown HNM to consistently outperform TD-PSOLA in diphone based concatenative speech synthesis. The results of one such test, conducted by Stylianou et al. [72], are presented. Using the same diphone set and target prosody (extracted from natural speech) both HNM and TD-PSOLA based systems were used to produce synthetic speech which was then compared. HNM was found to produce smoother, more natural speech than TD-PSOLA. Voiced fricatives, breathy vowels and, in general, voiceless speech were better modelled by HNM. This finding is not surprising given the extra flexibility afforded by HNM's separate and parametric modelling of the deterministic and stochastic speech signal components. TD-PSOLA's non-parametric and static "modelling" of speech offers very limited manipulation possibilities. Consequently artifacts, a result of unwanted periodicities in voiceless or partially voiced speech, are often apparent in synthetic speech produced using this method. Listeners' general preference for HNM has also been borne out by a formal listening test carried out by Syrdal et al. [45] where synthetic speech produced by HNM out-scored that produced by TD-PSOLA in terms of naturalness, intelligibility and pleasantness.

5.2 The ABS/OLA TTS System

Macon [67] [42] [41] has proposed a concatenative TTS system based on the ABS/OLA model (see Sections 2.4 and 3.4). Given its high quality prosodic modification capabilities, its parametric representation of the speech signal (thus facilitating smoothing) and the fact that re-synthesis is implemented using an efficient inverse-FFT procedure, the ABS/OLA model would seem an excellent candidate for use in concatenative TTS synthesis.

5.2.1 *Synthesis*

Macon's TTS system [41] draws its frontend components from British Telecom's Laureate TTS system [73] and employs a synthesis backend based on the ABS/OLA model. Once a sequence of diphones to be concatenated has been retrieved from the database, an energy normalisation algorithm is applied to smooth large energy fluctuations in voiced frames. Perceptually significant short-time energy mismatches may remain at dipphone boundaries and these are smoothed during synthesis by calculating appropriate gain terms for frames to the left and right of each concatenation point. Spectral smoothing is implemented by interpolation of the cepstral features that describe the spectral envelope of each frame to the left and right of a dipphone boundary.

While the above procedures are reasonably effective at smoothing energy and spectral discontinuities, phase mismatch and the consequent non-coherent OLA remain a serious problem. In [41] Macon presents an algorithm, based on pitch pulse onset time estimation, for maintaining inter-frame synchronisation during pitch and time-scale modification of continuous speech. In the latter case, assumptions can be made about the location of onset times in frame $l+1$ based on knowledge of those in frame l . However, when concatenating speech taken from disjoint utterances no such assumptions can be made, resulting in a more complex algorithm which is also presented in [41]. Importantly, and as pointed out by Macon [41], the performance of these algorithms, particularly in the concatenative case, is critically reliant on highly accurate pitch pulse onset time estimation. Even small errors will result in non-coherent OLA and synthetic speech quality degradation. This reliance on onset times, given the non-robust nature of the algorithms used to determine their location, seriously detracts from the ABS/OLA synthesis

model.

5.2.2 *Experimental Results*

In a formal listening test [41] listeners were presented with speech synthesised using a TD-PSOLA implementation and speech synthesised from the same diphones using the ABS/OLA model and smoothing techniques outlined above. Synthetic intonation was generated using a phonological model and listeners were asked to rate the speech in terms of “overall sound quality”. Results showed a statistically non-significant 52% versus 48% preference for the ABS/OLA model. One reason put forward by Macon [41] for the equivalent ratings was the fact that no duration model was used during the test - during synthesis only a diphone’s pitch was modified, its duration remaining unchanged. Duration modification (particularly time-scale expansion) is a strength of the ABS/OLA model but can be a weakness in TD-PSOLA. Failure to bring this superiority to light may have led to equal ratings being assigned by listeners to both systems. We can, however, conclude that at least in terms of their pitch modification performance both the ABS/OLA and TD-PSOLA models were assigned an equal rating by listeners.

5.3 The SHMDCU TTS System

We have developed a speech synthesis system based on the pitch and time-scale modification algorithms presented in Chapter 4 and incorporating smoothing techniques necessary for concatenative synthesis. The SHMDCU (Simple Harmonic Model at Dublin City University) system uses a diphone database made available with the Festival TTS installation*. Each diphone

*<http://www.cstr.ed.ac.uk/projects/festival/festival.html>

is coded as described in Section 4.1 i.e. voiced frames are coded as a set of harmonics and voiceless frames as a set of peak frequencies.

5.3.1 *Synthesis*

At synthesis time the sequence of diphones necessary to synthesise the target speech are retrieved from the speech unit database. Unit selection is a trivial process as the database contains only a single copy of each di-phone. Target prosody (currently extracted from natural speech) is used to assign pitch, time-scale and energy modification factors to each frame. Each modification factor, μ , is simply given by

$$\mu = \frac{\textit{Target value}}{\textit{Measured value}} \quad (5.2)$$

Once scaling factors have been assigned and harmonic amplitudes scaled to match to the target energy level, synthesis is carried out as described in Chapter 4. Spectral smoothing across voiced concatenation points, as in AT&T's Next Generation TTS system, is implemented by simple linear interpolation of harmonic amplitudes. Similarly, across (partially) voiceless concatenation points peak frequency amplitudes are also interpolated linearly.

As was the case for the other concatenative systems presented in this chapter, phase mismatch at concatenation points is a problem. In those systems phase mismatch leads to non-coherent OLA, resulting in garbled speech quality. Using the system presented here, however, the problem is slightly different. At di-phone boundaries target phase values bear no relation to start phase values. As a result, frequencies may "struggle" to meet their target phase and the consequent contorted frequency tracks can lead to waveform dispersion and noisy transitions. The resulting synthetic speech quality is seriously degraded.

The solution proposed here to this problem is to simply make frequency track transitions across diphone boundaries as smooth as possible. Let frame l be the last frame of diphone m and frame $l + 1$ be the first frame of diphone $m + 1$. As was the case with the pitch and time-scaling algorithms put forward in Chapter 4, we again rely on the fundamental to keep waves “locked together” in phase. Let ω_0^l and ω_0^{l+1} be the pitch-scaled fundamentals in frames l and $l + 1$ respectively. Discarding the measured target phase value, ψ_0^{l+1} , the smoothest frequency track between fundamentals, $\dot{\theta}_0(n)$, is calculated by simple linear interpolation and is given by

$$\dot{\theta}_0(n) = \lambda^l \omega_0^l + \frac{\lambda^{l+1} \omega_0^{l+1} - \lambda^l \omega_0^l}{\rho S} n \quad (5.3)$$

where S is the analysis frame interval, λ^l and λ^{l+1} are the pitch modification factors associated with frames l and $l + 1$ respectively and ρ is the time-scale modification factor. Integrating Equation (5.3) and adding back the start phase, ψ_0^l , gives the new target phase value $\psi_0^{l+1'}$. Again, as was the case for pitch and time-scale modification, the amount by which all other target phases must be adjusted in order to calculate their new value is given by δ where

$$\delta = \frac{\psi_0^{l+1'} - \psi_0^{l+1}}{\omega_0^{l+1}} \quad (5.4)$$

The phases of all harmonics in frame $l + 1$ are then made consistent with that of their fundamental by applying

$$\psi_k' = \psi_k + \omega_k \delta \quad (5.5)$$

Synthesis may then be carried out by computing the smoothest track from each harmonic's start to target parameters (see Section 2.2).

Note that the above solution to phase mismatch relies on the assumption that across voiced concatenation points phase relations between harmonics

and their fundamental will be quite similar. Thus, by making the fundamental's transition as smooth as possible all other transitions will also be made smoother.

It should be pointed out that this approach, adds no computational overhead to the synthesis procedure presented in Chapter 4. When performing pitch and time-scale modification the same process must be followed i.e. a new target phase configuration must be generated based on the pitch and time-scaled fundamental. Thus our solution to the problem of phase mismatch fits neatly into and follows directly from our existing pitch and time-scale modification algorithms.

That this approach is indeed effective at removing phase mismatches is illustrated by Figures 23 through 28. In Figures 23 and 25 a section of speech waveform traversing a diphone boundary where phase mismatch correction has not been applied is presented. Waveform shape dispersion is evident in both cases. In contrast, in Figures 24 and 26 the phase mismatch procedure outlined above has been applied and waveform shape is well preserved. Also presented in Figures 27 and 28 are the spectra (up to 1kHz) of the speech in Figures 25 and 26. Clearly, individual harmonic frequencies are better resolved when phase correction has been applied.

Lastly, this solution to the phase mismatch problem suggests a similar approach could be adopted during synthesis. Rather than interpolate the pitch and time-scaled fundamental frequency in order to estimate target phase parameters, using the above approach a target phase set (consistent, of course, with the measured target phase set, taken from the FFT analysis) could be imposed for each frame. A consequence of such a synthesis technique would be that even in the absence of any modification (i.e. $\rho^l = 1$ and $\lambda^l = 1 \forall l$) the measured phase values would be adjusted at synthesis time such that the

smoothest F0 transition from one frame to the next was obtained. Furthermore, using this approach would mean that diphone concatenation points would not be a special case, the same synthesis procedure being applied for all frames. Synthesis then would not be strictly bound to measured phase values, rather each set would serve as a template for all valid target phase configurations. This approach has been tested and achieves both high quality re-synthesis and modification.

5.3.2 Experimental Results

A copy synthesis experiment was carried out using natural prosody (taken from a speaker different to that used to produce the diphone database) to generate the phrase "I need to arrive by 10:30am on Saturday". The resulting synthetic speech was found to be of a high quality, close to that of natural speech.

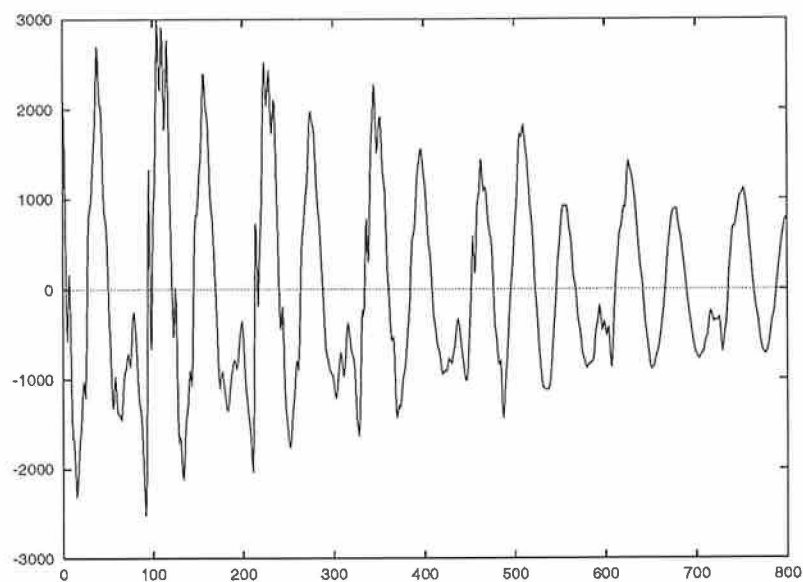


Figure 23 Diphone boundary waveform with phase mismatch

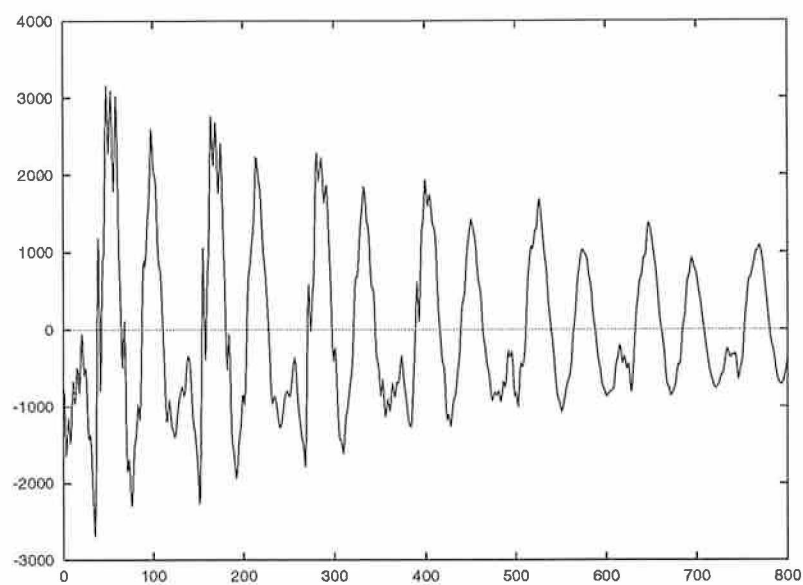


Figure 24 Diphone boundary waveform with phase mismatch correction

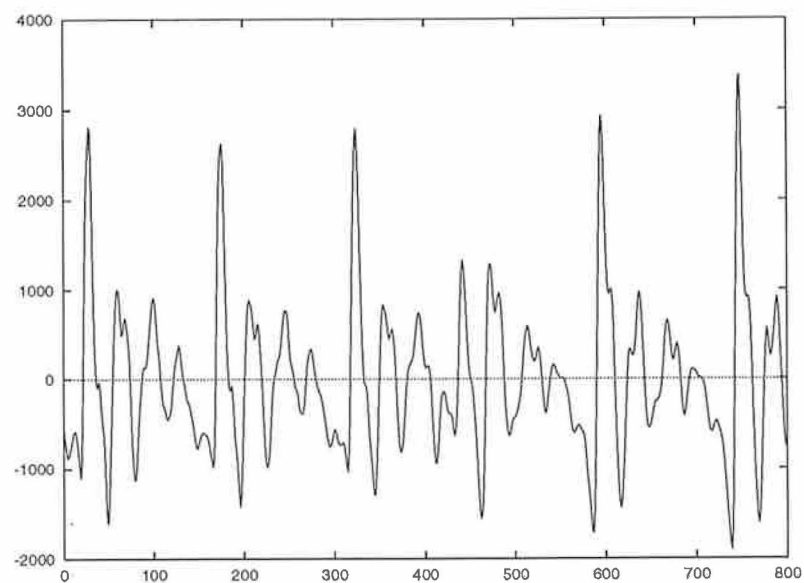


Figure 25 Diphone boundary waveform with phase mismatch

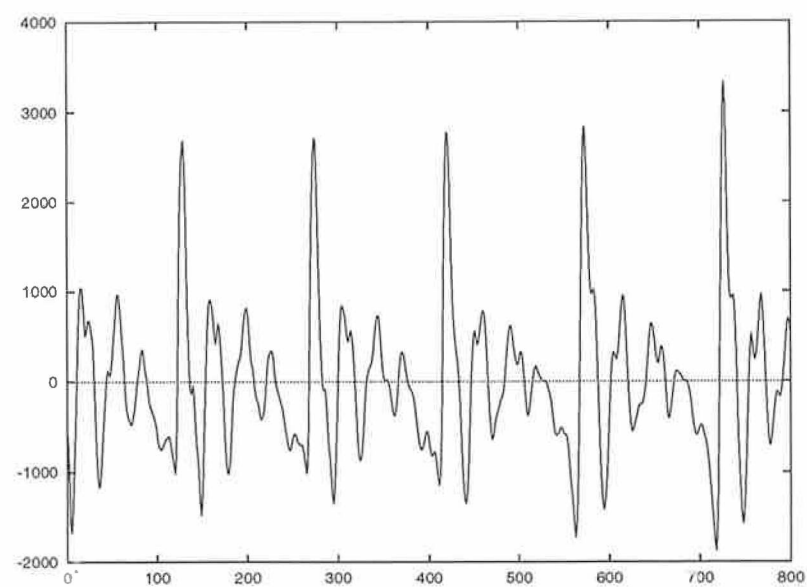


Figure 26 Diphone boundary waveform with phase mismatch correction

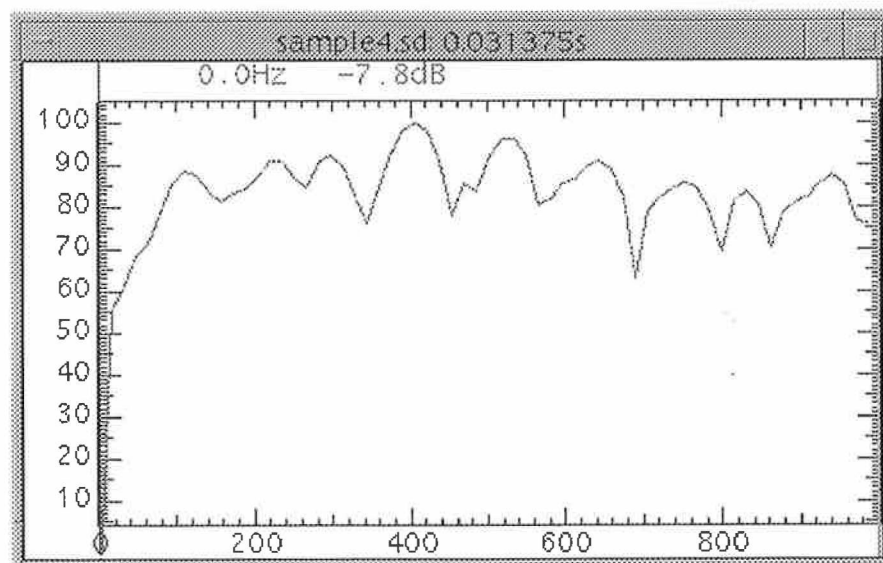


Figure 27 Diphone boundary spectrum with phase mismatch

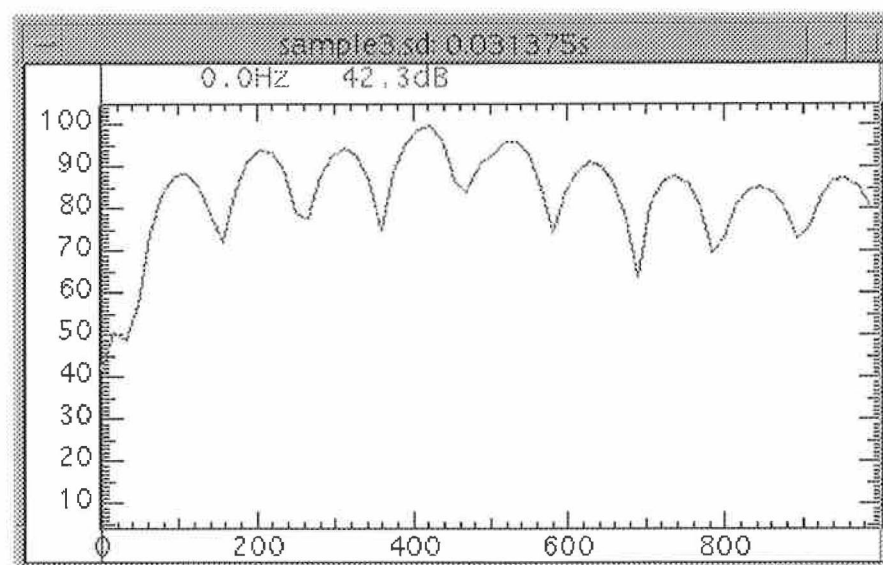


Figure 28 Diphone boundary spectrum with phase mismatch correction

5.4 Summary

Three approaches to concatenative speech synthesis based on a sinusoidal model have been examined in this chapter. Presented in Section 5.1, the AT&T Next Generation TTS system combines a CHATR-style unit selection algorithm with a HNM synthesis backend to produce very high quality synthetic speech. AT&T's system also incorporates a simple and importantly off-line solution to the problem of phase mismatch. Macon's TTS system, described in Section 5.2, uses an ABS/OLA synthesis backend with frontend components being taken from British Telecom's Laureate TTS system. Although capable of high quality synthesis the system's dependence on high precision pitch pulse onset time estimates is a serious drawback.

Finally, a new system was presented in Section 5.3. Based on the pitch and time-scaling algorithms put forward in Chapter 4, phase mismatches are corrected during synthesis by simple linear interpolation of the fundamental frequency. Informal listening tests have shown our system capable of high quality synthesis. Importantly, phase mismatch correction is implemented as an extension to our prosody modification techniques and adds no extra computational cost to the synthesis process.

It is planned to incorporate our synthesis system into the Festival framework at a later date in order to permit formal comparison with other TTS systems. However, a formal evaluation of the prosodic transplantation performance of HNM (as implemented by the Institut de la Communication Parlée, Grenoble, France) and Banga et al. 's, Pitch Synchronous Sinusoidal Model (PSSM) [75] [74], compared with that of SHMDCU has been conducted and the results of this study are presented in the next chapter.

CHAPTER 6

RESULTS

A listening test was conducted in order to allow a formal comparison of the performance of three sinusoidal based speech coders in a prosodic transplantation task. The results of the test form the basis of this chapter. The systems compared were HNMICP (HNM as implemented by Institut de la Communication Parlée, Grenoble, France), PSSVGO (a pitch synchronous sinusoidal model developed at the University of Vigo, Spain) and our own sinusoidal harmonic model, SHMDCU, incorporating the pitch and time-scale modification techniques which were presented in Chapter 4. Results show speech, pitch- and time-scaled using our approach, compares very favourably in terms of overall quality with speech which has been similarly scaled using the HNMICP and PSSVGO systems.

This chapter is organised as follows. In Section 6.1 the test suite used in the listening test is described. The experiment itself is described in Section 6.3. Results and discussion are presented in Sections 6.4 and 6.5 respectively.

6.1 The COST 258 Coder Evaluation Server

The COST 258 coder evaluation server* was created to provide a means of comparing the prosodic transformation performance of speech coders. Transformation tasks can be downloaded from and results uploaded to the server which thus provides a database of speech which has been pitch- and time-scaled using a number of approaches but with the same target prosody in

*http://www.icp.grenet.fr/cost258/evaluation/server/cost258_coders.html

mind. Formal comparisons between systems are therefore greatly facilitated. The various transformation tasks, designed to emulate those commonly encountered in concatenative speech synthesis, are described in more detail below.

Provided by the server is a set of natural speech samples with “flat” prosody (i.e. where the speaker has attempted to maintain a constant pitch level across the utterance) and for each one a set of associated target prosodic contours. Each set of target parameters is derived from natural speech where the speaker who produced the original flat utterance repeats its content but deliberately alters his prosody. The task for each coder is to transform the flat utterance such that its modified pitch and time-scales match those of the target utterance. Speech samples to be transformed include vowels, fricatives (voiced and voiceless) and continuous speech (in French and Czech). As mentioned, results from other coders are also made available on the server and including those produced by the HNMICP and PSSVGO systems. These were downloaded and included in the experiment described in Section 6.3.

6.2 Prosodic Transplantation

In each case the original flat utterance was analysed and coded as described in Section 4.1. Based on the supplied target prosodic parameters, to each frame was assigned both a pitch and time-scale modification factor. The prosodically modified speech was then synthesised using the algorithms described in Chapter 4. To illustrate this procedure some example transformations (carried out on vowels and continuous speech) are presented below. In each case the resulting speech was found to be of high quality.

Given in Figures 29 and 30 are, respectively, an original speech waveform (of a vowel) and its pitch contour. Pitch and time-scale modification was

carried out to produce the speech shown in Figure 31. As can be seen from Figure 32 the vowel's pitch contour has been successfully scaled to match that of its target. In Figures 33 through 36 similar data is supplied for a transformation carried out on continuous speech.

Lastly, the quality of time-scale expanded fricatives, voiced and voiceless, was found to be high. The frequency dithering technique presented in Section 4.2.2 was effective at preventing tonality.

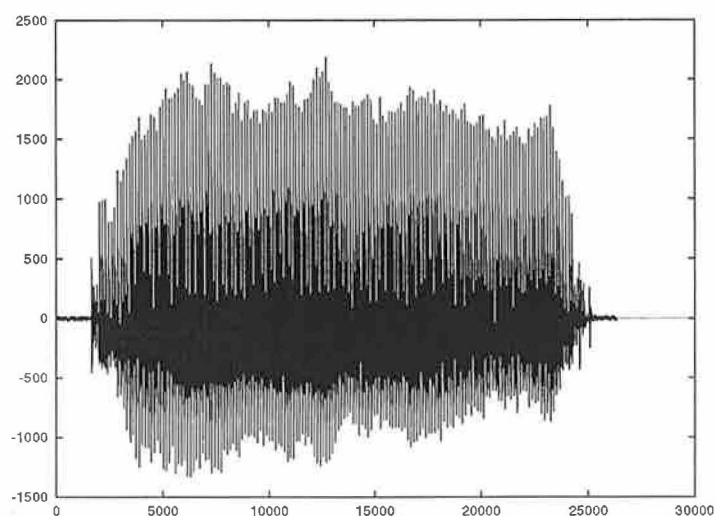


Figure 29 Original waveform

6.3 Experimental Procedure

Twenty-eight utterance-length target contours were selected from the prosodic transplantation task, and the versions produced by the SHMDCU model were compared against the versions from the HNM implementation from Grenoble (HNMICP) and the Vigo technique (PSSVGO). It should be

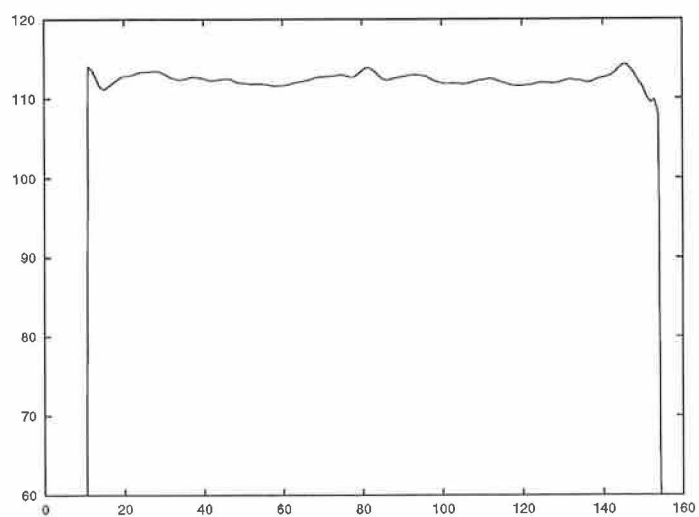


Figure 30 Original pitch contour

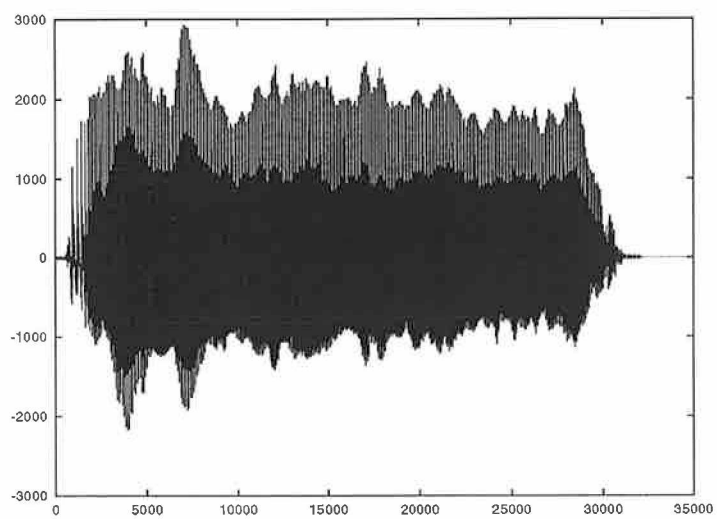


Figure 31 Pitch- and time-scaled waveform

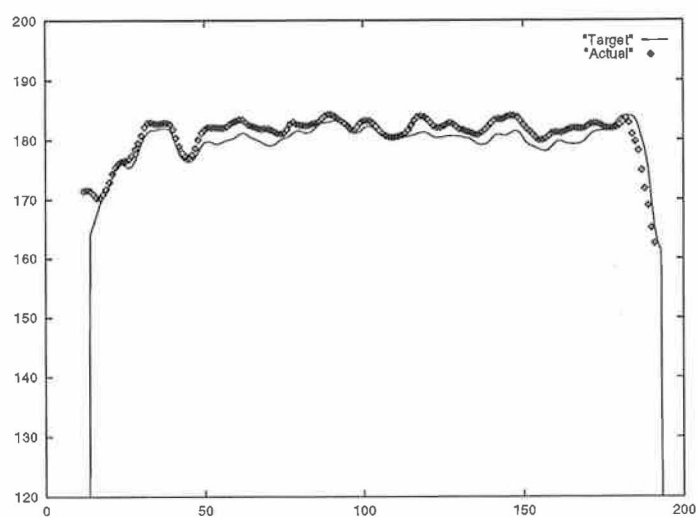


Figure 32 Target and actual pitch contours

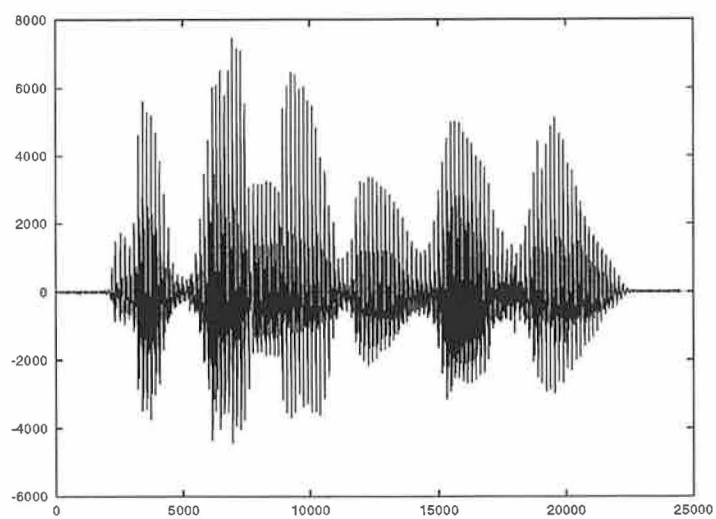


Figure 33 Original waveform

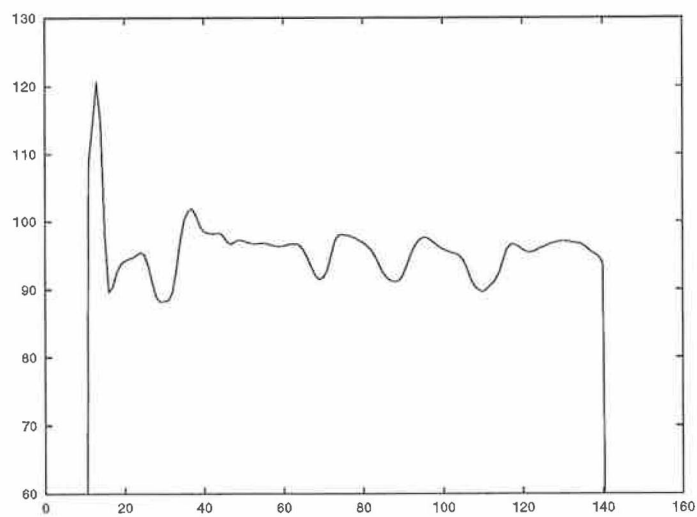


Figure 34 Original pitch contour

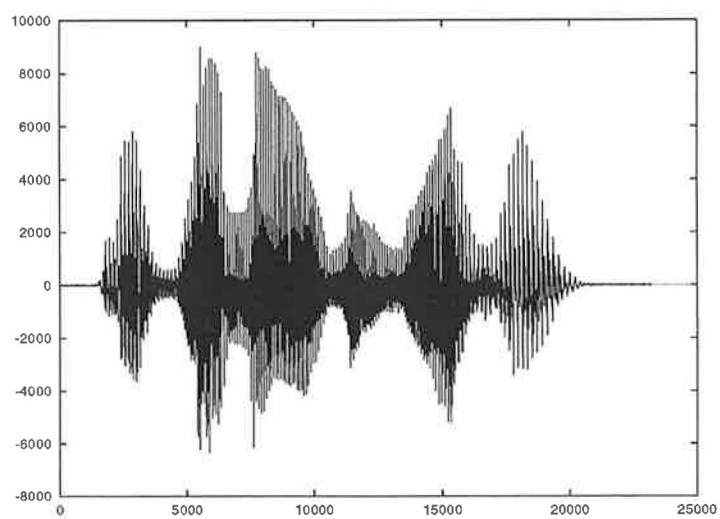


Figure 35 Pitch- and time-scaled speech

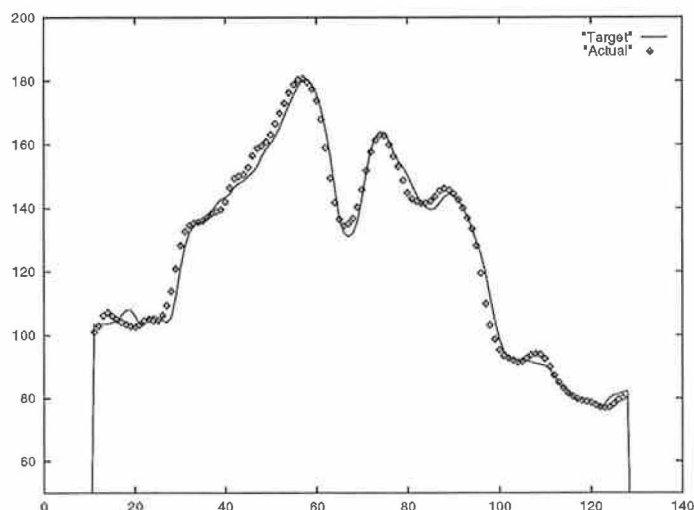


Figure 36 Target and actual pitch contours

emphasised that all three models are development systems only, and are constantly being improved and extended. The versions tested here were those existing in August 1999.

In a pairwise comparison, listeners were asked to indicate which of two versions of the same target utterance they considered to be of higher quality. The utterances were in French and Czech, and the listeners were not native speakers of those languages nor did they hear the natural utterance which provided the target contours, so that all they could judge was acoustic quality rather than intelligibility or closeness to the natural target.

Listeners were split into experts (those with a high level of familiarity with synthetic speech) and non-experts. There were four experts and ten non-experts. The experts judged two sets of pairs: SHMDCU v HNMICP, and SHMDCU v PSSVGO. The non-experts judged only one set of pairs. The stimuli were arranged in two different balanced random orders, where one order was the reverse of the other (i.e. BA-BA-AB became BA-AB-AB):

expert listeners judged one set of pairs in each order, and non-expert listeners were divided between the two orders.

Listeners were allowed to listen to each pair as many times as they liked before making their judgement. They were not allowed to judge both elements of a pair to be equally good. Listeners' comments indicated that in some cases the decision was difficult but in most cases there was a clear difference in quality. A spot check on consistency of judgements showed that the same listener would make the same choices a few minutes later with an agreement of between 75% and 80%.

6.4 Results

The results showed a high degree of consistency between listeners. There was no appreciable difference between the judgements of experts and non-experts, or between the different orders of presentation. The overall average scores are therefore presented without further analysis.

Average scores for the SHMDCU v HNMICP comparison were 87.4% for SHMDCU and 12.6% for HNMICP.

Average scores for the SHMDCU v PSSVGO comparison were 70.9% for SHMDCU and 29.1% for PSSVGO.

It should be emphasised that all these stimuli were prepared in the institutions concerned, and were freely available on the internet. In this respect, our evaluation represents a very fair test. However, as we mentioned above, these systems are not commercial products, nor are they stable versions: they are part of the continuing research programme on signal generation within COST 258, and they have probably all been modified since these stimuli were produced. Nevertheless, we consider that these results show the SHMDCU

model to be at least as good as other state-of-the-art systems for pitch and time-scale modification of speech.

6.5 Discussion

Although only the quality of continuous speech samples were compared in the listening test described above both transformed vowels and fricatives were also of high quality. A voicing cut-off frequency of 2000kHz was used in voiced fricatives above which the frequency dithering technique presented in Section 4.2.2 was applied. Good quality results were produced using this method. Frequency dithering was used in the continuous speech case only in purely voiceless regions. As a result, an unnatural tonal quality was imparted to time-scale expanded voiced fricatives in a small number of cases. However, our algorithms have been designed for concatenative speech synthesis applications where voiced fricatives can be assigned a cut-off frequency a priori. Our frequency dithering technique, the efficacy of which has been borne out by tests on voiceless and voiced fricatives contained in the COST 258 test suite, can then be applied above the cut-off frequency.

In summary, both the informal and formal listening tests described in this chapter have demonstrated our system to be capable of very high quality pitch and time-scale modification. Although producing results comparing favourably with those produced by other speech coders there remains room for improvement. The following chapter concludes the thesis, presenting our conclusions and suggestions for future work.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

The development of a new concatenative TTS system based on a sinusoidal speech model has formed the basis of the research presented in this thesis. During the course of that development existing synthesis systems have been examined in detail and new methods for pitch and time-scale modification, frequency dithering and phase mismatch correction at concatenation points have been proposed. Experiment has shown our algorithms to be capable of high quality copy synthesis and to perform very well in prosodic transplantation tasks.

Chapter 2 served to introduce the original sinusoidal model and two of its derivatives - the ABS/OLA model and the hybrid HNM approach. In Chapter 3 various prosodic modification techniques based on those systems discussed in Chapter 2 were presented.

New algorithms for pitch and time-scale modification of speech were presented in Chapter 4. Relying neither on pitch-synchronous analysis nor pitch pulse onset time estimation, these algorithms exploited the malleability of the harmonic representation used to code each (partially) voiced frame to ensure post-modification phase coherence. Unlike with TD-PSOLA, where certain frames are omitted from or duplicated in the modified speech, in our approach each frame makes a single contribution to the pitch and/or time-scaled version. Furthermore, rather than resorting to a hybrid approach (and having to deal with the perceptual fusion problems such an approach often entails) a simple but effective frequency dithering technique was developed

for use when time-scaling voiceless or partially voiced regions. By splitting each “noisy” sinusoid in two, thereby widening its bandwidth, perceptual randomness was maintained and spurious tonality eliminated.

Chapter 5 presented synthesis systems based on the HNM and ABS/OLA models with particular attention being paid to their approaches to smoothing at concatenation points. The algorithms developed in Chapter 4 were incorporated into our TTS system. This led to the development of a new phase correction technique which, again relying on the versatility of the harmonic representation used to code each frame, was simply and naturally amalgamated into the existing pitch and time-scale modification techniques. Importantly, this was achieved at no extra computational cost.

Results of an evaluation were presented in Chapter 6. In a prosodic transplantation experiment the performance of our pitch and time-scale modification algorithms was compared to that of two other systems and a clear preference for our method emerged.

In summary, the major contribution of this thesis is the development of new pitch- and time-scaling algorithms which are conceptually very simple, intuitively appealing and have been shown to produce excellent quality results. The development of such techniques is essential if high quality concatenative synthesis is to be achieved without recourse to extremely large speech databases and the complicated unit retrieval algorithms such approaches necessitate. Lastly however, it is worth noting that the prosody manipulation routines developed could easily be accommodated into a hybrid model such as HNM.

Although capable of high quality results, there is still room for improvement in our model. Analysis is extremely straightforward and a more sophisticated approach to choosing harmonic parameters such as that employed in

ABS/OLA or HNM would undoubtedly lead to more accurate modelling. Furthermore, the LPC inverse-filtering technique used to estimate the glottal excitation is rudimentary and results would surely benefit from a more refined approach. During analysis a crude voicing decision algorithm, which frequently makes errors, is used to differentiate voiced from voiceless regions: there is thus room for improvement in this area too. Each of these improvements, if implemented would only affect the analysis phase and would not incur any computational cost during synthesis.

Further work remains to be done in comparing our TTS system against the output of others. In an effort to facilitate such comparisons it is planned to incorporate our model into the Festival system. Such a framework would also allow for testing of the efficacy of our phase mismatch correction algorithm compared to those employed by other systems. Given the excellent results from prosodic transplantation and copy synthesis experiments, we are confident that our synthesis system will produce speech of quality comparable to that of the best synthesisers currently available.

7.1 References

- [1] Klatt Dennis H. Review of text-to-speech conversion for english. *Journal of the Acoustical Society of America*, 82(3):737–793, September 1987.
- [2] Allen J. B. Synthesis of speech from unrestricted text. *Proceedings of the IEEE*, pages 422–433, 1976.
- [3] Monaghan A. I. C. What determines accentuation? *Journal of Pragmatics*, 19:559–584, 1993.
- [4] Monaghan A. I. C. Generating intonation in the absence of essential information. In Ainsworth and Holmes, editors, *Speech 88: Proceedings of the 7th FASE Symposium*, pages 1249–1256, 1988.
- [5] Monaghan A. I. C. Heuristic strategies for higher-level analysis of unrestricted text. In Bailly G. and Benoit C., editors, *Talking Machines*, pages 143–161, Amsterdam, 1992. Elsevier.
- [6] Campbell W. N., Isard S. D., Monaghan A., and Verhoeven J. Duration, pitch and diphones in the CSTR TTS system. In *Proceedings of ICSLP*, pages 825–828, Kobe, Japan, November 1990.
- [7] Hart J. and Cohen A. Intonation by rule: a perceptual quest. *Journal of Phonetics*, pages 309–327, 1973.
- [8] Monaghan A. I. C. A system for left-to-right intonation specification from text. In *Proceedings of EUROSPEECH*, volume 2, pages 25–28, 1987.
- [9] Monaghan A. I. C. *Intonation in a text-to-speech conversion system*. PhD thesis, University of Edinburgh, 1991.
- [10] Fant G. *Acoustic theory of speech production*. Mouton, 's-Gravenhage, The Netherlands, 1960.
- [11] Rothenberg M., Carlson R., Granstrom B., and Gauffin J. A three-parameter voice source for speech synthesis. *Speech Communication*, 2:235–243, 1975.
- [12] Klatt D. H. Software for a cascade/parallel formant synthesiser. *Journal of the Acoustical Society of America*, pages 971–995, 1980.

- [13] Fant G., Lin Q. G., and Gobl C. Notes on glottal flow interaction. QPSR 2-3, 21-45, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, 1985.
- [14] Combescure P., Le Guyader A., Jouvét D., and Sorin C. Le traitement du signal vocal. *Ann. Télécommun.*, 50(1):142-164, 1995.
- [15] Mermelstein P. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, pages 1070-1082, 1973.
- [16] Coker C. H. A model of articulatory dynamics and control. *Proceedings of the IEEE*, 93:452-459, 1976.
- [17] Larar J. N., Schroeter J., and Sondhi M. M. Vector quantization of the articulatory space. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pages 1812-1818, December 1988.
- [18] Bailly G., Castelli E., and Gabioud B. Building prototypes for articulatory synthesis. In *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, pages 9-11, 1994.
- [19] Titze I. R. The human vocal cords: a mathematical model. *Phonetica*, pages 1-21, 1974.
- [20] Titze I. A four-parameter model of the glottis and vocal fold contact area. *Speech Communication*, 8:191-201, 1989.
- [21] Harris C. M. A study of the building blocks of speech. *Journal of the Acoustical Society of America*, 25:962-969, 1953.
- [22] Peterson G. E., Wang W., and Sivertsen E. Segmentation techniques in speech synthesis. *Journal of the Acoustical Society of America*, 30:739-742, 1958.
- [23] Beutnagel M., Conkie A., Schroeter J., Stylianou Y., and Syrdal A. The AT & T Next-Gen TTS system. In *Joint Meeting of ASA, EAA and DAGA*, Berlin, Germany, March 1999.
- [24] Beutnagel M., Conkie A., and Syrdal A. K. Diphone synthesis using unit selection. In *The 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, NSW, Australia, November 1998.
- [25] Itakura F. and Saito S. Analysis-synthesis telephony based on the maximum likelihood method. *Proceedings of the 6th International Congress on Acoustics*, 1968. paper C-5-5.

- [26] Atal B. S. and Hanauer S. L. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, pages 637–655, 1971.
- [27] Markel J. D. Digital inverse filtering: a new tool for formant trajectory estimation. *IEEE Transactions on Audio and Electroacoustics*, pages 129–137, 1972.
- [28] Makhoul J. Spectral analysis of speech by linear prediction. *IEEE Transactions on Audio and Electroacoustics*, AU-21:140–148, 1973.
- [29] Olive J. P. and Spickenagle N. Speech resynthesis from phoneme related parameters. *Journal of the Acoustical Society of America*, pages 993–996, 1976.
- [30] Atal B. S. and Remde J. R. A new model of LPC excitation for producing natural-sounding speech at low bit rates. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 614–617, 1982.
- [31] Charpentier F. and Moulines E. Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Proceedings of EUROSPEECH*, 2:13–19, September 1989.
- [32] Moulines E. and Charpentier F. Pitch synchronous waveform processing techniques for text-to speech synthesis using diphones. *Speech Communication*, 9:453–467, December 1990.
- [33] Moulines E. and Laroche J. Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 16:175–206, February 1995.
- [34] Dutoit T. High quality text-to-speech synthesis: a comparison of four candidate algorithms. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1:565–568, April 1994.
- [35] Moulines E., Emerard F., Larreur D., Le Saint Milou J. L., Le Faucheur L., Marty F., Charpentier F., and Sorin C. A real-time french text-to-speech system generating high quality synthetic speech. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 309–312, 1990.
- [36] McAulay R.J. and Quatieri T.F. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34(4):744–754, August 1986.

- [37] Quatieri T.F. and McAulay R.J. Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, 40(3):497–510, March 1992.
- [38] George E.B. and Smith M.J.T. A new speech coding model based on a least-squares sinusoidal representation. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 1641–1644, April 1987.
- [39] George E. B. *An analysis-by-synthesis approach to sinusoidal modeling applied to speech and music signal processing*. PhD thesis, Georgia Institute of Technology, November 1991.
- [40] George E.B. and Smith M.J.T. Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. *IEEE Transactions on Speech and Audio Processing*, 5(5):389–406, September 1997.
- [41] Macon M.W. *Speech synthesis based on sinusoidal modeling*. PhD thesis, Georgia Institute of Technology, 1996.
- [42] Macon M.W. and Clements M.A. Speech concatenation and synthesis using an overlap-add sinusoidal model. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1:361–364, May 1996.
- [43] Serra X. and Smith J.S. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–23, 1990.
- [44] Laroche J., Stylianou Y., and Moulines E. HNS: Speech modification based on a harmonic + noise model. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2:550–553, April 1993.
- [45] Syrdal A., Stylianou Y., Garrison L., Conkie A., and Schroeter J. TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 273–276, 1998.
- [46] McAulay R.J. and Quatieri T.F. Magnitude-only reconstruction using a sinusoidal speech model. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 27.6.1–27.6.4, April 1984.

- [47] Stylianou Y., Laroche J., and Moulines E. High quality speech modification based on a harmonic + noise model. *Proceedings of EUROSPEECH*, pages 451–454, September 1995.
- [48] Hedelin P. A tone-oriented voice-excited vocoder. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, page 205, 1981.
- [49] Hedelin P. A representation of speech with partials. In Carlson and Granstrom B., editors, *The Representation of Speech in the Peripheral Auditory System*. Elsevier Biomedical Press, 1982.
- [50] Ohm G.S. Über die definition des tones. *Ann. Phys. Chem*, 59:513, 1843.
- [51] Schroeder M. R. Models of hearing. In *Proceedings IEEE*, volume 63, page 1332, September 1975.
- [52] Cox R. C. and Robinson D. M. Some notes on phase in speech signals. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, page 150, April 1980.
- [53] Serra X. *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition*. PhD thesis, CCRMA Department of Music, Stanford University, October 1989.
- [54] Griffin D. W. and Lim J.S. Multiband excitation vocoder. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36:1223–1235, August 1988.
- [55] Quatieri T.F. and McAulay R.J. Speech transformations based on a sinusoidal representation. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 13.5.1–13.5.4, 1985.
- [56] Quatieri T.F. and McAulay R.J. Speech transformations based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34(6):1449–1464, December 1986.
- [57] McAulay R.J. and Quatieri T.F. Phase modeling and its application to sinusoidal transform coding. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 1713–1715, April 1986.

- [58] Quatieri T.F. and McAulay R.J. Mixed-phase deconvolution of speech based on a sine-wave model. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 649–652, April 1987.
- [59] Quatieri T.F. and McAulay R.J. Pitch estimation and voicing detection based on a sinusoidal model. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, April 1990.
- [60] Paul D.B. The spectral envelope estimation vocoder. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pages 786–794, August 1981.
- [61] Pollard M.P., Cheetham B.M.G., Goodyear C.C., and Edgington M.D. Shape invariant pitch and time-scale modification of speech by variable order phase interpolation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 919–922, 1997.
- [62] Stylianou Y. Removing phase mismatches in concatenative speech synthesis. In *The 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, NSW, Australia, November 1998.
- [63] O'Brien D. and Monaghan A. I. C. Shape invariant time-scale modification of speech using a harmonic model. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 381–384, Phoenix, Arizona, March 1999.
- [64] O'Brien D. and Monaghan A. I. C. Shape invariant pitch modification of speech using a harmonic model. *Proceedings of EUROSPEECH*, 3:1059–1062, 1999.
- [65] Alku P., Vilkmann E., and Laine U. K. Analysis of glottal waveform in different phonation types using the new IAIF-method. *International Congress on Phonetic Sciences*, 1991.
- [66] Stylianou Y. Concatenative speech synthesis using a harmonic plus noise model. In *The 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, NSW, Australia, November 1998.
- [67] Macon M.W. and Clements M.A. Speech synthesis based on an overlap-add sinusoidal model. *Journal of the Acoustical Society of America*, 97:3246, May 1995.

- [68] Black A. and Taylor P. The Festival speech synthesis system: system documentation. Technical report HCRC/TR-83, Human Communications Research Centre, University of Edinburgh, Scotland, UK, January 1997.
- [69] Black A. and Taylor P. CHATR: A generic speech synthesis system. In *Proceedings of Coling 94, the 15th International Conference on Computational Linguistics*, pages 983–986, Kyoto, Japan, 1994.
- [70] Black A. CHATR, version 0.8, a generic speech synthesis. System documentation, ATR - Interpreting Telecommunications Laboratories, Kyoto, Japan, March 1996.
- [71] Hunt A. and Black A. Unit selection in a concatenative speech synthesis system using a large speech database. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1:373–376, 1996.
- [72] Stylianou Y., Dutoit T., and Schroeter J. Diphone concatenation using a harmonic plus noise model of speech. *Proceedings of EUROSPEECH*, pages 613–616, 1997.
- [73] Page J. H. and Breen A. P. The Laureate text-to-speech system: Architecture and applications. *British Telecom Technology Journal*, 14(1), January 1996.
- [74] Banga E. R. and García-Mateo C. Shape-invariant pitch-synchronous text-to-speech conversion. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1:656–659, May 1995.
- [75] Banga E. R., López-Gonzalo E., and García-Mateo C. A text-to-speech system for spanish with frequency domain prosodic modification algorithm. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, II:183–186, April 1993.