Amendments to thesis titled
Pitch Modification Techniques for Sampled Voice

- Chapter 2, pp19, 20, 21, 23, figure captions: the Hanning window is a misnomer and should be called the Hann window after von Hann who proposed its use in Astronomy.
- Chapter 4, p33: A more comprehensive literature review could include a search of patent references. This would include some of the commercial pitch shifters and time scalers in the review. An example is A.M Agnello, "Pitch changer with glitch minimizer", US patent 4464784, 1984.
- Chapter 4, p33: Note that the literature review does not include any analog pitch shifting techniques such as the "Springer Machine".
- Chapter 4, p33, last paragraph: Time domain techniques may, of course, be quite computationally expensive due to, for example, repeated autocorrelation searching. This complexity is still less than that of the frequency domain algorithms presented herein, such as the iterative approach of Griffin and Lim [2].
- Chapter 4, p40, last paragraph: In order to achieve a real signal in the time domain, it is necessary that the magnitude and phase have certain symmetry properties. These cannot be imposed directly on an arbitrary MSTFT without destroying the formant structure of the synthesis signal and hence losing the individuality and recognisability of the voice signal. The search phase is necessary to find a real signal with both the required frequency domain characteristics and a formant shape preserved from the original signal.
- Chapter 6, p82, last paragraph: Shepard tones need not use constantly varying pitches. It is possible to make a 12 tone equal tempered version of the Shepard illusion.
- Chapter 6, p82, last paragraph: The suggestion of using Schroeder's "tricks" to pitch shift speech is problematic because they require special fractal waveforms.
- Chapter 6, p83, first paragraph: The final sentence should read "It seems to be self evident that in performing high quality pitch modification, ...".

Michael Brooks

September 1999

Pitch Modification Techniques for Sampled Voice

A thesis submitted for the degree of Master of Engineering

Michael Brooks

B.E. (Honours), University of Sydney

Telecommunications Engineering Group

Research School of Information Sciences and Engineering

The Australian National University

December 1998

Declaration

The contents of this thesis are the result of original research and have not been submitted to any other university or institution for the purpose of obtaining a postgraduate degree. The direction of the research presented herein has benefited greatly from the guidance offered by my supervisors, Dr. Rodney A. Kennedy and Dr. Bob Williamson.

M. J. Brooks.

Michael John Brooks
10 December 1998

Abstract

Altering the way someone sounds when singing or speaking has been the subject of increasing scrutiny over the past twenty years. Modification algorithms based on transforms such as the Discrete Fourier Transform and the Cepstrum have been popular with academic researchers, but computationally efficient time domain techniques are most widely used in commercially available implementations. This thesis attempts to orientate the reader with respect to the disparate array of proposed methodologies, and considers some new procedures for varying the pitch of human singing and speech.

One of the most widely sought after modifications to speech is that of altering the rate of articulation without reducing the intelligibility or the personal characteristics of the original speaker. Conversely, when modifying the pitch of a human voice, either singing or talking, these more nebulous personal characteristics, sometimes referred to as "naturalness", are often more disturbed than the intelligibility. Published work revolves around the Phase Vocoder, which models the production apparatus of the human species, but it is found that attention to subjective evaluation is lacking.

Two new techniques for modifying the pitch of a voice by means of low complexity time domain structures are proposed, developed and implemented. The first is based on a filter bank interpretation of the Fourier Transform, implemented in the time domain as a set of FIR filters. Several methods for generating the set of filters are compared and the best performing method is incorporated into a system for the real-time transformation of sung pitch by a time-varying rate. The proposed system is computationally efficient and capable of acceptable quality, especially when compared to techniques from the literature.

Another pitch modification technique is suggested by the methods used to encode pitch information in speech coding systems. An implementation based on linear prediction filtering and autocorrelation-based pitch extraction is developed and found to be of insufficient quality in comparison to the filterbank structure.

Table of Contents

Chapter 1 - Introduction

1.1. Thesis Scope

Devices designed to modify the perceived characteristics of the human voice have been commonplace in the entertainment industry for centuries. It is only in relatively recent years that science and engineering have attempted to understand the mechanics of production, the limits of intelligibility and, in the modern era, the psychology of perception of modified voice signals.

Two manifestations of the human voice are considered in this thesis: speech and singing. Samples of both are subjected to modifications in playback rate, and to modifications in pitch. For the first vocal phenomenon considered, that of speech, the ability for a listener to comprehend the meaning of speech subjected to a modification is considered most important. Singing, the other vocal exercise examined, attaches less importance to the intelligibility of the treated voice and more to the perceived pitch of the sung note. Nevertheless, assessments of quality are inherently subjective.

A wide range of solutions have been proposed over the last twenty years to the practical implementation of time scaling and pitch modification of sampled signals, in particular speech. It is the intention of this thesis to assess the state of the field, and to propose several new algorithms for pitch scaling speech and singing. The proposed methods are compared with high quality algorithms from the literature. In particular, real-time, time-domain solutions exhibiting high quality are sought.

Both time-scaling and pitch-scaling of speech will be considered, with algorithms proposed only for the purpose of altering the perceived pitch of a voice. To this end, samples of human speech and singing are employed to assess the quality of the methods.

When the term "pitch shifting" is used, we are referring to the modification of a voice waveform whereby the fundamental perceived pitch is changed without altering the

temporal features of the signal. Here it is desirable to retain the intelligibility, the formant structure, to some degree, but most applications of pitch modification are intended for use on sung vocal performances and so the naturalness, or prosody, of the voice is most important. So, too, is the relation between the vocal pitch and the notes of any accompanying instruments.

Some common applications of pitch scaling include correcting the pitching of singers in recording studios and karaoke machines which correct the singer as they sing. Other applications which have been suggested or are in development are systems designed to allow normal conversation between deep-sea divers in a helium atmosphere, improving the quality of speech synthesis systems such as text-to-speech systems and improving the perceived "humanness" of excitation systems for laryngectomy patients.

A complementary modification is time scaling the voice waveform. One of the problems associated with time scaling speech may be illustrated by the well known "chipmunk" effect which occurs when speech recorded at one speed is replayed at another, higher speed. The speech will sound higher in pitch and as the factor is increased, the intelligibility of the signal is severely degraded. To generalise, the task of time scaling speech is to alter the playback rate of a sampled waveform without altering the pitch content of the signal and without degrading its intelligibility.

Some typical applications of time-scaling algorithms include synchronising dubbed dialogue in video/film production and editing, and adaptive rate changes for speech for use with variable bit-rate communications channels. Retaining the intelligibility in high speed play back of recordings of dictation, lectures or voicemail messages, and bit rate reduction in telecommunication transmission are two other uses to which time scaling systems have been put.

Two notions affecting the subjective quality of a modified vocal waveform have been developed here; the first being that changes in pitch affect the "naturalness" of the speech, the second being the preservation of the intelligibility of the speech. The former characteristic, the pitch, may be viewed as the *fundamental frequency* of the signal and its *harmonics*. The latter feature, the *formant structure*, arises from the *spectral*

envelope of the signal. These two effects are characterised in the most common model of speech production discussed briefly in Section 2.

We shall concentrate our efforts on investigating real-time time-domain techniques for altering the perceived fundamental frequency of a sung vowel. The sung vowel is a distillation of that portion of human speech from which the pitch is determined by the listener's brain. These voiced sections are modelled by sampling a human male singing the 5 long vowels at discrete pitches. These samples are used to assess the efficacy of a method by comparing the modified pitch-scaled signal with a sample of the same singer singing naturally at the target pitch.

Many of the algorithms which have been proposed in the literature for performing these modifications are, as we shall see, not based explicitly on the assumption that the signal is human speech. Even in cases where this assumption is made, the model of speech production can quite adequately serve as a model for other audio sources, such as musical instruments. The speech production model consists of an excitation source concatenated with a slowly time-varying filter, and is the basis of many speech processing algorithms in coding, speech recognition and other fields. It can, with little or no modification, be seen to model single instruments such as the guitar (strings as excitation, body as filter) quite well. Modelling the sounds of multiple instruments, known as polyphonic signals, can be far more demanding, however.

The next section provides a more detailed overview of the structure of the thesis as a whole.

1.2. Thesis Overview

After this introductory chapter, a background chapter details some basics of speech production models, and analysis and synthesis systems. A formalised definition of time and pitch scaling follows in Chapter 3, while Chapter 4 presents a thorough review of the previously published work in the field. Chapter 5 develops several new techniques for pitch scaling sung vowels and speech, supported by analysis and the author's

implementations. Specific aims and contents will be detailed next, and Section 1.3 which follows has a point form summary of the main contributions of this thesis.

Chapter 1 - Introduction

Thesis Scope

A brief explanation of what time and pitch scaling of speech is. Some examples of applications for both modification methods. An explanation of the motivation behind the project and some of its goals.

Thesis Overview

The structure of the thesis explained. The section you are reading now.

Thesis Contributions

The main contributions of the thesis.

Chapter 2 – Background

Models of Speech Production

The most commonly used speech production model, the source/filter model, is introduced. Some example snippets of speech waveform are examined with reference to the model. The peculiarities of various methods of examining speech signals are noted; an example being the effect of the shape of the window used to segment the speech. A discussion of various examples of the frequency response of the vocal tract filter and the excitation waveforms. Introduction to the quasi-stationary assumption.

The Short Time Fourier Transform

Explanation of the Short Time Fourier Transform, or STFT, including discussion of appropriate choices for frame length and frame shift. This Time-Frequency representation is the most frequently used theoretical basis for representing sampled speech waveforms. This thesis proposes alternative analysis systems which do not suffer some of the limitations of the STFT technique.

Time Frequency Representations

Discussion of the general form of joint time-frequency transforms, of which the STFT is a good example. It has been suggested that by making the transform itself adaptive to the local characteristics of the signal the uncertainty inherent in the use of the STFT can be minimised. By developing this technique, the validity of the quasi-stationary assumption may be compromised. In this case, the STFT, which treats the sampled speech as a time series of independent, one-dimensional spectral snapshots is inappropriate.

Deconvolution

An exposition of various methods for deconvolving the vocal tract filter response from the excitation waveform. Linear prediction and its uses in speech coding. Homomorphic deconvolution and the cepstrum. Improved cepstral liftering. Pitch scaling is usually carried out on the excitation waveform alone.

Issues

A recap of the preceding sections main points. A block diagram of the general form for speech processing implementations. Particular attention is given to the similarity of speech coding methods to speech modification methods. Some of the implementation limitations and assumptions of the previous sections are discussed.

Chapter 3 - General Formulation of Speech Scaling

With acknowledgments to Moulines and Laroche [6], a general formulation of speech modification describing the various methods is developed along with a consistent notation.

Chapter 4 - Literature Review

Two main groups of published techniques are examined: those modification methods based on models of speech production and those based on no particular production model. Each sub-group is further classified into time-domain and frequency-domain implementations. A discussion of the evaluation methods used and their general inadequacy, with rare exceptions. A table classifying of all the reviewed papers. Two of the methods were implemented as part of this research. References to sampled sounds on the accompanying disk.

Chapter 5 - An Investigation into Alternative Scaling Methods

Filter Banks

The general case of a bank of uniformly spaced filters, used to isolate signals which may be modulated to alter their frequency response. The approach is refined to improve the quality, but requires explicit extraction of pitch from the original. The results of these experiments are included on the accompanying disk. Finally, a complete pitch-modification system is presented and discussed. References sampled waveforms and source code on the accompanying diskette.

Adaptive Codebook Modification

Inspired by the method of pitch extraction in the speech-coding field, a method is assessed using the Long Term Predictor technique in conjunction with Linear Predictive Coding to modify the pitch of the excitation. This method proves less successful than the filter bank approach but yields some useful insight into the use of the LPC method of deconvolving source from filter.

Chapter 6 – Summary and Conclusion

The similarities and differences between the modification of speech and singing are examined to assess the limitations of the results achieved. The use of model based modification techniques are compared against methods not based on a model of signal production. The issue of subjectivity is mentioned in passing.

1.3. Thesis Contributions

We list the main contributions of this thesis:

- A comprehensive review of the available publications relating to time and pitch scaling voice signals is reported. Several of the published techniques have been implemented by the author, and the results of applying these algorithms to various input sentences and sung vowels are included on an accompanying disk and used in subjective comparisons with the new techniques which are introduced later.

- The filterbank interpretation of the Fourier Transform is used as a basis for investigating time domain approaches to pitch scaling. Two techniques are developed for shifting discrete segments of the frequency domain by either filtering their complex envelopes, and shifting each to a new target centre frequency or by using Single Side Band modulation directly on each sub-band output from the filter bank.

- It is discovered that tracking the pitch of the source is essential to ensuring a subjectively acceptable result at pitch scaling factors between ½ and 2. This is confirmed with reference to the variety of published techniques, most of which embed a pitch extraction operation explicitly or implicitly.

- A new high quality time domain system is proposed based on the use of a deconvolution system delivering the excitation component to a filterbank system which pitch scales the excitation by a time-varying amount subject to the input pitch and a target pitch.

Chapter 2 – Background

2.1. Introduction

The purpose of this background chapter is to demonstrate a simple model of speech production and to examine how the various components and parameters of a segment of speech may be obtained using frequency domain and cepstrum techniques. This material forms the basis for the discussion of previously published voice modification techniques which follows in Chapter 4, and covers the range of assumptions which underpin the new work in Chapter 5.

Many of the concepts in this chapter will be familiar to readers, but it is suggested that at least the final section of this chapter, Section 2.6, be read in order to understand some of the distinctions between the various published methods, and also the directions taken in developing new pitch modification techniques.

2.2. Models of Speech Production

We begin by outlining the model of the physical aspects of speech production used most often in signal processing applications. Basically, the vocal tract forms an acoustic transmission system which, during speech production, is subjected to at least two different forms of excitation signals. These are specifically

- Voiced sounds produced by the periodic opening and closing of the glottis. These are commonly vowel sounds such as the "a" in "sand". These are the sounds responsible for the perceived pitch of a signal.
- Non-voiced sounds produced by constriction or closing off of the vocal tract with the resulting variations in air flow producing noise-like excitations. These are sounds such as "s" and "t" and are also known as sibilants.

As the vocal tract changes shape relatively slowly with time, it can be modelled as a slowly time varying linear filter, with characteristics of which are stationary for periods

up to 20 ms typically. The vast majority of the literature treats speech in this "quasi-stationary" manner. This model of speech production is illustrated in Figure 2-1 below.

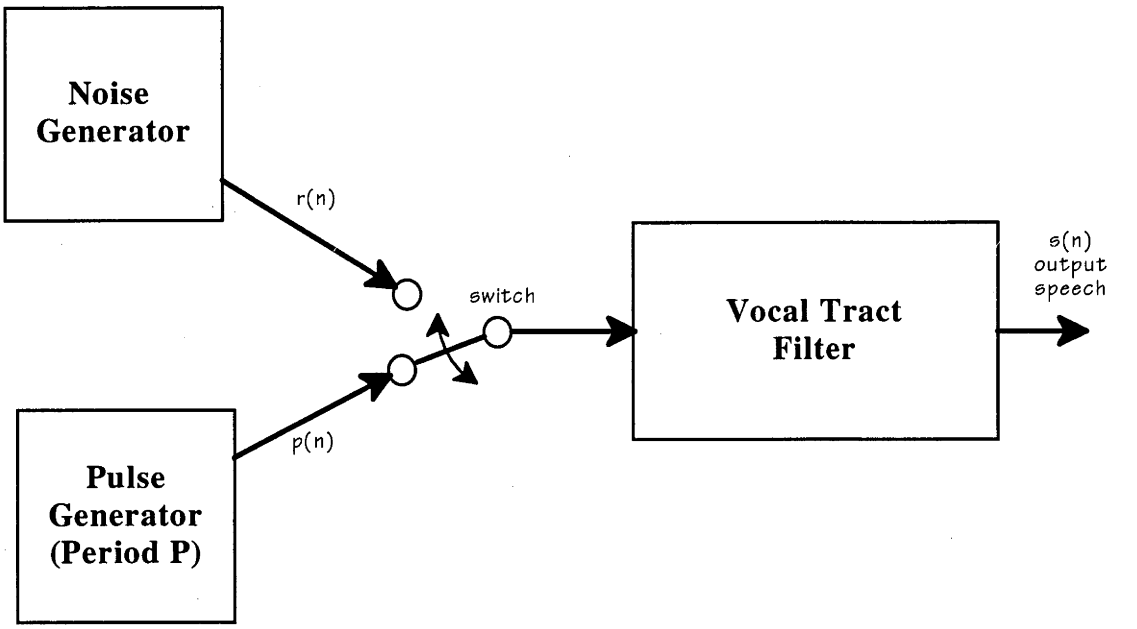

Figure 2-1: The most basic model of speech production process. A time-varying vocal tract filter is alternately excited by either a noise source (for sibilants) or a pulse generator (for vowels).

If a sampled speech signal is of a short enough time interval to be considered stationary but is still long enough to contain sufficient pitch information, the entire signal can be represented by the following convolution:

$$s[n] = v[n] * p[n] \text{ for } 0 \le n \le L-1 \qquad (2.1)$$

where L is the segment length in samples, $v[n]$ is the impulse response of the vocal tract and $p[n]$ is either the periodic, pulsed excitation of a voiced speech segment or the random noise excitation of unvoiced speech.

Throughout this thesis, the emphasis is on the voiced portion of speech or singing. This is because of the importance of this element of the signal to the human perception of pitch. No effort is expended on the consideration of methods of voiced/non-voice classification or extraction. In many cases, treating the non-voiced segment as though it

were voiced results in only minor artefacts. The most important of these is the introduction of a periodic "hum" in aperiodic, noisy sections of the speech. This issue is found predominantly in segmented-in-time techniques and is discussed later.

2.3. The Short Time Fourier Transform

In a pair of fundamental papers in 1981, Portnoff [23, 24] laid the theoretical basis for the representation of a speech signal in terms of its short-time Fourier transform. Starting with the speech production model outlined above, he developed a relationship between the parameters of a speech waveform and the Fourier transform of the speech. The quasi-stationary assumption is again applied and the vocal tract is modelled as a linear, time-varying transfer function.

Rather than assuming a time invariant filter, the vocal tract is characterised by its time-varying unit-sample response, $t(n,m)$, which is defined as the response of the system at time n to a unit sample applied m samples before, at time $(n-m)$.

Defining the Fourier transform of this sample response, with respect to the second index, m, and with the symbols for the independent frequency variable, ω, and the square root of minus one, j, taking their usual places, to be

$$T_2(n,\omega) = \sum_{m=-\infty}^{\infty} t(n,m)e^{-j\omega m} \qquad (2.2)$$

Portnoff showed that voiced segments of speech, $x(n)$, can be represented as a linear combination of harmonically related complex exponentials. In equation (2.2), the subscript 2 is used to denote that $T_2(n,\omega)$ is a partial Fourier transform with respect to the second argument. The time variation of the filter response corresponds to the dependence of the two functions $t(n,m)$ and $T_2(n,\omega)$ on the index n.

If we define the local pitch period of a pulsed excitation signal in the neighbourhood of n to be $P(n)$, then using the convolution of equation (2.1), Portnoff showed how quasi-stationary voiced speech may be represented as a sum of complex exponentials, thus

$$x(n) = \sum_{k=0}^{P(n)-1} c_k(n) e^{jk\phi(n)} \qquad (2.3)$$

with coefficients $c_k(n)$ given by

$$c_k(n) = \frac{1}{P(n)} T_2(n, k\Omega(n)) e^{jk\phi_0} \qquad (2.4)$$

such that the terms $\Omega(n)$ and $\phi(n)$ are functions depending only on $P(n)$, the instantaneous, or local pitch period.

A non-voiced speech waveform is represented by its "time-varying power spectrum",

$$S_x(n,\omega) = \sigma_u^2 \left| T_2(n,\omega) \right|^2 \qquad (2.5)$$

where the σ_u is the second moment of the zero-mean stationary white noise process modelling the non-voiced excitation waveform. Portnoff showed that the time varying power spectrum can be written in a Fourier transform pair with the autocorrelation function R_x of the speech as follows,

$$S_x(n,\omega) = \sum_{\tau=-\infty}^{\infty} R_x(n,\tau) e^{-j\omega\tau} \qquad (2.6)$$

The short-time Fourier transform (STFT) of a discrete signal, can then be defined as the sum

$$X_2(n,\omega) = \sum_{m=-\infty}^{\infty} h(n-m)x(m)e^{-j\omega m} \qquad (2.7)$$

where *x(n)* is the voiced speech signal from (2.3) and *h(n)* is the analysis window restricting the input speech signal to a segment of limited duration over which the signal can be considered stationary. Portnoff further demonstrated that the voiced and unvoiced representations presented above in equations (2.3) and (2.5) can be extracted directly from the STFT. Importantly, he uses both narrow- and wide-band approaches to the analysis of speech with respect to the bandwidth of the window transfer function. This generalises the effect of the window transfer function bandwidth to cases where it is both greater and less than the fundamental pitch frequency of the speech. He noted that voiced and unvoiced speech exhibit similar narrow-band instantaneous amplitude and frequency components, but finds that the major difference between the two speech modes is the different underlying harmonic structures. In the unvoiced case there is a lack of discernible form as opposed to the regular inter-harmonic maxima found in voiced speech.

This method of STFT-based analysis of speech is often used to justify the use of the STFT in a variety of speech processing applications including speech analysis/synthesis systems such as phase vocoders, channel vocoders and sub-band coders [2, 3, 29, 33, 36]. It also provides a mathematical model of the speech spectrogram, a widely used tool in speech analysis. There are many other time-frequency representations outlined in the literature, but almost all of them are based on Portnoff's work on the use of the STFT. This class of systems we shall refer to henceforth as *segmented in time* systems.

2.4. Time Frequency Representations

In his book, Riley [25], takes an even more general approach to time-frequency representations of speech than does Portnoff. In fact, he calls into question the validity of the quasi-stationary assumption for speech. Riley demonstrates that speech is not always quasi-stationary, even during voiced segments.

The short-time Fourier transform has fundamental limitations as discussed by de Bruijn [27]; these limitations are known as the *uncertainty principle*. The basis of the argument is that a sampled waveform must have a long enough duration for adequate frequency resolution, but be short enough to allow adequate time resolution. Namely, given the Fourier transform pair

$$h(x) \Leftrightarrow H(\omega) \tag{2.8}$$

if the variances of the two sequences are

$$\begin{aligned} \mathrm{var}|h(x)|^2 &= (\Delta x)^2 \\ \mathrm{var}|H(\omega)|^2 &= (\Delta \omega)^2 \end{aligned} \tag{2.9}$$

then the variances are related by the inequality

$$\Delta x \Delta \omega \geq \frac{1}{2} \tag{2.10}$$

All of this relates directly to the quasi-stationary assumption for speech in that the segment used for the STFT must be short enough so that the vocal tract response is stationary, yet long enough to contain adequate pitch information.

In fact, there are classes of signals for which no window length for the STFT is adequate. One set of such signals are called chirps, and are of the form

$$x(t) = e^{\frac{jmt^2}{2}} \qquad (2.11)$$

where the modulus, *m*, causes the instantaneous frequency to increase linearly with time. For sufficiently large m, the quasi-stationary assumption breaks down. In fact, the human voice can and does produce sounds of this nature.

Riley also makes use of neuro-physiological work which has found that a large amount of auditory nerve endings in the mammalian cochlear do not respond optimally to continuous tones but instead to a swept tone within particular range of frequency slope. Linguistics has also shown that the movement of formant frequencies are important - their transitions and discontinuities contain important semantic content. That is, the human ear exhibits a directional resolution, which varies with time and pitch frequency.

Thus, in many applications the approach of Portnoff, treating the sampled speech as a time series of independent, one-dimensional spectral snapshots is inappropriate. Instead we should consider any analysis of speech as a joint time-frequency representation of the signal.

The spectrogram is a widely used tool in speech analysis, and may be visualised as a time sequence of STFTs. We can express the spectrogram as a joint time-frequency transform of a voiced speech signal, *x(t),* which has been segmented in the time domain by a window function *w(t).*

$$S_x(t,\omega) = \left| \int_{-\infty}^{\infty} w(\tau)x(t+\tau)e^{-j\omega\tau}d\tau \right|^2 \qquad (2.12)$$

Because of the limitations of the spectrogram, namely the implication of the uncertainty relation on the simultaneous time and frequency resolution, solutions are sought for other representations which may better approach the optimal resolution.

If we define the marginals of an arbitrary signal representation $F_x(t,\omega)$ to be

$$\pi_1(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F_x(t,\omega)d\omega$$
$$\pi_2(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F_x(t,\omega)dt \qquad (2.13)$$

then, by (2.9) and (2.10), "perfect" time and frequency resolution can be achieved provided these marginals, or projections, satisfy the following relations.

$$\pi_1(t) = |x(t)|^2$$
$$\pi_2(\omega) = |X(\omega)|^2 \qquad (2.14)$$

One time-frequency representation which satisfies the criteria of equation (2.14) is the Wigner distribution,

$$W_x(t,\omega) = \int_{-\infty}^{\infty} e^{j\omega\tau} x(t+\tau/2)x^*(t-\tau/2)d\tau \qquad (2.15)$$

Riley goes on to propose a whole class of representations, based on the Wigner distribution, which satisfy several criteria such as shift invariance and superposition. The reason for all this is to optimise the resulting time-frequency representation such that the transform is directionally localised, exhibiting better resolution in some directions of the time-frequency plane than others.

Speech transforms could, in principle, be made to behave optimally in the local direction of movement for particular formants so that they may be recognised more easily during feature extraction. The features concentrated on are time discontinuities and time frequency ridges, and the application is intended to be useful in speech recognition tasks.

The representations presented are of limited interest when considered for real-time analysis of speech. However, the discussion of the quasi-periodic nature of speech and the departure of the signal from this assumption are of direct impact on the suitability of segmented-in-time approaches to the analysis of speech.

2.5. Deconvolution

As promised, our analysis proceeds assuming that the segment of speech to be analysed is voiced. When the signal is segmented in time, the production model becomes invalid at the edges of the analysis interval because of pulses falling outside the interval. At the start of the window, the voice waveform may be non-zero, a boundary effect unlike any natural response of the human vocal articulators. The effect of this discontinuity can be lessened by applying a tapering window (*w[n]*) to the interval. Several different types of windows are commonly used, Hamming and Hanning windows being examples from the literature [29, 30].

From Oppenheim and Schafer [29], the window effect on *v[n]*, the vocal tract impulse response, is ignored because it (*w[n]*) varies slowly with respect to *v[n]*. The vocal tract response is considered stationary for the duration of the window. Then the windowed speech is effectively the convolution

$$x[n] = v[n] * p_w[n] \qquad (2.16)$$

where the windowed excitation is given by the superposition of the window *w[n]* on the pulse train *p[n]* with local pitch period *P(n)*,

$$p_w[n] = w[n]p[n] \qquad (2.17)$$

In order to achieve a characterisation of the sampled speech signal in terms of the vocal tract filter *v[n]* and the excitation signal *p[n]*, it is necessary to isolate one from the other. If we were to undo the convolution of (2.16) we could separate the filter response from the windowed excitation as long as the stationarity assumption holds over the window length. As will be seen later, many techniques exist for achieving this separation, including linear prediction methods. The discussion here follows the approach of Oppenheim and Schafer [29] in using a method known as homomorphic deconvolution to demonstrate the separation of the vocal tract response from its excitation.

Homomorphic deconvolution may be implemented by making use of the properties of the so-called cepstrum transformation. The cepstrum for a sequence *x[n]* with Fourier transform $X(e^{j\omega})$ is defined as

$$\hat{x}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log[X(e^{j\omega})]e^{j\omega n}d\omega \qquad (2.18)$$

Using the cepstrum defined (2.18), it can be shown that the convolution (2.16) may be mapped into an addition operation in the cepstrum domain. That is

$$\hat{x}[n] = \hat{v}[n] + \hat{p}_w[n] \qquad (2.19)$$

The analysis in Oppenheim and Schafer [29] goes on to show that for speech, the periodic excitation will show up as a peak in the cepstrum domain at the pulse interval period. This can then be isolated by subtraction. These concepts will be illustrated with an example.

Figure 2-2 following shows 256 samples of speech sampled at 8 kHz. The section represents a part of the "oa" voiced segment in the word "oak". The first plot shows the

original speech (or to use the parlance, rectangularly windowed speech), while the remaining plots demonstrate the application of two popular window weighting functions, the Hamming and Hanning windows.

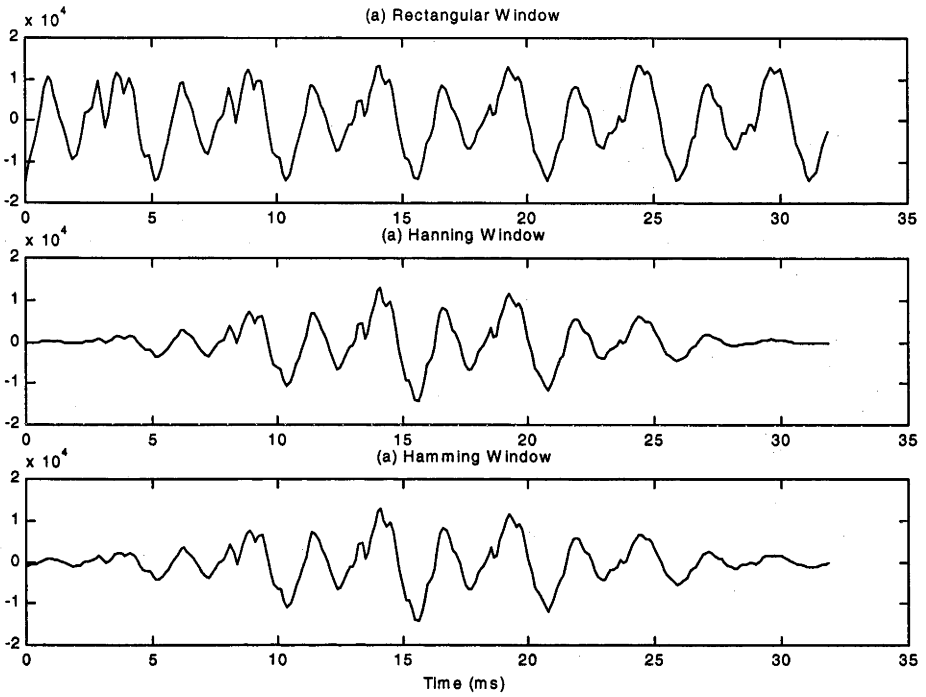

Figure 2-2: A voiced section of speech is shown subjected to various window functions. (a) is the original, rectangularly windowed speech waveform, (b) shows the effects of the Hanning window, while (c) is the result of a Hamming window. Note the zeroed sections, or pedestals, at either end of the Hanning window.

As can be seen in the original speech, the waveform appears to be periodic as expected for a voiced segment. The purpose of the weighting windows is to taper the speech segment so that the edge discontinuities have less effect when the discrete Fourier transform is calculated. Power spectra for three windows, the rectangular window, the Hanning window and the Hamming window are shown in Figure 2-3 below. Both the tapered windows can be seen to provide a clearer picture of the excitation pitch and its harmonics than in the rectangular window.

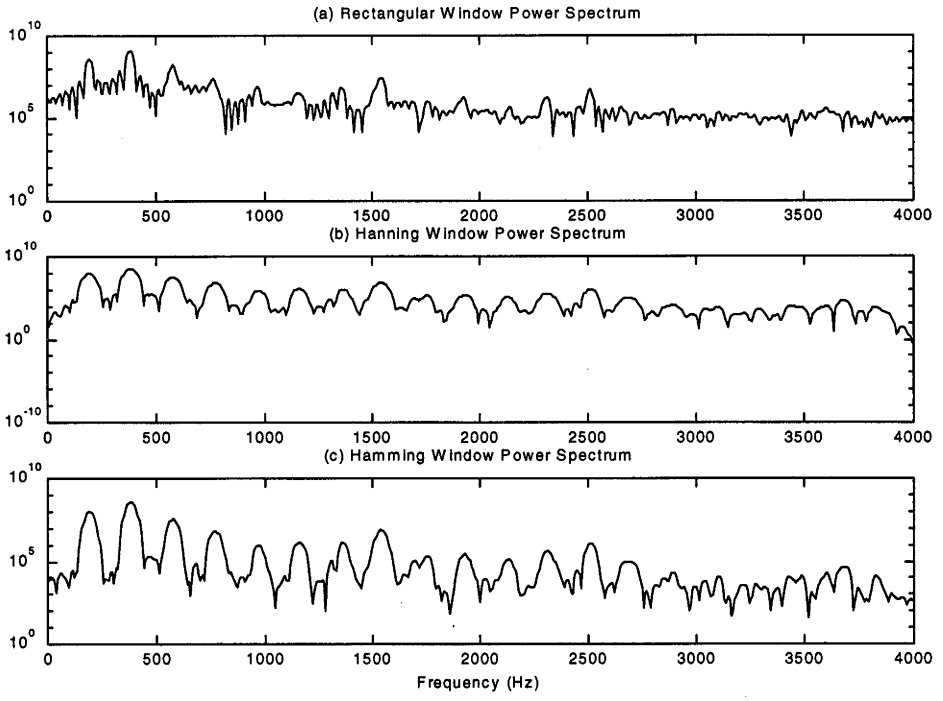

Figure 2-3: Power spectra for windowed speech from Figure 2-1. (a) shows the magnitude spectrum of the rectangular window, (b) is the Hanning window spectrum and (c) is the Hamming window. The inter-harmonic spacings show a fundamental pitch of ~200 Hz, corresponding to a 5 ms pitch period as demonstrated by the peak to peak time differences in Figure 2-2.

Leaving aside the discussion of the merits of one window shape over another, and continuing with the Hamming windowed example, the possibilities of decomposing a speech segment into the constituent parts of our model are now demonstrated. The complex log of the Fourier transform of the segment is calculated and the inverse DFT is formed.

The result is known as the cepstrum and its magnitude is shown in the first plot of Figure 1.4. Note the peak in the cepstrum at about 5 ms; this corresponds to the fundamental pitch in the voiced speech segment, which can be seen by comparing the peaks in the original signal at the top of Figure 2-2. The bulk of the information about the slowly varying vocal tract response *v[n]* is contained near the origin.

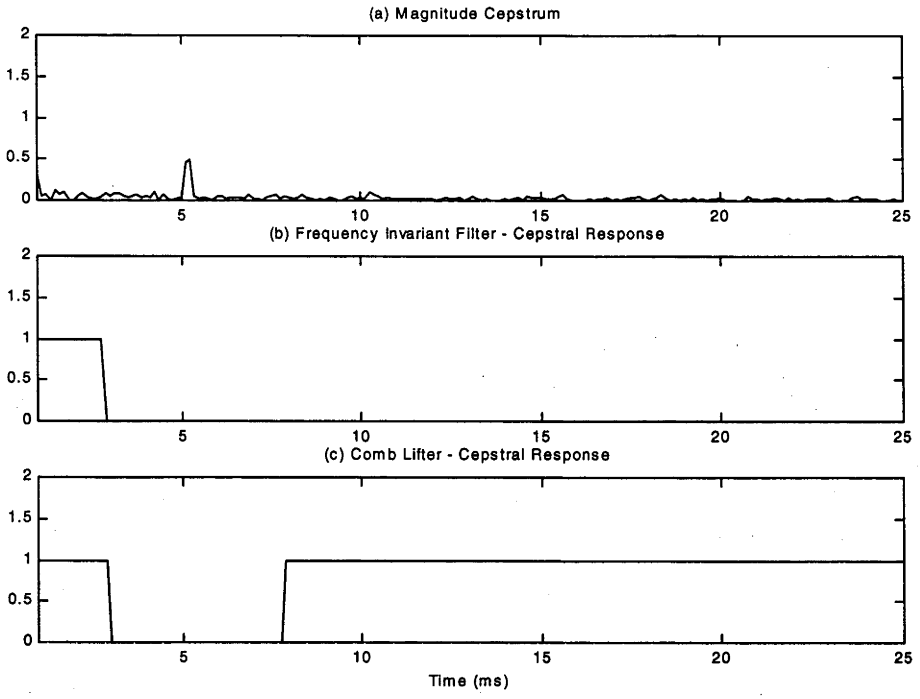

Figure 2-4: Magnitude cepstrum and lifters used in deconvolution. In (a), the Hanning windowed magnitude spectrum, from Figure 2-3 (c) has been transformed into the cepstrum domain. The lifter in (b), together with the complementary high pass lifter, is used to separate the 5ms pitch information (the peak) from the vocal tract response. The lifter proposed by Abe et al. ([20]) is shown in (c).

This leads to a method of separating the excitation from the vocal tract filter response by applying a mask such as the one shown in the middle plot of Figure 2-4. Similarly, the excitation may be obtained by applying the complement of the mask. If the Fourier transform of the modified cepstrum obtained from this "frequency-invariant filtering" is calculated, we get the magnitude spectra shown in Figure 2-5. The original windowed spectrum is plotted first followed by the frequency response of the vocal tract and finally, the frequency content of the excitation harmonics, or difference spectrum.

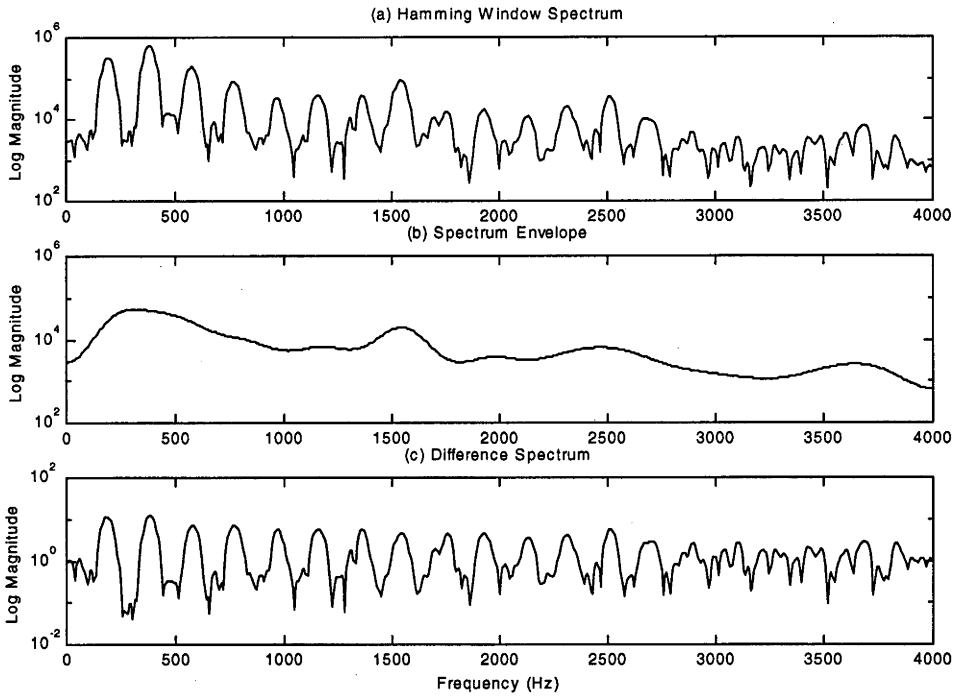

Figure 2-5: The spectrum of the Hamming windowed signal from Figure 2-3 (c) is repeated for comparison in (a). The "smoothed" vocal tract response is shown in (b) and the harmonic frequency content of excitation is shown in (c). The spectrum envelope and the excitation were separated by means of the lifter from Figure 2-4 (b).

The smoothed vocal tract response or spectrum envelope clearly shows the poles (or formants) of the filter characteristic with peaks at approximately 300, 1500 and 2500 Hz. These formants are considered to be the "carriers" of the information content of the speech segment. The semantic meaning of speech is contained in the movement (in the frequency domain) of the formants. The spectrum of the excitation displays a fundamental at ~200Hz (for the period of 5ms) with evenly spaced harmonics thereafter.

This simple cepstral mask or "lifter" and its results, shown in Figure 2-5, are optimised for obtaining the smoothed estimate of the spectrum envelope. This envelope is useful in calculating formant movement information in speech recognition or parameterisation tasks and forms the basis of the familiar spectrogram.

However, in tasks involving the fundamental pitch of the excitation waveform, a clearer estimate of the excitation spectrum is required. A method for implementing this is proposed by Abe et al.[21], as part of a scheme for modifying the fundamental pitch of sampled speech. They propose a comb lifter to extract the excitation spectrum and the form of this lifter is shown at the bottom of Figure 2-4. The results in the frequency domain of applying the comb lifter appear in Figure 2-6 with the extracted excitation spectrum shown in the bottom plot. The excitation spectrum appears to be almost sinusoidal, indicating that only the pitch frequency and its harmonics were extracted.

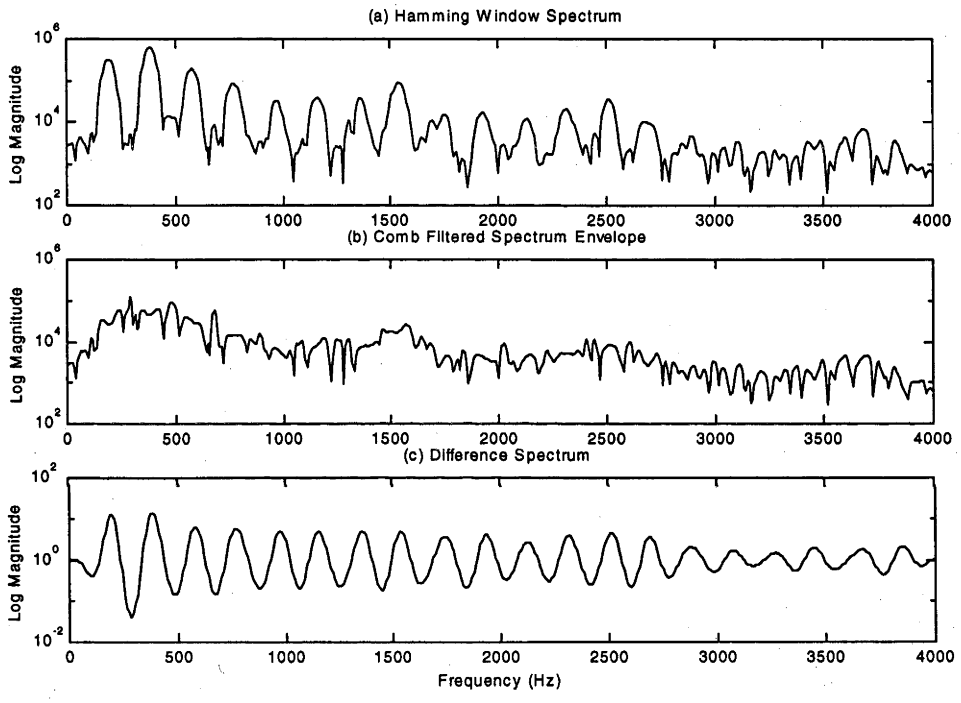

Figure 2-6: The Hanning windowed spectrum from Figure 2-3 (c) is again repeated for comparison purposes. The smoothed vocal tract response (b) and the frequency content of excitation (c) here were obtained with the comb lifter of Figure 2-4 (c).

2.6. Issues

Modification algorithms, similarly to other signal processing systems, are commonly presented in terms of the following simple model, known as the analysis/synthesis model. A preliminary analysis stage is followed by a modification step and finally the signal is resynthesised. This model is shown in block diagram form in Figure 2-7.

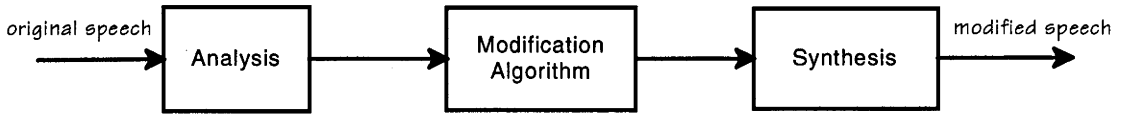

Figure 2-7: The analysis-synthesis model of speech processing. The initial analysis step usually involves the characterisation of the signal in terms of a production model. After modification of certain signal characteristics or parameters (like pitch), a synthesis step attempts to reproduce a "natural" sounding voice.

Typically, the analysis stage segments the signal in the time-domain into what are known as analysis frames. The modification is performed on these analysis frames individually and in the last step, the output signal is synthesised by some form of inverse of the analysis step. We describe implementations adopting this approach as having *segmented-in-time* architectures.

These terms and concepts are directly analogous to those in the speech coding analysis, transmission and synthesis model. Similarly to that case, we require that the signal should be stationary over a single analysis frame. This is often referred to as the quasi-stationary assumption. In general, both the vocal tract filter and the excitation source are assumed to be stationary over this period, often 20ms.

In some of the parametric approaches from the literature ([35, 36]), this requirement is relaxed and only the vocal tract filter (sometimes referred to as the Short Time Fourier Transform spectral envelope) is required to be stationary over an analysis frame.

In addressing an alternate pitch-shifting algorithm in Chapter 5, we shall look to a different view of the Fourier transform than the one presented above. Specifically, we shall examine the interpretation of a set of filters (or filter-bank) as calculating the frequency response of the signal on a sample-by-sample basis. These filters may be implemented in the time-domain, leading to a *segmented-in-frequency* architecture.

These alternate modification implementations are considered in Chapter 5, but first we present an analysis of the desired modifications in terms of the speech production model. We then examine the existing literature in order to understand the problems and advantages of the segmented-in-time approach.

Chapter 3 - General Formulation of Speech Scaling

3.1. Speech Production

In most engineering models of human speech production the speech waveform is represented as the convolution of a time-varying linear filter and an excitation signal. The excitation sequences used are different if we are considering the voiced or unvoiced parts, but we will concentrate on the voiced part as it is the most important section with regard to pitching and pitch perception.

For voiced speech then, the excitation is expressed as a sum of narrow-band signals with harmonically related instantaneous frequencies, usually complex exponentials with unit amplitude. The instantaneous fundamental frequency is given at time *n* by

$$F_f = \frac{2\pi}{P(n)} \tag{3.1}$$

where the local pitch period is *P(n)* and is varying slowly with *n*. The exponentials are summed to form the excitation thus:

$$e(n) = \sum_{k=0}^{P(n)-1} \exp[j(\phi_k(n))] \tag{3.2}$$

where the excitation phase of the *k*-th harmonic is defined as the integral of the time-varying harmonic frequency:

$$\phi_k(n) = \sum_{m=0}^{n} \omega_k(m) = \sum_{m=0}^{n} \frac{2\pi k}{P(m)} \quad (3.3)$$

The time-varying filter representing the vocal tract articulation is commonly defined in terms of the Fourier transform of its impulse response which are shown in polar form in equation (3.4).

$$G(n,\omega)\exp(j\psi(n,\omega)) \quad (3.4)$$

G(n,ω) and ψ(n,ω) are referred to as the time-varying amplitude and phase of the vocal tract system.

Given these definitions, the voiced speech segment may be shown to be

$$x(n) = \sum_{k=0}^{P(n)-1} G(n,\omega_k(n)) \times \exp[j(\phi_k(n) + \psi(n,\omega_k))] = \sum_{k=0}^{P(n)-1} G_k(n) \times \exp[j\theta_k(n)] \quad (3.5)$$

This form of expressing a speech signal is dependent on several assumptions to do with the convolution of time-varying signals. The primary assumption maintains that the signal may be considered stationary for short periods of time. As discussed in Section 5.2.5, this assumption holds true in most practical situations involving the human voice.

3.2. Time Scaling of Speech

The object of time scaling the human voice is to cause the rate of articulation to change without the spectral content being affected. Consider a time-varying relation between the time instants in the original signal, t, and the time instants in the time scaled version, t':

$$t' = D(t) = \int_0^t \beta(\tau) d\tau \qquad (3.6)$$

The second part of this is an integral definition based on the time-varying modification rate, $\beta(t)$. The relation amounts to specifying a mapping between time in the original signal and time in the modified signal. Where the modification rate is constant, $D(t)$ becomes simply βt.

Using this time scale in the parametric quantities of (3.5), we have the modified pitch contour, expressed in the modified time-scale, n', given by (3.7). $D^{-1}(.)$ denotes the inverse mapping from the modified time-scale back to the original time-scale.

$$P'(n') = P(D^{-1}(n')) \qquad (3.7)$$

The new pitch contour is a time-warped version of the original. Equation (3.8) shows the modified vocal tract magnitude response.

$$G_k'(n') = G(D^{-1}(n'), \omega_k(D^{-1}(n'))) \qquad (3.8)$$

The new magnitudes of the vocal tract system are also a time scaled versions of those in the original signal. Similarly, the filter phases, shown in equation (3.9), consist of temporally relocated copies of the original phases.

$$\theta'_k(n') = \phi'_k(n') + \psi(D^{-1}(n'), \frac{2\pi k}{P(D^{-1}(n'))}) \quad (3.9)$$

The instantaneous source frequencies are similarly translated in time.

$$\phi'_k(n') = \sum_{m=0}^{n'} \frac{2\pi k}{P(D^{-1}(m))} \quad (3.10)$$

For the simple case of time scaling by a constant factor, t, the system parameters reduce to the following set of equations.

$$P'(n') = P(n'/t) \quad (3.11)$$

$$G'_k(n') = G(n'/t, \omega_k(n'/t)) \quad (3.12)$$

$$\theta'_k(n') = \phi'_k(n') + \psi(n'/t, \frac{2\pi k}{P(n'/t)}) \quad (3.13)$$

$$\phi'_k(n') = \sum_{m=0}^{n'} \frac{2\pi k}{P(m/t)} \quad (3.14)$$

Throughout the remainder of the text, we consider the time scaling rate to be a constant and equal to β.

3.3. Pitch Scaling of Speech

The scaling of the pitch of a voiced signal is defined as altering the perceived fundamental frequency without affecting the spectral envelope or its time evolution. If we define a mapping of pitch periods in terms of a slowly time varying, always positive function $\alpha(n)$, such that

$$P'(n) = \frac{P(n)}{\alpha(n)} \tag{3.15}$$

then we have the local pitch frequency being increased if $\alpha(n) > 1$; the signal's pitch is decreased when $\alpha(n) < 1$. An ideal pitch scaling operation would require the modifications to the speech parameters as follows.

The pitch contour is scaled by the time-varying factor.

$$P'(n') = \alpha(n')P(n') \tag{3.16}$$

The amplitudes of the excitation harmonics are sampled at the shifted pitch harmonic frequencies

$$G'_k(n') = G(n', \alpha(n')\omega_k(n')) \tag{3.17}$$

$$\theta'_k(n') = \phi'_k(n') + \psi'(n', \alpha(n')\omega_k(n')) \tag{3.18}$$

Finally, the excitation harmonics are scaled by the pitch factor.

$$\phi'_k(n') = \sum_{m=0}^{n'} \alpha(m)\omega_k(m) \tag{3.19}$$

In scaling pitch, it may be seen that the vocal tract system amplitudes and phase must be recalculated at the new excitation frequencies and so must be estimated explicitly. This is in contrast to the time scaling case where the filter system response was unchanged.

If the scaling factor is a constant, p, then the model parameters become

$$P'(n') = pP(n') \tag{3.20}$$

$$G'_k(n') = G(n', p\omega_k(n')) \tag{3.21}$$

$$\theta'_k(n') = \phi'_k(n') + \psi'(n', p\omega_k(n')) \tag{3.22}$$

$$\phi'_k(n') = \sum_{m=0}^{n'} p\omega_k(m) \tag{3.23}$$

Throughout the remainder of the text, we consider the pitch scaling rate to be constant and equal to α. The symbol used to denote this constant pitch scaling factor will be p.

Chapter 4 - Literature Review

4.1. Introduction

A problem which presents itself in reviewing the literature in the audio processing field is that many of the methods in common use in recording studios are housed in commercial equipment. Because the methods themselves are considered to be the intellectual property of the commercialising party the techniques are rarely allowed into the public domain by way of journal publication or peer review.

One advantage obtained by commercial developers is the imperative placed on subjective evaluation of competing methods. It is the contention of this review of the academic literature that in most cases the assessment processes used are inadequate. It is, unfortunately, beyond the scope of this thesis to address an issue which would require a large group of subjects in double blind tests at this stage. These tests should be designed to use comparative scales such as those described in Thorpe and Sheldon [19] and make use of control groups and impartial test supervisors.

A tabulation of the most important papers from the literature is included below. Each contribution is presented chronologically with details of distinguishing criteria for comparison purposes. A discussion of the criteria used in the table follows.

The first criterion used is whether the method outlined makes explicit use of the Fourier Transform in some part of the algorithm. This can be a reasonable guide to complexity as time domain approaches tend to be of less complexity than those requiring a transform, due partly to the number of multiplications required by the transform. In addition, the need to invert the transform to return to the time domain once the modification is complete may require a computationally expensive search for a corresponding real signal in the time domain.

Method	Transform Domain	Speech Specific	Parametric or Non-Parametric	Polyphonic or Non-Polyphonic	Explicit Pitch Extraction
Malah [3] (1979)	Time Domain	Yes	Non-parametric	Monophonic	Yes
Portnoff [24] (1981)	Frequency Domain	Yes	Non-parametric	Monophonic	Yes
Seneff [33] (1982)	Frequency Domain	Yes	Non-parametric	Monophonic	Yes
Griffin & Lim [2] (1984)	Frequency Domain	Yes	Non-parametric	Polyphonic	No
Roucos & Wilgus [32] (1985)	Time Domain	Yes	Non-parametric	Polyphonic	No
Quatieri & McAulay [36] (1986)	Frequency Domain	Yes	Parametric	Polyphonic	No
Abe, Tamura & Kuwabara [20] (1989)	Cepstral Domain	Yes	Non-parametric	Monophonic	No
Lent [18] (1989)	Time Domain	No	Non-parametric	Polyphonic	Yes
d'Alessandro [1] (1991)	Frequency Domain	Yes	Parametric	Monophonic	Yes
Moulines & Charpentier [5] (1992)	Time & Frequency Domains	Yes	Non-parametric	Monophonic	Yes
Seiyama, Takagi, Umeda & Miyasaka [28] (1992)	Cepstral Domain	Yes	Non-parametric	Monophonic	Yes
Yim & Pawate [34] (1996)	Time Domain	No	Non-parametric	Polyphonic	Yes

A majority of the approaches in the literature use the Short Time Fourier Transform (STFT) as the basis for modifying both temporal and pitch related characteristics. The separation of the excitation source from the vocal tract filter may also be accomplished by means of a further transform to the so-called cepstral domain as in Abe et al. [20] and Seiyama et al. [28]. STFT techniques are also used in many parametric approaches to estimate the parameters of sinusoidal waveforms.

A problem caused by modifying waveforms in the STFT domain is that the resulting Modified Short Time Fourier Transform (MSTFT) may not correspond to a real-valued time-domain signal when the inverse transform is applied. Griffin & Lim [2] offer a solution (albeit an iterative and complex one), and theirs is the most cited paper in the field. Their solution to minimising a distance measure on the Short Time Fourier Transform Magnitude (STFTM) is guaranteed to converge to a, possibly non-global, minimum. Roucos and Wilgus [32] take the approach a step further, eliminating the need for an STFT at all. This is discussed in more detail below.

The second column in the table illustrates whether the proposed method is specific to speech signals or not. This is based on the observation of the use of the quasi-stationary source/filter model in the algorithm presented. Although this model may be valid for single instruments as well, the literature contains no conclusive demonstration of this. In only two cases was testing carried out on signals other than monophonic (single source) human speech: Lent [18], and Quatieri and McAulay [36].

A sub-class of the speech specific methods are those approaches using parametric models of the human speech production apparatus. These are those of d'Alessandro [1] and Quatieri and McAulay [36], as shown in column 3. Both papers employ similar parametric models, where speech is represented as the sum of sinusoids with slowly time-varying amplitudes and instantaneous frequencies. The parameters are found by use of the STFT and a rule based transformation performed on the parameters. Other models incorporating stochastic components also exist. A thorough comparison with non-parametric approaches is yet to be published.

The fourth column represents whether the proposed algorithm can support multiple speakers in the case of voice or complex music consisting of more than one instrument known as *polyphony*. The support for polyphony may be deduced from the model (or lack of it) assumed in the algorithm. Most treatments make extensive use of finding the parameters of a single source model (a *monophonic* source). The authors claims in this regard were also taken into account. However, as the results of Quatieri and McAulay [36] demonstrate, the effectiveness of an algorithm on polyphonic sounds may be difficult to deduce.

The final column explores whether the technique in question requires the fundamental harmonic of the excitation signal (hereafter referred to as the pitch of the signal) to be determined as part of the algorithm. Virtually all of the approaches use a fixed window for their time or STFT domain modifications and the pitch of the signal generally needs to be estimated during this frame. This leads to a resolution problem, where the frame needs to be long enough to cover several pitch periods, but short enough so that the pitch does not vary too much during the frame. As pitch varies quite a lot in one persons speech, and significantly between the sexes and different languages, this presents a significant problem to solve.

A better solution, it seems, is to dispense with the fixed frame altogether and, using the extracted pitch, create the analysis frame centred on the maximum signal (corresponding to the glottal pulse) and with a frame width equal to an integral multiple of the pitch period. These pitch synchronous methods (Moulines and Charpentier [5]) are quite low in complexity and are also high quality. They reduce artefacts caused by poor frame positioning with respect to temporal aspects of the speech signal but require accurate pitch calculation and frame placement with respect to the glottal maximum.

The more important contributions are now examined, beginning with those methods based explicitly on models of speech production. Section 4.2 considers those methods explicitly based on the model of speech production previously outlined. In section 4.3, methods which employ heuristic algorithms and do not rely on any assumptions about the physical signal production apparatus are examined.

4.2. Model Based Methods

4.2.1. Time Domain Implementations

4.2.1.1. Malah

In 1979, David Malah published the seminal paper for parametric methods in the time and pitch scaling fields. Although the method is non-parametric itself, it provides the basis for several later contributions (Quatieri and McAulay [36] and d'Alessandro [1]).

Malah demonstrated that by choosing STFT basis functions such that they form a bank of filters equally spaced in the frequency domain, the desired pitch scaling can be performed by scaling the instantaneous frequencies of the output of each filter and recombining to form the output signal. Time scaling may then be performed by altering the playback rate. Note that no separation of excitation and vocal tract characteristics is performed, and this certainly would affect the intelligibility of the output speech. It appears that no further work has been carried out on this technique, so further investigation is certainly warranted.

By employing several assumptions, such as FIR filter implementations and choosing a suitable windowing function, Malah developed an approximation to the parametric modification, which may be implemented wholly in the time-domain and requires only one multiplication and two additions per output sample. Malah reported only informal claims of acceptable quality in scaling speech. The algorithm appears to be very sensitive to accurate pitch extraction and requires the scaling factors to be recalculated as the pitch of the original signal changes.

No formal listening tests were reported in this article. The author claimed that with time scaling factors of up to 2.0 for different speakers and texts, the method was "informally judged to be very good". Other methods for assessing the quality were by examination of spectrograms, generally considered to be a poor indicator of the naturalness and quality of synthesised speech.

4.2.2. Frequency Domain Implementations

4.2.2.1. Seneff

The initial work on joint time-frequency representations of speech carried out by Portnoff [23, 24] included an algorithm for using the STFT to parameterise a model of speech production, similar to Malah's [2] in the use of band-pass filters equi-spaced in the STFT domain. Unlike Malah, the parameters were explicitly extracted in the STFT domain and altered before playback to achieve the desired modification. This technique was computationally costly and of questionable quality requiring a complex phase unwrapping procedure. Subsequent contributions built on this formulation of the specific properties of the STFT, but not on the parametric solution proposed. The first to add to Portnoff's work was Seneff [33] in 1982.

Seneff presented an approach which does not require that the individual pitch periods be extracted. In using a static analysis frame and performing all modifications in the STFT domain, she avoided having to determine the pitch explicitly (Portnoff required that instantaneous frequencies and phases for the fundamental and all its harmonics be calculated). This significantly reduced the complexity of the algorithm although the phase spectrum still needs to be unwrapped and modified.

The approach also deconvolves the excitation and source filter in the STFT domain, using a 17-point raised cosine filter to smooth the STFTM and obtain the spectral envelope. Pitch modifications are then performed on the excitation spectrum left by dividing out the spectral envelope. The modified excitation is finally recombined with the spectral envelope to form the Modified STFT (MSTFT). A block diagram of the entire process is shown in Figure 4-1.

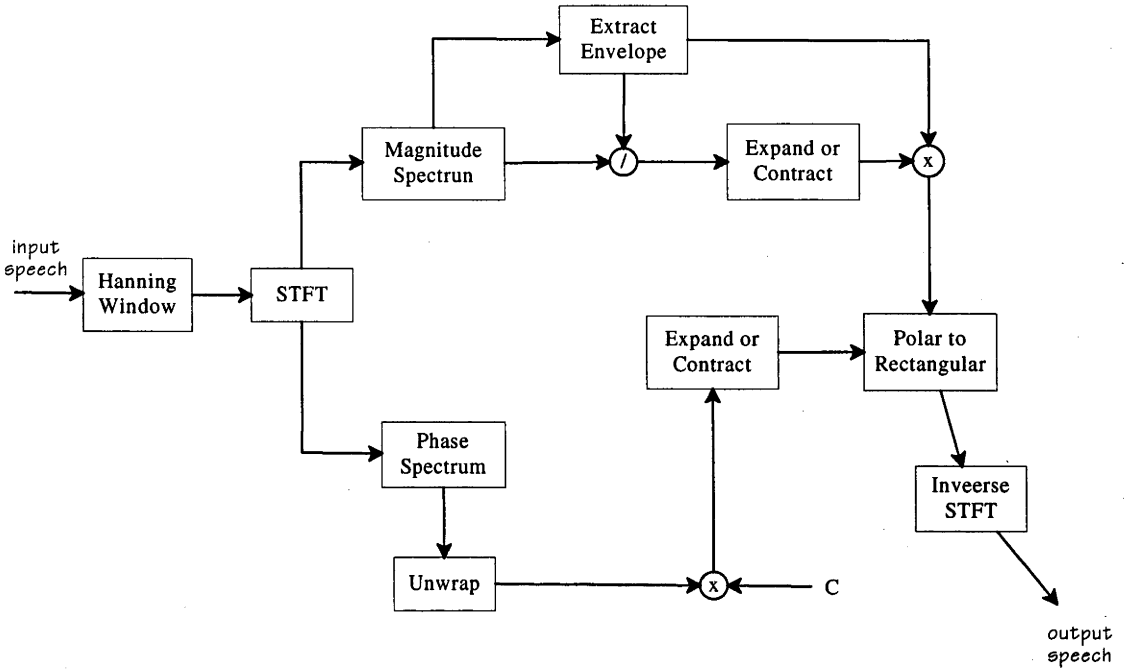

Figure 4-1: Block diagram of Seneff's algorithm. STFT frames are split into their phase and magnitude components with each modified separately. The phase component is unwrapped and pitch modified. The magnitude component has the "smoothed" envelope extracted before undergoing the pitch modification step. The modified phase and magnitude, and the unmodified vocal tract response, are combined and returned to the time domain.

A general technique for scaling signals in the Fourier domain is then explained. In order to increase the pitch of a signal, a resampling method is used to interpolate the spectrum magnitude of the excitation. A side-effect of this operation is that the width of the harmonic lobes of the excitation is increased. When the pitch is to be decreased, this method leads to another problem where the higher frequencies of the spectrum are left blank. The solution suggested by Seneff is to copy up the low frequency portions of the spectrum to complete the high frequency end as required. These problems and the proposed solution to the second problem are illustrated graphically in Figure 4-2.

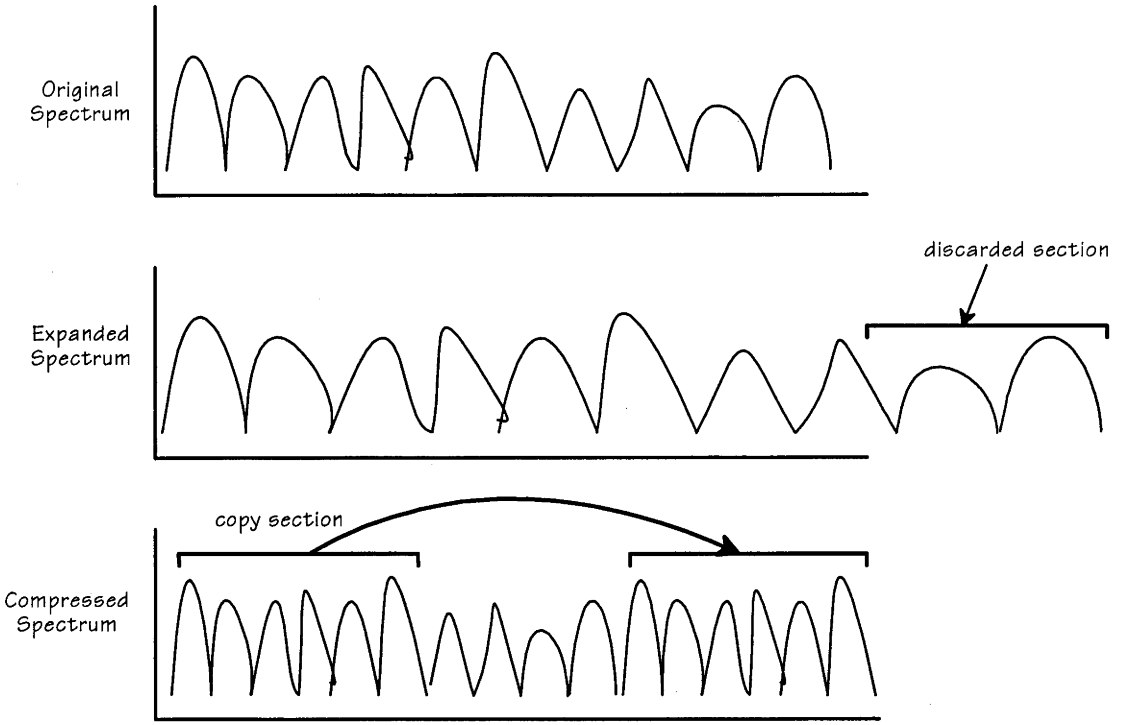

Figure 4-2: The problem of spectral remapping and Seneff's solution. When p>1, interpolation methods result in discarded information. When pitch lowering (p<1) is performed, then a section in the upper part of the spectrum is left blank. Seneff's proposal was to copy some of the baseband spectrum into the blank area.

No analytic justification is provided for this operation, and the approach is repeated in all of the derivative literature. It will be referred to as the *expansion-contraction* method. The only alternative approach to this is given by Moulines and Charpentier [5] and is known as the *elimination-repetition* method. A little more analytic effort is given to the problem there, although the primary citation is a doctoral thesis in French and therefore difficult to track down and comprehend. Moulines and Charpentier also outlined a variation on the expansion/contraction method where the high frequencies of the compressed spectrum are folded in the missing region. The paper indicates that both variants have equivalent subjective qualities.

Seneff, although the results of her testing were poor, has nevertheless introduced some key concepts. One is the deconvolution of excitation and vocal tract response (spectral envelope). Another is her method of scaling the pitch by linear interpolation in the frequency domain. An important idea introduced was the summation of the real parts of

the MSTFT to ensure the inverse transform provides a real signal in the time-domain. This has little foundation and is further examined in Chapter 5.

Twelve test sentences of 2 second duration were modified and informal listening tests used to judge the quality of the output speech. No intelligibility tests were performed, but spectrographic and spectral analyses were carried out.

4.2.2.2. Griffin and Lim

In 1984, Griffin and Lim [2] , proposed a solution to the problem of the invalidity of the MSTFT, and which gives a real-valued signal in the time domain from an arbitrary STFT magnitude. The proposal is included in a paper describing a system to time scale speech signals based on Portnoff's STFT work.

The method outlined for time scaling a signal involves using the STFT to determine the spectral characteristics at time instants on the original signal, and then recreating the spectral characteristics at different time instants in the output signal. It is, in fact, the FD-PSOLA technique of Moulines and Charpentier [5] but with a fixed analysis and synthesis frame and a pitch asynchronous placement of those frames with respect to the original signal. Figure 4-3 illustrates the system in block diagram format.

Of greater interest is the generalised technique used to generate the output time series from the modified STFTM. Note that the magnitude only is used, phase information is effectively discarded and reconstructed by the iterative synthesis step.

The paper describes a magnitude distance function to measure the difference between two STFTMs. An iterative approach is then used which guarantees that for a given modified STFTM and an arbitrary time series, the STFTM of the time series will converge to minimise the distance function. An initial estimate of the output signal is required to "seed" the iteration; the authors proposed a gaussian noise sequence.

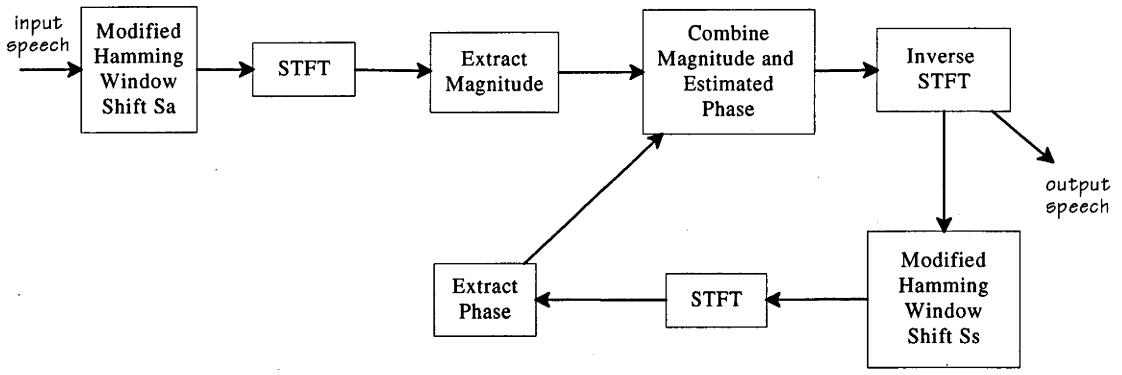

Figure 4-3: Block diagram of Griffin and Lim's modification algorithm. The overlapping STFT frames are analysed and synthesised at different shift lengths to realise the time expansion by a factor $t=S_s/S_a$. The iteration loop consists of the rewindowing the output speech and combining the STFT phase with the magnitude response of the original speech.

The synthesis frame shift must be at least 1/4 the size of each frame (Oppenheim and Schafer [29]), so in the case of the time scaling algorithm the iterative technique attempts to reduce the phase inconsistencies between successive frames as discussed above. The technique, however, has powerful uses for any modified STFTM and has gained a widespread use despite its complexity and slow convergence.

Figure 4-4 illustrates the method with reference to each frame for the case of time-expansion of a signal. Note that the shift between successive analysis frames, S_a, is scaled by the desired time-scaling factor to calculate the synthesis shift value, S_s. The iterative process for producing the output signal is then performed at the new synthesis instants.

The iterated step consists of taking the frame in the output signal which is centred on the current synthesis instant, and computing its STFT. The output signal is initially filled with Gaussian noise as a seed for the process, and each frame may contain some signal from overlapping frames computed immediately preceding the current one.

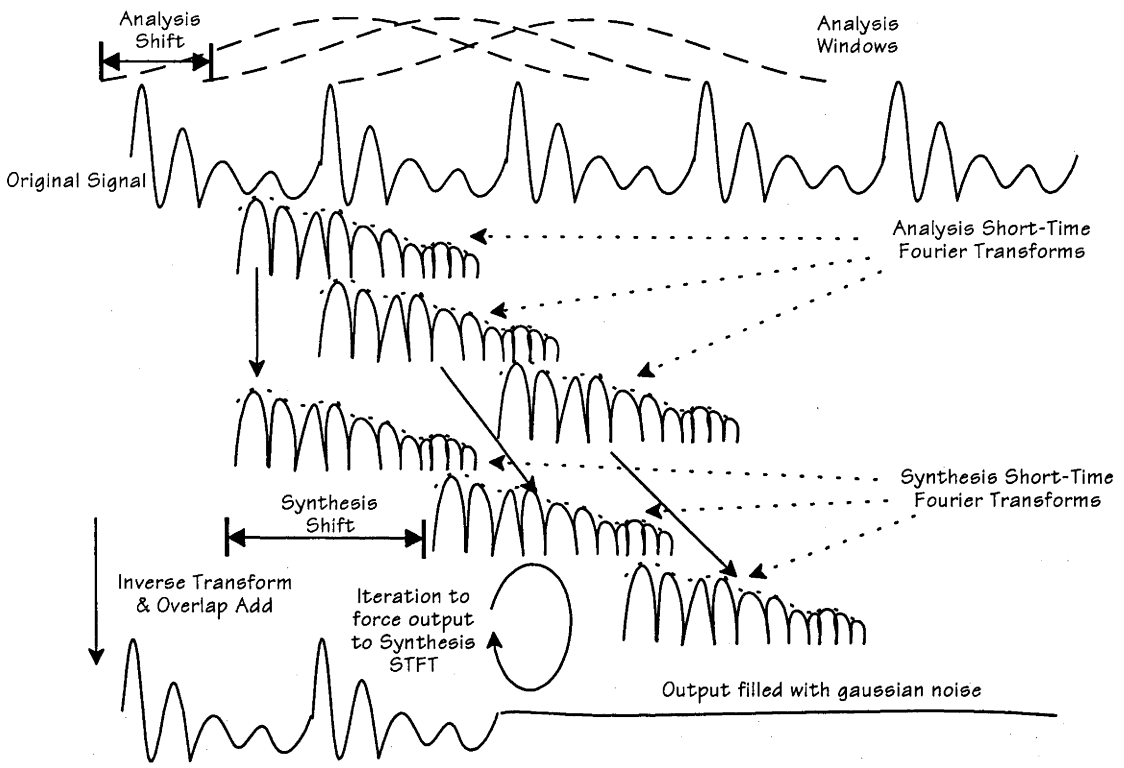

Figure 4-4: Graphical explanation of Griffin and Lim's algorithm for time-scaling. The synthesis step is performed in an iterative fashion with the STFT overlap add occurring at a different shift length than in the analysis step.

The STFT phase computed from the output array synthesis frame, is combined with the unmodified STFT magnitude of the original analysis instants to produce a modified STFT which is guaranteed to be real. Griffin and Lim showed that this iteration converges to minimise a distance measure in the STFT domain between the desired STFT response and the actual output STFT response.

This algorithm presented for time scaling speech has been implemented in a C++ program and using MATLAB M-files by the author. The source code, and the results of time-scale modifications, for both speech and singing, are included on the accompanying disk. Please refer to Appendix C for an index of files on the disk.

No formal listening tests were carried out by the authors. An informal assessment of the three files on the disk indicates that the process is quite high quality, certainly in comparison to Lent's algorithm. A "reverberation" artefact is noticeably present, even

on the unscaled file. The complexity of this algorithm is very high typically requiring 100 iterations to achieve acceptable quality.

4.2.2.3. Abe

Abe and his associates [20, 21] presented a refinement of the Seneff method for pitch scaling signals. Again, the interpolation method of the expansion/contraction method is used in the STFTM domain with spectral copying used to replace the high frequency components left blank when the pitch is compressed. In Seneff [33], however, the phase had to be explicitly determined and unwrapped with a scaling operation applied separately to that applied to the magnitude. The procedure is complex and critical to the success of the Seneff method; it is replaced here by a shift correction applied in the time domain prior to the overlap-add procedure. Interestingly, although they cite Roucos and Wilgus [32] (the SOLA method) there is no evidence of them using a cross-correlation to evaluate the required shift value. Instead, they imply that the shift is simply related to the pitch modification factor.

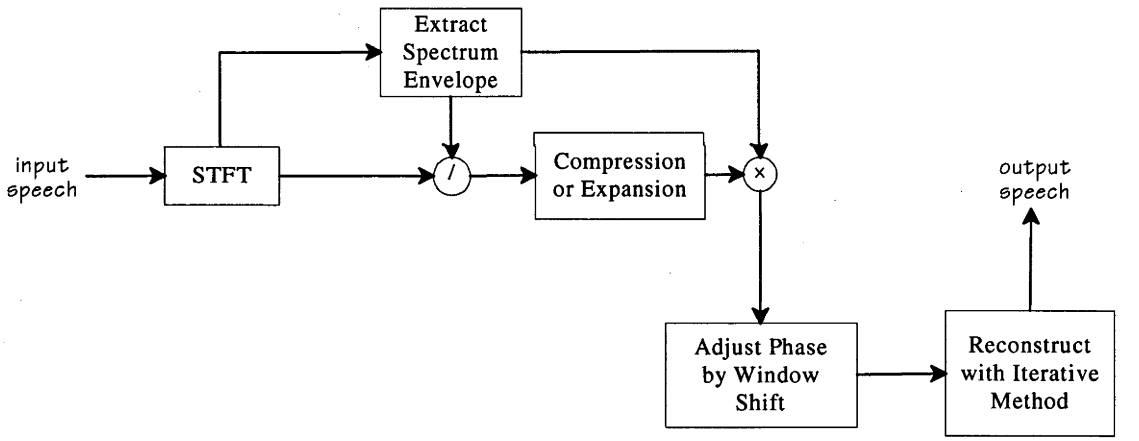

Figure 4-5: Block Diagram of Abe's algorithm. This method is a combination of Seneffs spectral expansion/compression and the Griffin and Lim iterative synthesis step. The separation of source and vocal tract is achieved by means of a new cepstrum technique. The STFT is divided by the extracted spectrum envelope to produce the excitation response for pitch modification.

The method outlined here makes use of the Griffin and Lim approach to synthesising the output signal from the modified STFTM. In addition, the deconvolution operation is refined in the STFT domain. Seneff used a *smoothing* filter to extract the spectrum

envelope. Abe and associates make use of *liftering* (see section 2.5 for more details) in the cepstral domain with a high order comb type lifter. This approach is designed to produce an excitation spectrum which is continuous, presumably resulting in higher quality modifications. The system is shown as a block diagram in Figure 4-5.

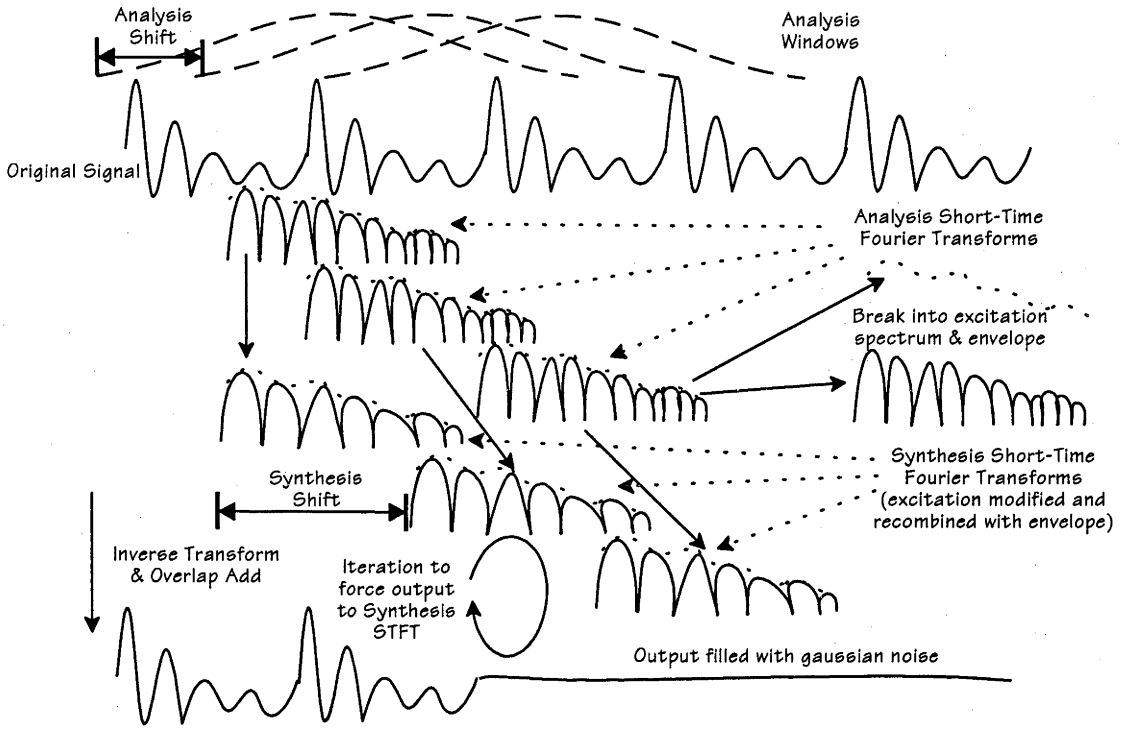

Figure 4-6: Graphical explanation of the algorithm of Abe et al. The analysis STFT frames have the excitation independently modified and then recombined at the synthesis rate. Time scaling can be performed simultaneously by making S_s vary from S_a.

As explained before, this continuous excitation spectrum is then expanded or contracted according to the pitch modification ratio. The modified excitation is then multiplied with the previously stored spectral envelope (this spectral envelope is discontinuous and does not exactly represent the vocal tract filter response - see Figure 4.3). Finally, the modified STFTM is used to estimate the output signal in the time domain using the Griffin and Lim algorithm with 20 iterations. No discussion is given of the initial estimate used, but a comparison of the convergence graphs with those from Griffin and Lim [2] suggests that the gaussian noise estimate is used in both.

Formal subjective tests were carried out with eight untrained subjects. The subjects rated the modified speech on a five point system compared to the original speech. A different method for synthesising speech was also compared to the original and the results showed a clear preference for the new method.

4.2.2.4. Other Parametric Approaches

McAulay and Quatieri [36] make use of time varying sinusoidal generators to model the speech production process of the excitation signal. Fixed length frames of 20ms length are used to segment input speech and the STFT is used to extract the parameters. Each sinusoidal frequency "track" contains components due to both the excitation and the vocal tract response. Homomorphic deconvolution was used to estimate the vocal tract contributions to the model parameters and a new method based on the Hilbert transform is used to estimate the phase contributions from the magnitude spectrum. By assuming that the vocal tract transfer function is minimum phase, the system magnitude and phase response form a Hilbert transform pair. If the phase estimate is derived from the logarithm of the magnitude estimate, the separation of vocal tract and excitation phase is greatly simplified. This minimum phase condition is an approximation because the vocal tract transfer function may contain zeros outside the unit circle in the z-plane [36].

Unlike Malah and Portnoff, the model parameters do not require that each sinusoid be equally spaced in the STFT domain. Instead, a peak picking algorithm is used and the between 20 and 40 frequency tracks assigned to the frequency peaks. From the STFT magnitudes and phases at each frequency peak, the excitation magnitudes and phases are found by removing the vocal tract response at these frequencies.

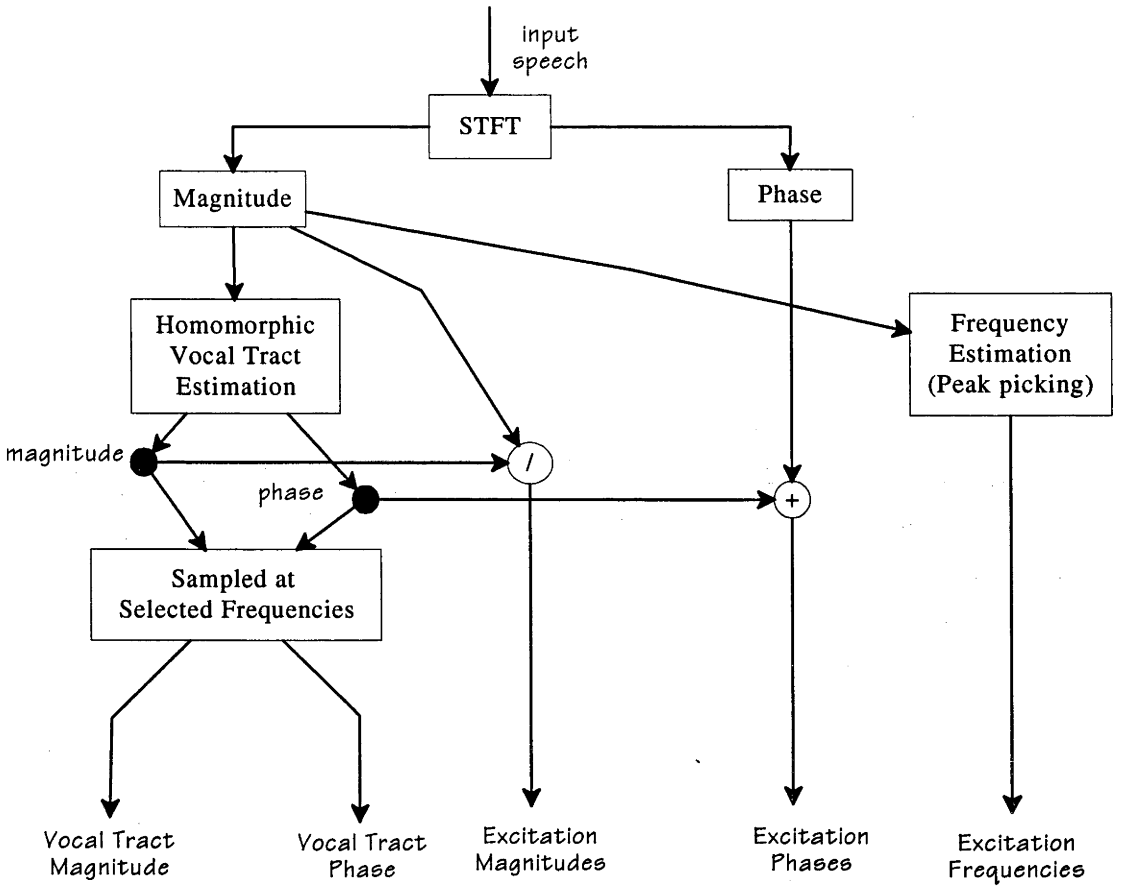

Figure 4-7: Block diagram of parametric analysis phase. Magnitude and phase are separated, and a vocal tract deconvolution is performed on the STFT magnitude. Specific parameters are then calculated for use in a parametric synthesis step, shown in Figure 4-8.

Synthesis is achieved by interpolating between the parameter values found for successive frames on a sample by sample basis, and filtering the instantaneous sinusoids through the vocal tract filter. Time expansion or compression is achieved by altering the number of points in the playback frame and interpolating the parameters across this new time scale, in order to preserve the temporal location of the frequency events. Pitch modification may be implemented by altering the instantaneous frequencies of the excitation sinusoids by the required pitch modification factor.

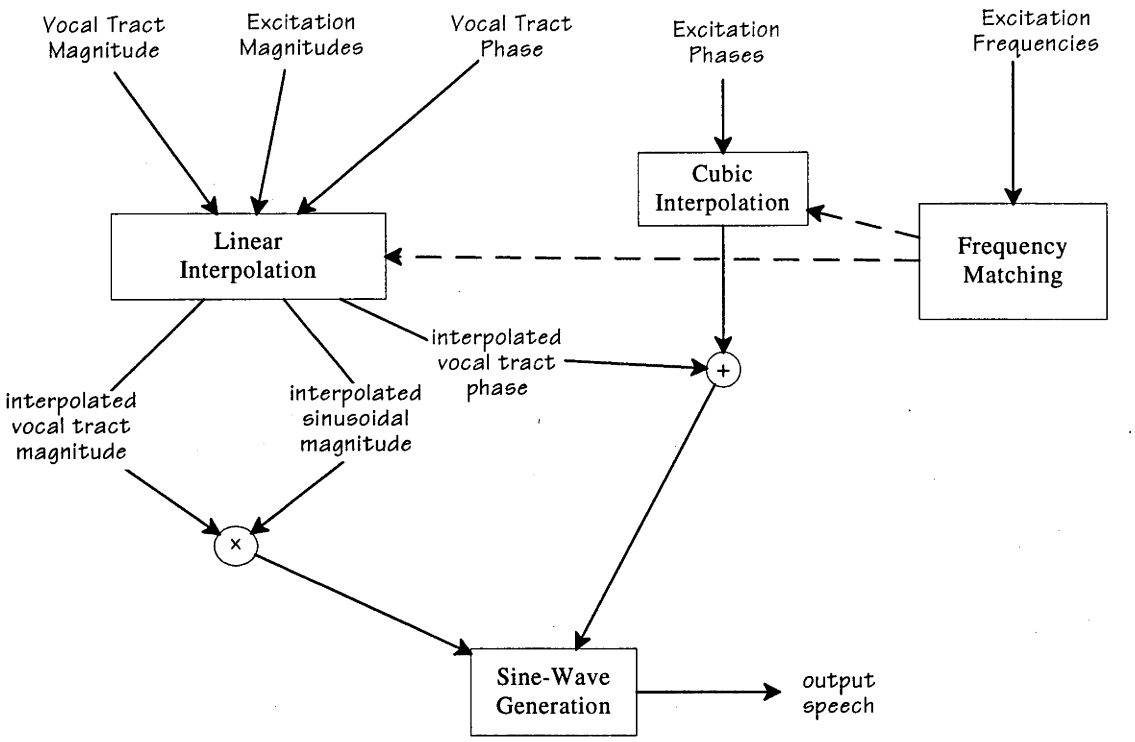

Figure 4-8: Block diagram of parametric synthesis step. A bank of sine wave generators, their fundamental frequencies, phases and amplitudes having been extracted from the STFT produce the output speech. Pitch scaling is achieved by changing the sine wave frequencies, and this necessitates an interpolation in order to find the new sine wave amplitude as dictated by the original vocal tract response.

One advantage of this approach is that non-uniform variations in time and pitch scale factors may be easily implemented. That is the modification factors may also change with time. This method is ideally placed to implement this.

Despite the apparent computational complexity of this method, the authors claimed in 1986 that a real-time implementation was possible using a DSP chip available at the time. Further investigation is warranted as to how this proceeded. Interestingly, despite the fact that the algorithm presented is intricately dependent on assumptions of the human speech production model, the authors claimed that the system performed successfully for non-speech sounds and speech with various types of interference. The list of signal types for which time-scaling was "smooth and without artefacts" includes music, multiple speakers, speech in noise, speech with musical background and even whale "speech"!

A related method is outlined by d'Alessandro [1]. This method is based on the Elementary Waveform Speech Model (EWSM) and requires knowledge of the spectral envelope to segment the STFT domain into different "harmonic regions" corresponding to the inter-formant intervals. Different waveforms are used to model the speech in each formant interval, for example, sinusoidal waveforms are used in the baseband (the lowest spectral region) and a technique utilising peak-picking gives the individual sinusoidal parameters as in McAulay and Quatieri [36]. A somewhat more complex method is used to model the higher spectral regions. The modifications proposed run into modifying locations and characteristics of individual formants, topics outside the scope of this survey.

In 1992, Quatieri and McAulay [35] suggested improvements to their original 1986 paper. They noted the objectionable "reverberant" quality introduced by frequency-domain-based transformation systems (also noted and demonstrated in the Griffin and Lim system above) and proposed methods to reduce this effect. The problem is blamed on the inability of frequency-domain systems to maintain the temporal structure of the original speech, in effect dispersing the modified waveform. The same sinusoidal model explained above is used, although the analysis frame is spaced at 10 ms and set to 2.5 times the average pitch period. This differs greatly from the original method as explicit pitch extraction is now required. In addition, the attempt to minimise the dispersion effect is to mark "pitch pulses" as those moments when the various sinusoids add in phase and to preserve the temporal location in the time-scaled version by ensuring that the sinusoids again add in phase at the time-scaled "pitch pulse" instant. This is remarkably similar to those methods employed by the PSOLA technique and by Abe et al [21] to preserve phase relationships.

No formal tests are referenced in either of the papers. One assumes that the authors opinions form the sole measure of the quality of the algorithms presented.

4.3. Non-Model Based Methods

4.3.1. Time Domain Implementations

4.3.1.1. Lent

In an article in the Computer Music Journal in 1989 Lent [18] proposed a computationally efficient technique which preserves formant characteristics. This heuristic algorithm included the ability to either time or pitch shift the sampled signal. The nature of the algorithm lends itself only to monophonic, or single source, pseudo-periodic sources.

The method proposed is similar to the later approach of Valbret, et al. [11] in attempting to synchronise the analysis frames with the periodic excitation source. In contrast to the approach of Valbret though, Lent provides no analysis of his method, the method of finding pitch markers is simplistic in the extreme and analysis frames are limited to a single pitch period. The claimed high quality of the modified waveforms is not evident from the investigations we have carried out. Pitch and time scaled samples are provided on the accompanying disk to illustrate this.

Analysis frame boundaries are found by low pass filtering the signal and marking zero crossings. The frames delineated by these zero crossings are windowed with a Hanning window of the appropriate length. Where time expansion or compression is to be performed, these windowed frames are repeated, in the case of expansion, or removed in the case of time compression. For pitch modification, the frames are shifted in time. That is, for pitch lowering, the frames are repeated at the new (or synthesis pitch) rate adding zero-valued samples in between the frames. Where the pitch is to be raised, each pitch period frame is played back overlapping (and added to) previous frames thus shortening the pitch period by the appropriate amount. The algorithm is presented as a block diagram in Figure 4-9.

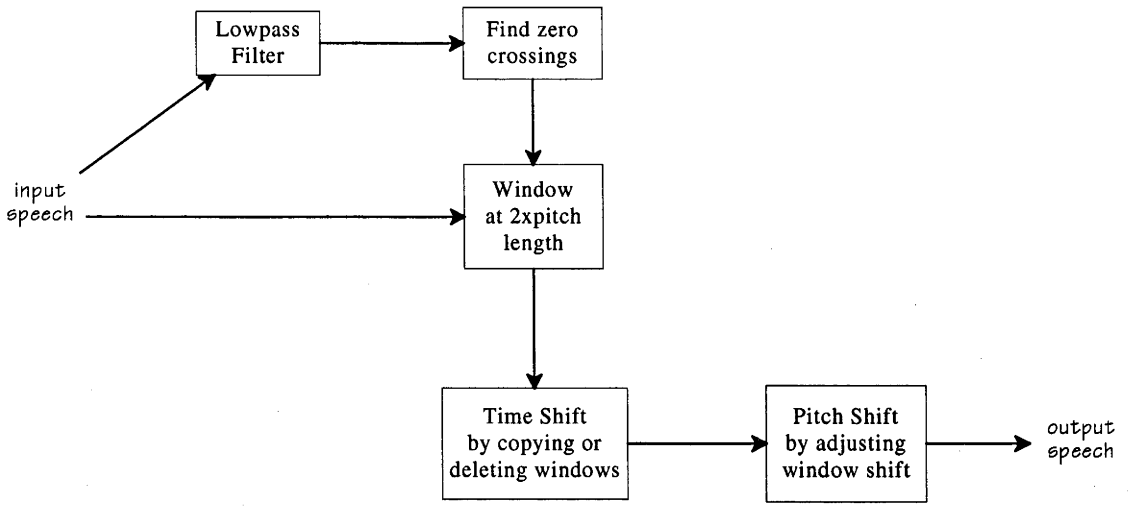

Figure 4-9: Block diagram of Lent's algorithm. Analysis frames are positioned according to a zero crossing technique for estimating pitch. Time and pitch scaling may be performed simultaneously; windows are repeated or deleted as necessary to achieve time scaling. The separation of the "pitch-synchronous" analysis frames is altered in the synthesis step by the pitch scaling rate when modifying the pitch.

The author has implemented this algorithm in both MATLAB scripts and a C++ program. The source code and WAV format files containing modified voice signals may be found on the accompanying disk. A detailed index of files on the disk is included in Appendix C.

The author reports that "the algorithm was found to work quite well on speech sounds". No formal comparative subjective tests were carried out. Further informal tests on piano tones were found to be acceptable with shifts of up to an octave (a pitch doubling). As can be heard from the test vectors, while the modified signals are intelligible, they can not be considered high quality

In a recent paper, Robert Bristow-Johnson [41] provided an interesting analysis of Lent's algorithm showing that it does, in theory, result in a shifting of the excitation harmonics. This is achieved whilst maintaining the formant locations from their original positions. This paper is interesting also in that the author has worked in the commercial music world for firms such as Fostex and Eventide (maker of commercially produced studio-quality pitch shifters). Lent's algorithm could well be improved by a better analysis frame positioning method. Bristow-Johnson also proposes better

windowing functions to improve the output quality. Given this, the method is equivalent to that of PSOLA, examined in a later section.

4.3.1.2. Roucos and Wilgus

In their 1985 paper [32], Roucos and Wilgus attempted to improve on the methods outlined by Griffin and Lim [2]. In doing so they discovered the basis for a new method of time-scale modification which is wholly time domain in its implementation. In both algorithms, the analysis frames are windowed and overlap the successive frame by 75%. Thus the analysis frame shift is 1/4 the frame length. At the synthesis stage, the frame shift rate is altered by the time scaling factor. In the Griffin and Lim approach, the STFT of the output signal is made to converge to the STFT of the current analysis frame by using an iterative technique requiring one FFT and one inverse FFT per iteration.

The justification for Roucos' and Wilgus' research was to develop techniques to reduce the computational overhead of the Griffin and Lim technique. They initially provided the details of a study into the convergence of the Griffin and Lim iteration for various initial output signal estimates. Noting that the criteria Griffin and Lim use to measure the convergence of the iteration accounts only for the magnitudes of the spectra of successive frames, and does not concern their phases, they try several approaches. A comparison is presented between the convergence curves of signals using initial output estimates formed from a gaussian noise sequence as presented by Griffin and Lim, a linear prediction residual and a frequency domain estimate obtained by zeroing the phase of the modified STFTM.

The results of this investigation showed that the fastest converging method used the LP residual as its initial estimate. However, the Griffin and Lim method still required 50 iterations before the synthesised speech was of an acceptable quality. In an attempt to reduce the number of iterations necessary for high-quality rate modification, they identify the primary problem of the Griffin and Lim approach to be the discarding of phase information in the modified STFTM. This causes the pitch pulses in successive overlapping frames to be misaligned in time; making for a large initial error in even the LP estimate of the speech and also resulting in the signature "reverberation" still present after many iterations. The zero-phase estimate was a first attempt at aligning the pulses of successive frames.

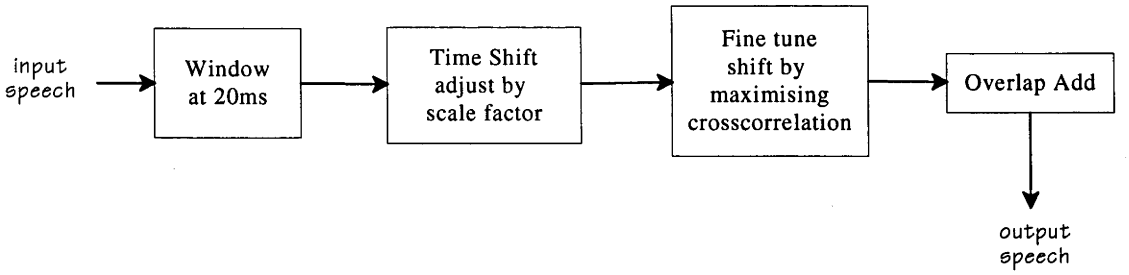

Figure 4-10: Block diagram of SOLA method. The overlap-add step is performed at a scaled synthesis instant, but the synthesis shift is "fine-tuned" by maximising a local cross-correlation sum. Note that the analysis frames are not pitch-synchronous, but both analysis and synthesis shift lengths must be at least one-quarter that of the frame length used.

The innovation introduced in this paper is to maximise the time-domain cross-correlation between successive windows before the overlap and add iteration step. This provides a small correction to the synthesis shift length, so that the time scaling is not achieved exactly. This Synchronised Overlap and Add algorithm (SOLA) provides such a dramatic improvement in the Euclidean distance between the target STFTM and the output signal from the Griffin and Lim technique that no iterations are necessary. The initial SOLA estimate is at least as high in quality as the LP excitation initial estimate after 100 iterations and the whole algorithm may be carried out in the time domain.

Roucos and Wilgus claimed the technique to be effective on speech in noise and speech passages including more than one speaker. In addition, the method requires only a fraction of the computations required by the Griffin and Lim approach for modifications of similar type and quality. No formal subjective tests were carried out, however the technique seems to be widely used as it is not speech specific. The authors also rely on the dubious STFT domain distance used to measure the convergence of the Griffin and Lim algorithm as an objective measure of quality.

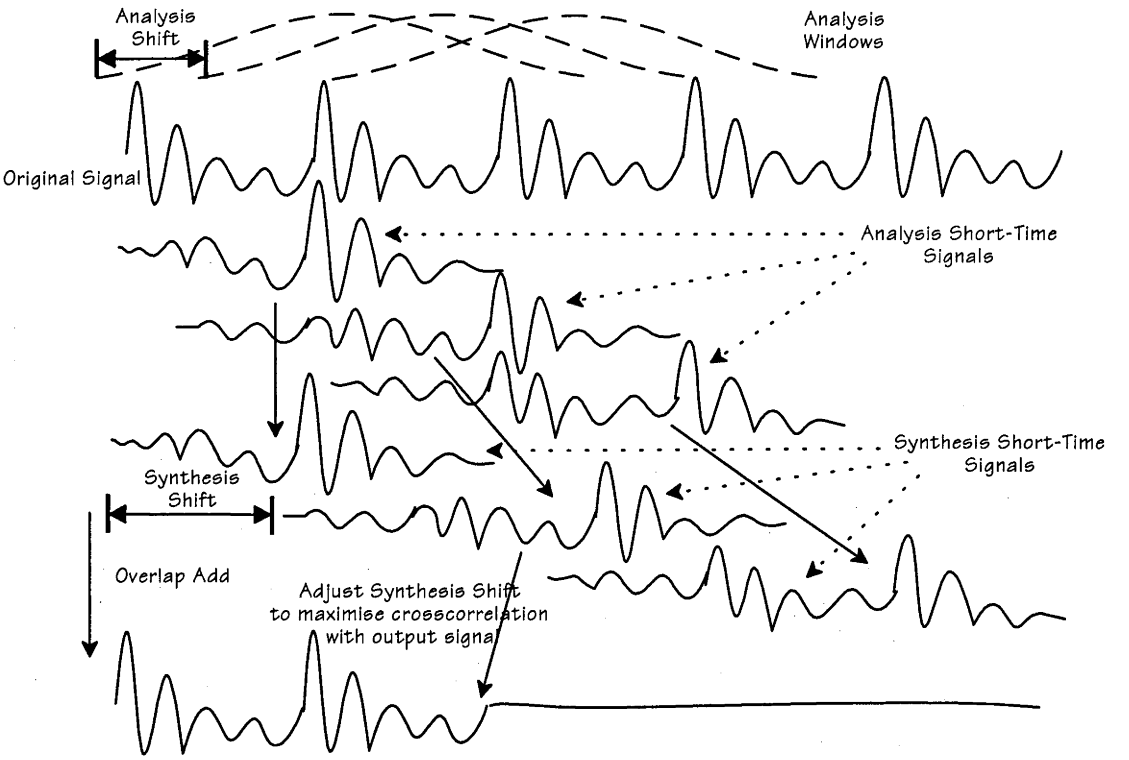

Figure 4-11: Graphical explanation of SOLA algorithm. In a single step, the crosscorrelation adjustment of the synthesis shift out-performs the 100 iteration, transform-domain approach of Griffin and Lim. The correlation ensures that waveform peaks in overlapping frames are correctly aligned and are not "blurred" in the time domain.

4.3.1.3. PSOLA

Moulines and Charpentier [5] proposed the Pitch Synchronous Overlap and Add (PSOLA) algorithm specifically for the improvement of text-to-speech synthesis systems but with emphasis on time and pitch scaling the signals to improve the subjective quality. Valbret, et al. [11] expanded on the subject, with more emphasis on speaker transformation where a speech utterance is modified to make it appear that a different speaker (the target speaker) is the source.

These papers introduced a similar approach to the other time domain approaches with a cut-and-paste method predominant. The important difference is that the portions to be cut in the analysis are the centred on successive instants, called pitch marks, which are set at a pitch-synchronous rate corresponding to glottal impulses. The window used is

the Hanning window of a length proportional to the local pitch period. The method is known as Time Domain PSOLA (TD-PSOLA) in order to distinguish it from the frequency-domain pitch shifting variant described below.

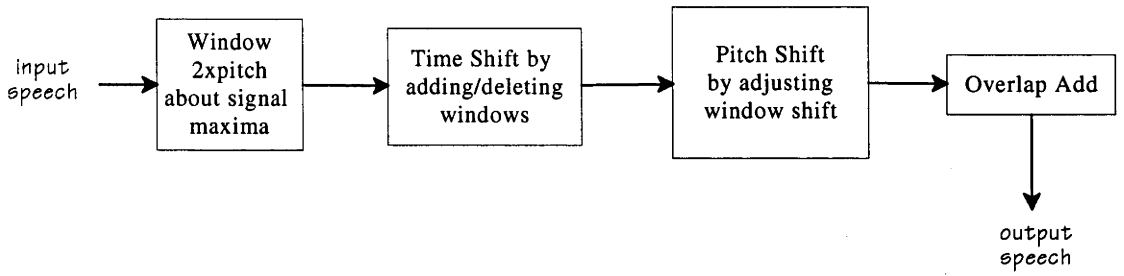

Figure 4-12: Block diagram of TD-PSOLA. An initial peak-picker positions the analysis frames about pitch-synchronous signal maxima. The modification of time scales by repeating or deleting frames is the same as in the Lent algorithm. Similarly, the pitch is scaled by altering the synthesis shift length. The pitch-synchronous nature of each frame precludes the need to fine-tune the new shift length as overlapping frames are centred on signal maxima.

The initial step, that of placing the pitch-marks, is acknowledged to be the most important in maintaining the high quality of the synthesis step. However, nowhere is the method of obtaining these pitch-marks explained. Indeed, it appears that in some cases, the pitch marks are positioned by hand! Once the analysis pitch marks have been obtained, a mapping to a set of synthesis pitch marks is determined. This mapping is similar to that proposed in Lent, although here the pitch synchronicity is far better realised.

Time scale modifications are achieved by copying or removing analysis frames as in the Lent case. They noted an acoustical artefact caused on unvoiced sections of the signal in cases where the speech is to be slowed down by a factor above 2, where a short term correlation is introduced in to the synthesised signal being perceived as a tonal noise.

The time compression case is illustrated in Figure 4-13. Note that the most central of the three analysis windows is discarded in order to maintain the pitch spacings of the original.

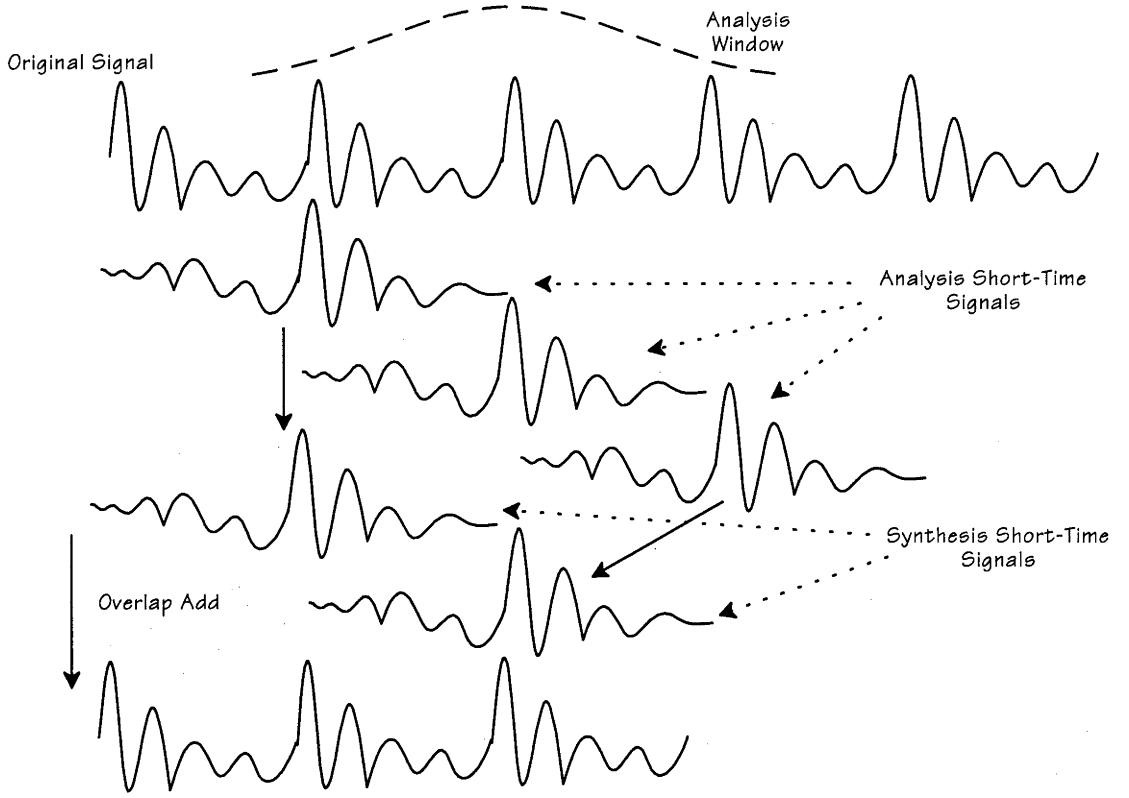

Figure 4-13: Graphical explanation of TD-PSOLA time compression. The length of the signal is to be shortened, and pitch-synchronous analysis frames are discarded before the overlap-add step to achieve this.

Pitch scaling is achieved in a similar fashion to that of Lent, with the synthesis pitch-marks being placed closer together for increases of the pitch and further apart for pitch lowering. An alternative to the time-domain method, called Frequency Domain PSOLA (or FD-PSOLA), is presented with each of the analysis frames being transformed to the STFT domain and modified before synthesis. This involves several of the problems discussed in the section on frequency domain methods in general.

This is further illustrated in Figure 4-14 where it can be seen that a discontinuity is introduced into the signal in the overlap add stage at the new pitch. It is apparent that with this method, as in Lent's, the simultaneous modification of both time and pitch is possible.

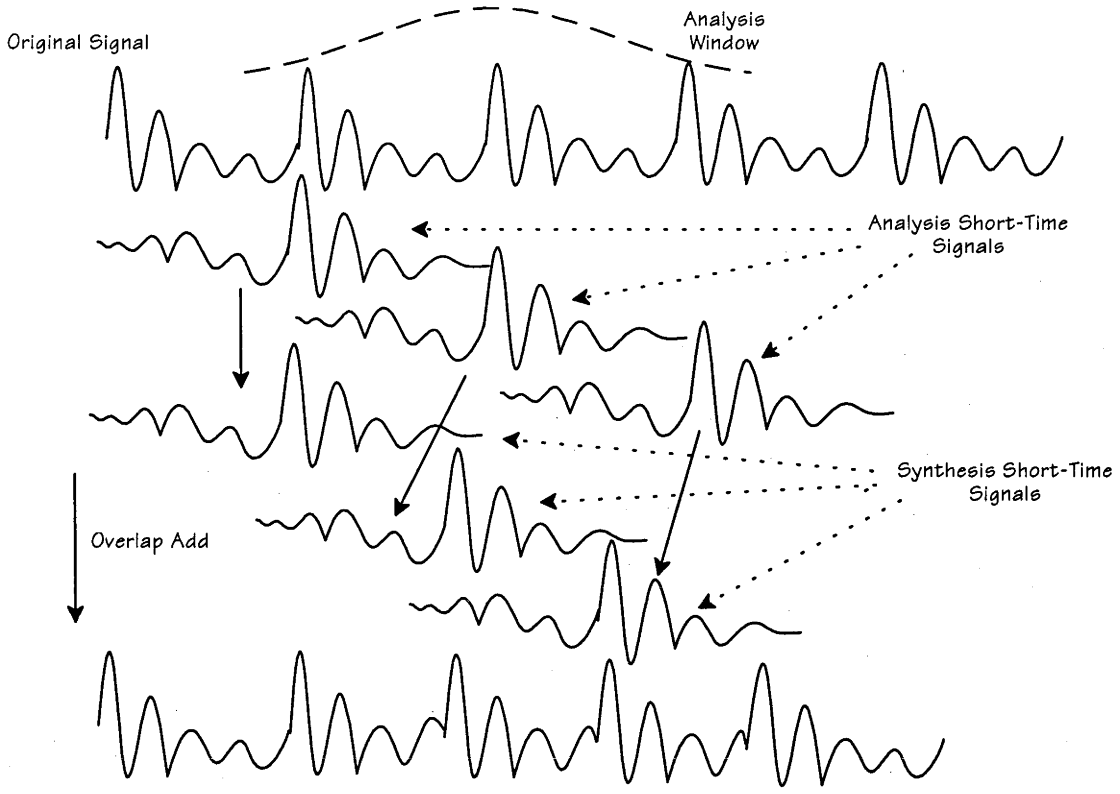

Figure 4-14: Graphical explanation of TD-PSOLA increasing pitch. The pitch-synchronous frames are overlap added at shifted synthesis instants. A pitch estimation method is inherent in the selection of analysis instants about the signal maxima.

The time-domain method may be concatenated with a linear prediction analysis stage and the TD-PSOLA technique applied to the residual excitation signal. This leads to very high quality modifications at a cost of further complexity. Because the synthesis pitch marks are not aligned with the frames used in calculating the LP filter coefficients, these must be recalculated for the synthesis instants before the refiltering process can occur. This is generally done by interpolating successive sets of LP coefficients at the synthesis instant.

These papers, and an associated survey paper (Moulines and Laroche [6]) succeed in presenting a unified framework for analysing all of the methods which involve segmenting the input speech signal in the time-domain. They also provide an of the TD-PSOLA method in terms of a parametric speech model. This model consists of a superposition of a deterministic periodic signal (for voiced segments) and a zero-mean

wide-sense stationary process (for unvoiced segments). They examine in detail the effect on formant bandwidths and harmonic magnitudes of the TD-PSOLA pitch modification algorithm.

The speech quality gain was evaluated using a formal test comparing speech synthesised using a conventional LPC synthesiser and speech synthesised by the FD-PSOLA, TD-PSOLA and LP-PSOLA methods. The tests used 16 subjects and 10 different sentences, and the systems were compared in pairs alternating A-B and B-A for preference. The results showed that the three algorithms performed much better than the LPC synthesis and were roughly equivalent to each other. No tests were performed on the modification of natural speech.

4.3.1.4. GLS-TSM

This method, proposed in 1996 by Yim and Pawate [34], is a computational improvement on the SOLA method described above. This method used the overlap and add principle to copy input frames in which the exact positioning of the new frame into the output signal is found by maximising a cross-correlation function. The analysis frames are static in size, unlike PSOLA, and bear no relation to the pitch pulses. However, the cross-correlation in SOLA substitutes for the difficulty of accurate placement of pitch-marks in PSOLA.

Yim and Pawate proposed an improvement on the cross-correlation step. Instead of performing the full cross-correlation across all possible pitch periods, a preliminary global similarity search is performed which narrows down the possible shift values. A search for local similarity is then performed in the area indicated by the preliminary search. This two-step approach produces high quality results and achieves an increase in computational efficiency of 40 times that of the SOLA rate according to the authors.

No formal comparison tests are cited.

Chapter 5 - An Investigation into Alternative Scaling Methods

5.1. Introduction

We have seen in Chapter 3 the analytic representation of what we mean to achieve when we talk of scaling a speech signal in the time or frequency domains. In Chapter 4 we saw how various techniques have attempted to address the problems in devising algorithms to satisfy these objectives. We now turn to alternative techniques which, we hope, can be implemented in order to avoid some of the complexity problems from which frequency domain methods suffer. The work in this chapter is entirely that of the author, with the exception of Section 5.2.3 which contains a digression to discuss a related, well understood topic.

Specifically, we seek a time-domain methodology which can be implemented in real-time to alter the pitch of an arbitrary signal. The possible uses of this method will then placed within the context of voiced speech, with particular reference to sung vowels.

We start by examining the Fourier domain scaling operation in terms of a bank of filters, a view which leads to a method of altering the pitch of a sung vowel with acceptable quality results. The method is refined and real-time implementation issues are addressed.

Another possible scaling method suggested by the duality of speech modification and coding techniques is also investigated, with somewhat less success.

The pitch of human speech, and indeed singing, varies rather slowly with time. A 5 ms pitch extraction window is commonly used in toll quality CELP (Code Excited Linear Prediction) implementations. The variation of the pitch fundamental with time is further explored in Section 5.2.5. For the purposes of this investigation we presuppose

that we are operating on a time-limited segment of speech where the pitch is constant. Indeed, we are assuming that the operation is being performed on an excitation signal alone; that is, the vocal tract effects have been deconvolved. Samples of a male singing vowels at constant pitch are used to examine the effects of the proposed schemes with reference to subjective quality and pitch perception and scrutiny of spectrograms.

5.2. Filter Banks

5.2.1. Arbitrarily Located Bandpass Filters

The interpretation of the Short Time Fourier transform as the output of a set of bandpass filters is well known (Oppenheim and Schafer [29]). This suggests a time-domain method of isolating and modifying segments of the signal's spectrum. If we modulate the outputs of each filter by an appropriate signal in order to shift these newly isolated frequency patches, we can add the shifted signals to synthesise a signal with an altered spectrum.

One advantage of implementing the analysis stage as a bank of filters is that we have greater flexibility in the choice of frequency responses. Where a system based on the STFT in essence provides filters with overlapping responses, the generalised filter bank system can be designed from filters with arbitrarily sharp cutoff, better interband isolation and unequal bandwidths.

Consider a filter bank system designed to modify pitch at a time-invariant rate p, shown in Figure 5-1. Here we are using with a cosine signal used as the modulation carrier. This is analogous to a single sideband modulation in the continuous domain. The scheme uses N filters of constant bandwidth to segment the STFT domain .

This scheme has been implemented by the author as a pair of MATLAB scripts which are included on the accompanying disk. Twenty filters are used in order to get 200Hz resolution at baseband. Smaller bandwidths result in too many filters and very high order. Each filter was designed using the Remez algorithm with passband ripple of 3dB and stopband rejection of 40 dB. The results of this scheme in pitch modifying a range

of signals are also included in WAV-file format on the disk. The naming scheme for these files and a full index and description of each file is included in Appendix C.

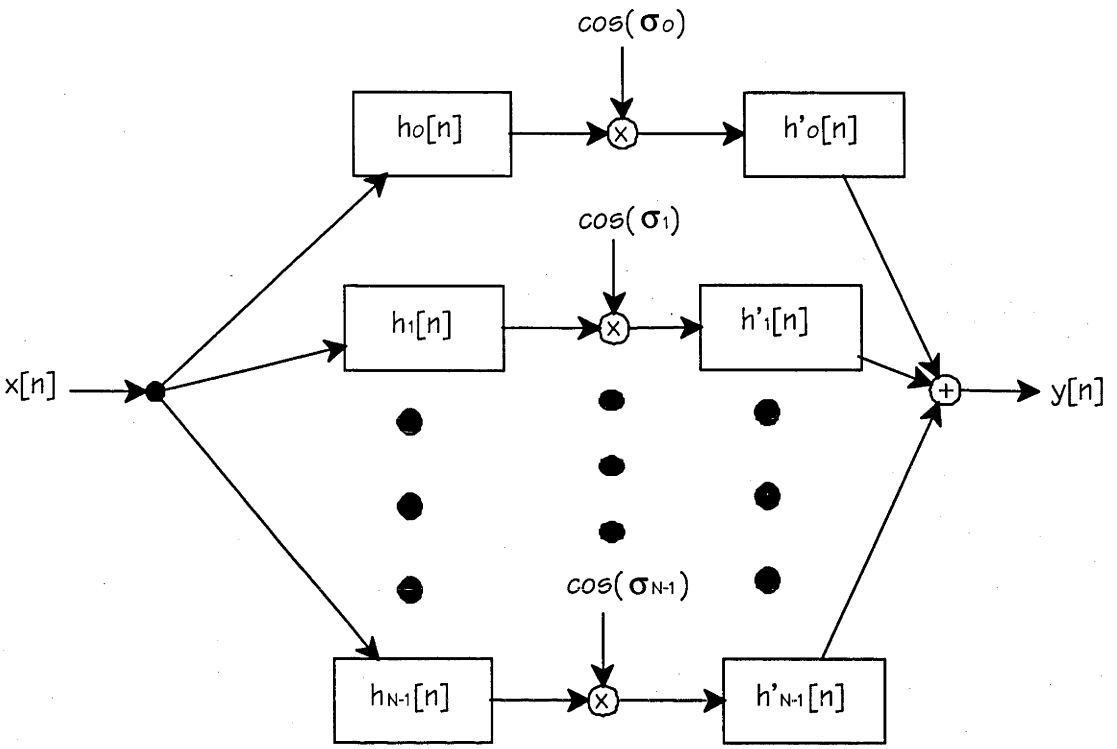

Figure 5-1: Bandpass filter bank with SSB modulation. N equi-spaced bandpass filters are used to isolate each frequency segment. Each segment is modulated to a target frequency with a cosine signal and the segments are resummed.

For the outputs of the N analysis filters, we have

$$x_k[n] = \sum_{m=0}^{L} h_k[n] * x[n-m] \qquad k = 0, 1, 2, \ldots, \text{N-1} \tag{5.1}$$

where the h_k represent the N bandpass filters, each of FIR length L, and equally spaced in the frequency domain with centre frequencies given by

$$f_k = \frac{2\pi k}{N} \qquad k = 0, 1, 2, \ldots, \text{N-1} \tag{5.2}$$

In order to perform the modulation step, we must define the modulation (or shift) factor. This amount is different for each frequency segment, since we are linearly scaling the frequency response, not shifting it by uniform amounts. Hence, we seek to translate each segment to a new centre frequency, f'_k. Before summing, we also need to refilter the individual "channels" at the new centre frequencies to avoid introducing aliasing effects from shifted mirror signals (negative frequency "mirror" components may have been accidentally moved into the bandwidth of another frequency segment). The calculated synthesis filters may be considered shifted versions of the analysis filters. The target centre frequencies are given by

$$f'_k = \frac{2\pi kp}{N} \tag{5.3}$$

This implies that the shift amounts are given by

$$\begin{aligned} \sigma_k &= f'_k - f_k \\ &= (p-1)\frac{2\pi k}{N} \end{aligned} \tag{5.4}$$

We see that for $p>1$, or pitch raising, the shift amount is always positive and dependent on k, the "band number". Conversely, if $p<1$, the shift amount is always negative for lowering of pitch.

Since the analysis filters are equi-spaced in the frequency domain, their bandwidths are constant and given by

$$BW_k = \frac{2\pi}{N} \tag{5.5}$$

When a frequency segment, k, is modulated, so is its negative frequency mirrored segment. In order to avoid an intersection between the sum and difference components

in the modulation, we must constrain the individual frequency shifts to be greater than a half filter bandwidth.

Thus we have

$$\sigma_k \geq \frac{BW_k}{2} \qquad (5.6)$$
$$p \geq \frac{1+2k}{2k}$$

The right hand side of the inequality is maximum at $k=1$, k being integral. We ignore the $k=0$ case as we will not frequency shift the "DC" signal. Hence for the constant bandwidth case, p must be greater than 3/2 or else unaliased reconstruction is not possible. This means that we cannot have a frequency scaling factor less than 1. That is, this method cannot reduce pitch of a signal without introducing overlapping between adjacent spectral chunks. In Figure 5-2, this "squishing" of the synthesis filters in the case of spectral compression is shown at the bottom, in comparison to the analysis filters, at top, and the spectral expansion case, depicted in the middle.

This type of problem was addressed by Moulines and Charpentier [5] who tried two methods of avoiding the overlap problem. The first is to reduce the width of each chunk when producing the synthesis signal. The second is to discard alternate chunks. In the case of Moulines and Charpentier, the modification was occurring in the Fourier domain, but we can achieve the same results in the time domain with our frequency-segmented signals.

The two methods of addressing the problem were implemented and compared. In the first, the down-shifted spectral bands are refiltered with a set of synthesis filters with reduced bandwidths. In the other method, for $p<1$, alternate spectral bands are not added into the synthesis signal. The reduction factor used is p. Subjectively, the effect of eliminating adjacent bands is better than the second method when tried on singing. For speech, the difference is harder to determine. These WAV-format files can be found on the accompanying disk, in the alfb subdirectory. Consult Appendix C for a full index and description of each file.

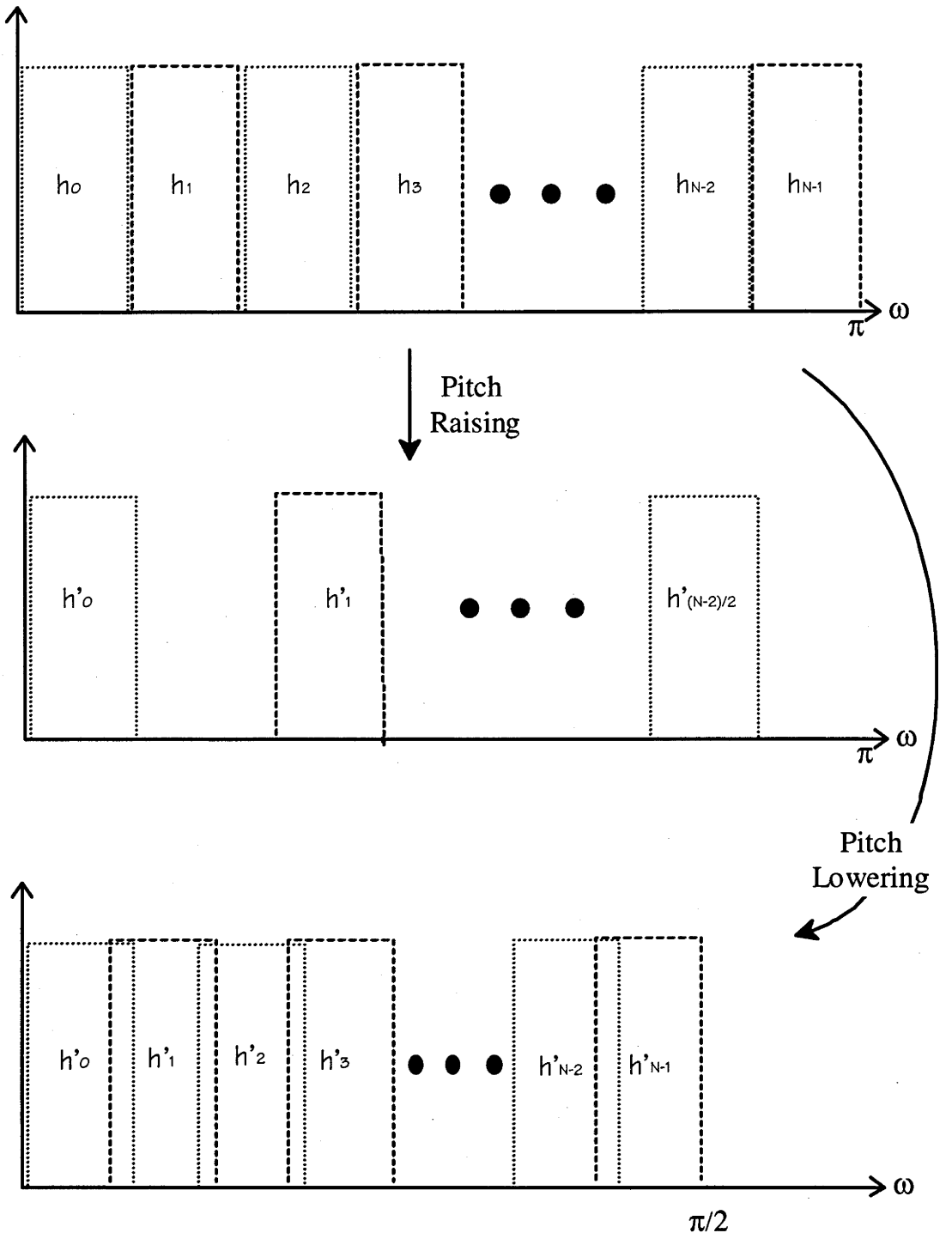

Figure 5-2: Filter placement when using filterbanks for pitch scaling. Analysis filters are shown at the top. The middle set of filters are synthesis filters for the $p>1$ case and the bottom set for the $p<1$ case. Gaps are left when the pitch is raised and undesirable overlap occurs when lowering the pitch.

Another constraint may be discovered with reference to spectrograms of the modification process. Namely, the analysis filter bandwidths must be smaller than one inter-harmonic spacing of the actual source harmonic frequencies. To see this, consider the spectrogram of a male singing a long "A" vowel at a fundamental pitch of approximately 200 Hz, shown in Figure 5-3. The signal was sampled at 8kHz.

The structure of the harmonic frequencies is clearly visible, and the inter-harmonic spacing is seen to be constant at 200 Hz. This inter-harmonic spacing (or interval) is crucial to the human ear when perceiving the pitch of a sound (Meddis and Hewitt [39]).

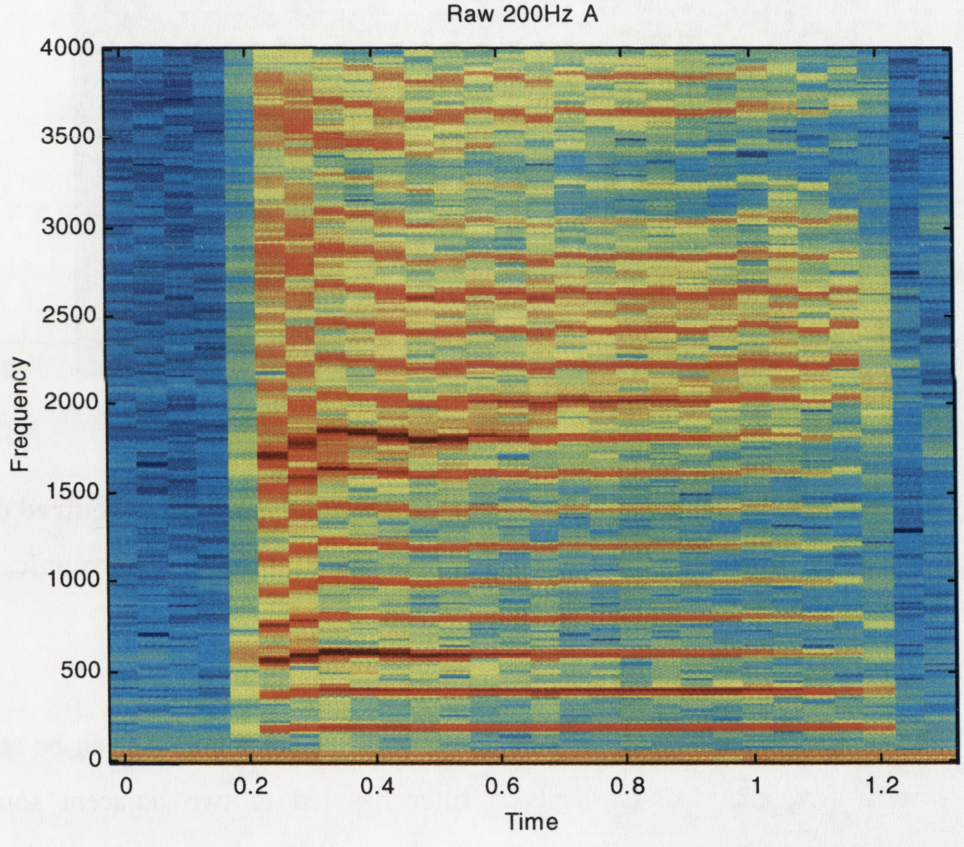

Figure 5-3: Spectrogram of sung A. The fundamental pitch is approximately 200 Hz. Pitch perception is primarily based on the spacings between the clearly visible harmonic frequencies.

The signal used for these spectrograms, and in the example WAV files was the original signal with no source/filter decomposition performed. Thus, in Figure 5-3, the first

couple of formants are clearly visible at the harmonic frequencies near 600 Hz and 1600 Hz. These formants will be shifted around and will decrease the subjective performance of the technique, but the implementation of an adequate source/filter decomposition is outside the scope of this thesis. The implications and requirements of an acceptable deconvolution scheme compatible with other aspects is further discussed in 5.2.5.

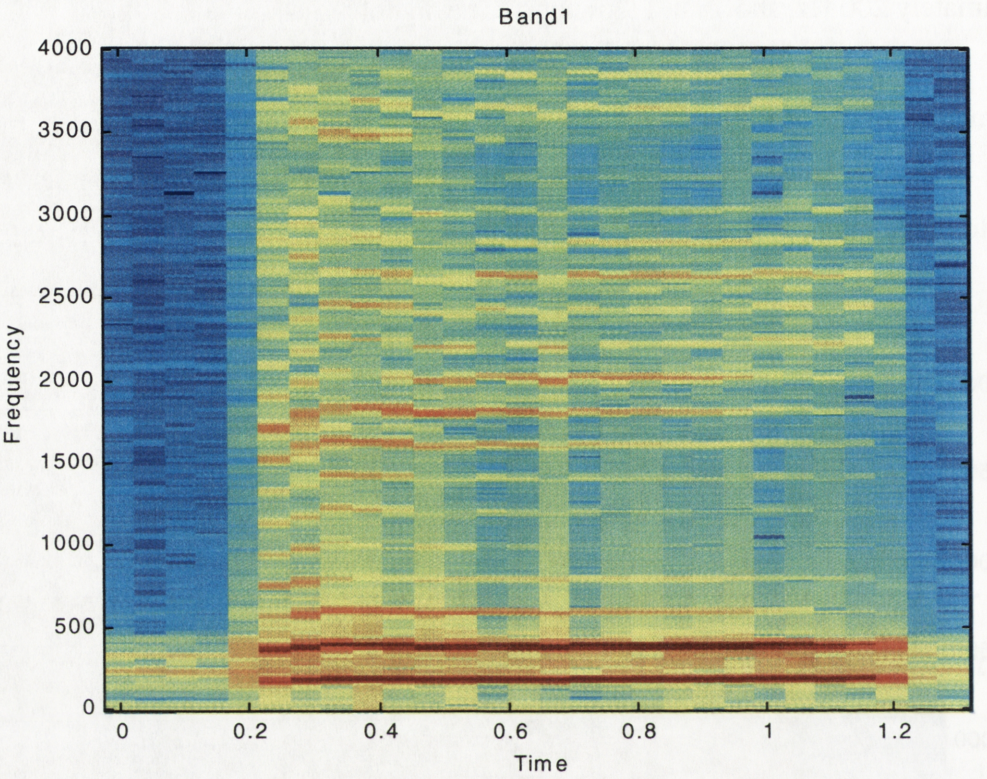

Figure 5-4: Analysis signal, band 1, 200-400 Hz. The analysis filter is centred on 300 Hz, and energy from both the 200 Hz fundamental and the first harmonic, at 400 Hz, are captured.

In Figure 5-4, the isolated spectral chunk is shown as a spectrogram. As can be seen, the placement in frequency of the analysis filter has led to two adjacent source harmonics being "captured". These harmonics are also well into the transition region of the filter, and hence are also attenuated somewhat.

The output signal for a pitch increase by a factor of two is shown in Figure 5-5. Compare the output signal with the same singer singing the same vowel at the target pitch, 400 Hz, displayed in Figure 5-6. The inter-harmonic spacing has been drastically altered, with new and attenuated harmonics appearing at new spacings.

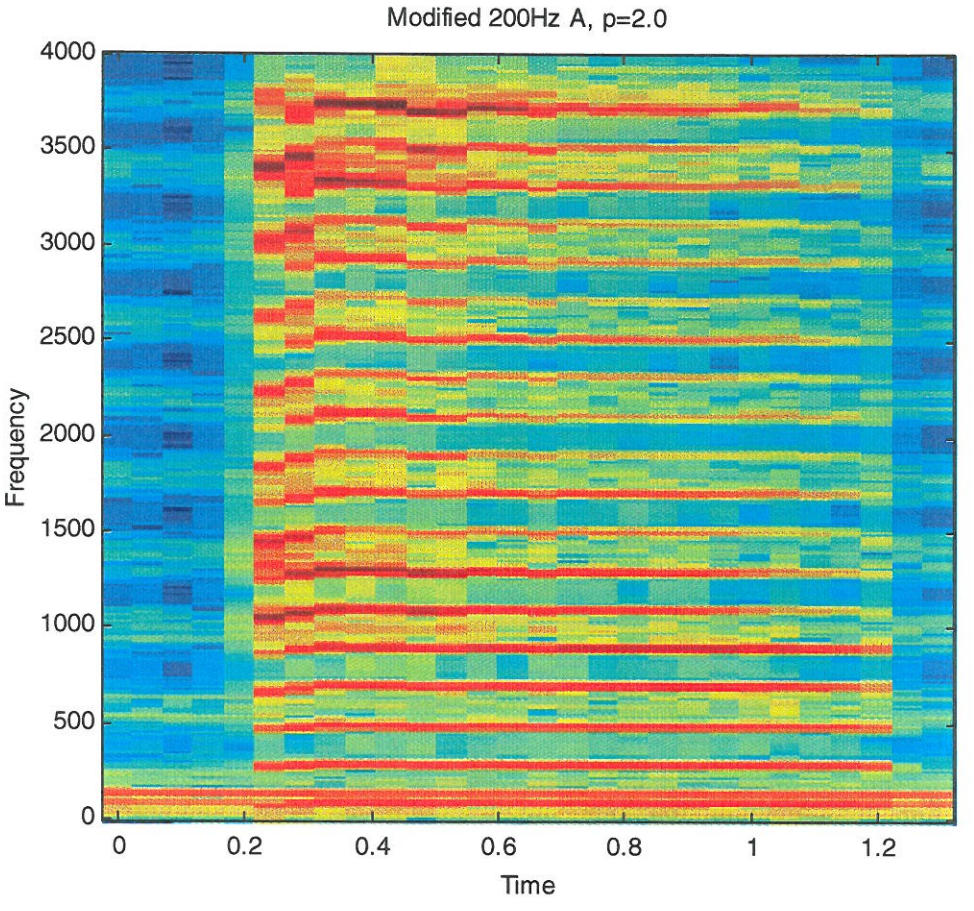

Figure 5-5: Output signal, pitch raised by a factor of two. Each frequency segment has been shifted up; the DC levels have been moved to a quite audible frequency range at the bottom of the spectrogram. Because the analysis filters were not centred on individual harmonics, the interharmonic spacing has not been correctly scaled.

This illustrates that the differences between an analysis filter centre frequency and actual harmonic frequencies within the pass-band of the filter, will cause the harmonics to be shifted instead of scaled properly by the modulation process.

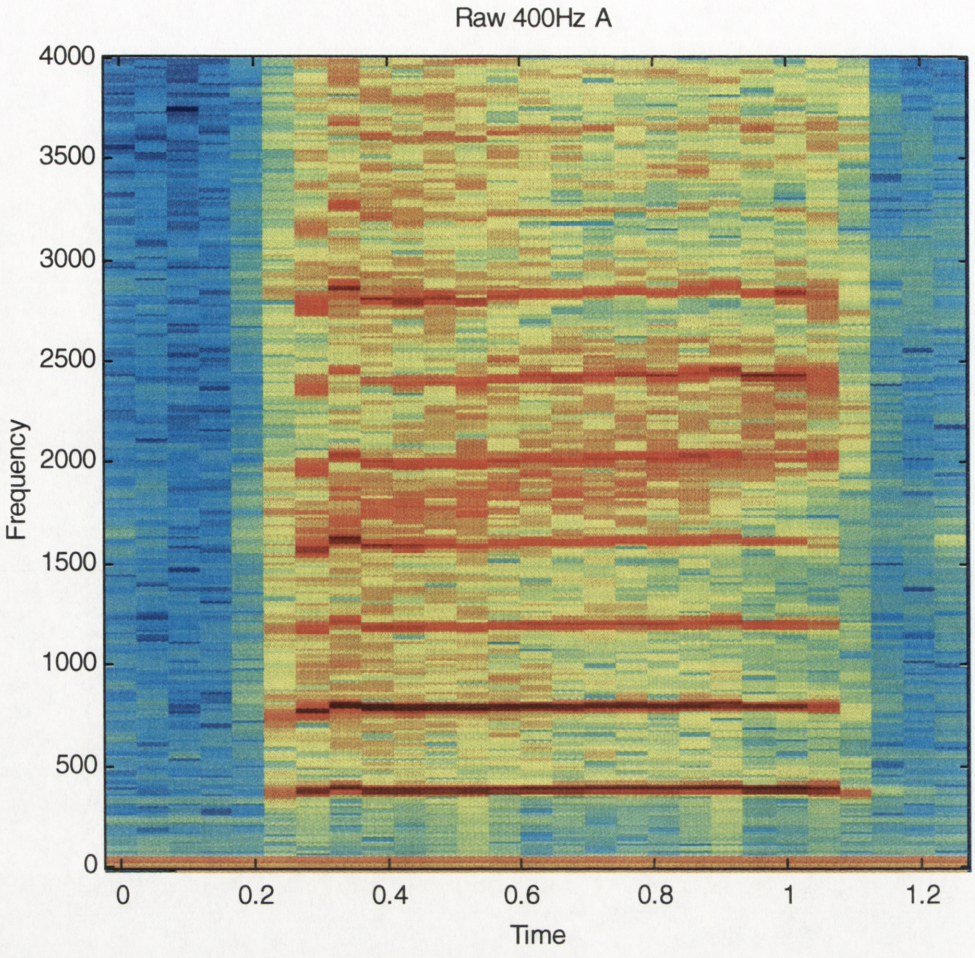

Figure 5-6: Sung A, at approximately 400 Hz. A slowly moving formant may be seen starting at 1500 Hz and drifting up to finish near 2000 Hz.

To analyse this further, we note that the maximum error between a shifted harmonic and a target harmonic (to revert to the nomenclature of Section 4) occurs when the unmodified excitation harmonic is at edge of the passband. Calling this harmonic, the *k*th harmonic in line with our numbering of frequency bands, the relative maximum error is given by

$$E_k = \frac{\phi'_k - (\phi_k + \sigma_k)}{\phi'_k} \qquad (5.7)$$

$$= \frac{p-1}{p(2k+1)}$$

where the ϕ_k and ϕ_k' are the harmonic frequencies of the source signal and the target harmonics respectively.

If p=1, then the expression is zero; that is, there is no error if no modification is performed. When pitch raising is occurring, the maximum error decreases with larger and larger scaling factors. Similar to human pitch perception, the error decreases with increasing frequency band (k).

Another point to note about this method is that if the harmonics are, by luck or design, included in the pass band of the analysis filter then the spectral shape of the harmonic component is retained. In the frequency domain techniques such as the Seneff/Abe compression/expansion method, the harmonic shapes are squeezed or stretched (see [5] , for example).

A low frequency hum may be noticed when the pitch is raised. This effect is caused by shifting up the "DC" component into a perceptible frequency range. A useful ability with bandpass filter banks is that we can treat the frequency bands independently of one another. For example, in the implementation pursued here, the lowest band is not shifted in frequency at all to minimise this degradation.

5.2.2. Harmonically Related Bandpass Filters

The error caused by the excitation harmonics not being centred in the bandpass filters suggests that we should dynamically tune the analysis and synthesis filter centre frequencies based on the instantaneous pitch and the desired target pitch. This requires the explicit extraction of the source fundamental frequency but many simple techniques now exist for calculating this quantity (see, for example Martinez-Alfaro and Contreras-Vidal [10] or Atkinson et al. [12]).

The scheme of section 5.2.1 was modified in order to centre the bandpass analysis filters on the harmonic frequencies of a particular signal, a sung long "a", for which the fundamental is approximately 200Hz. Constant bandwidth filters of 200Hz passband width were again designed using the Remez function of MATLAB. This

implementation is included on the accompanying disk as files hrfb.m and do_hrfb.m. The results can be heard in WAV files on the disk. See Appendix C for a list of locations and file names on the disk.

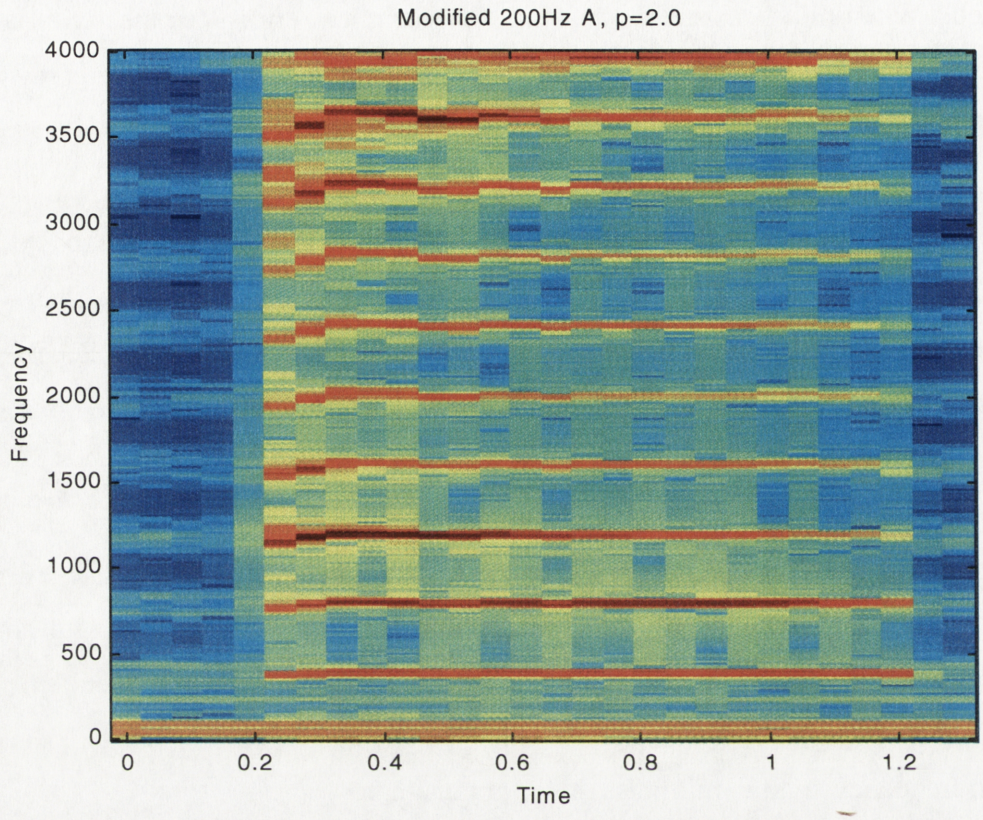

Figure 5-7: Sung A, with pitch raised by a factor of two. Because analysis filters were centred on the original harmonics, the modified inter-harmonic spacings are scaled correctly. The perceived pitch has been increased by a factor of two.

If we ignore the effects of the movement of the formants, we find that the pitching of the modified signal compares favourably with the same vowel sung naturally at the scaled fundamental. This is apparent in Figure 5-6.

In fact, both the pitch raised and pitch lowered signals exhibit very good perceptual pitch when compared with the naturally sung vowels of the same singer at the target pitch. In the pitch lowered case, the volume of the vowel sung at 100 Hz is substantially less than that of the one sung at 200 Hz due to the ability of the singer. This accounts for the stronger sounding voice of the modified ($p=0.5$) signal.

Thus, it appears that the best approach in the time-domain still relies on some form of parameter extraction, and must be based on a valid model of the speech production process.

5.2.3. QMF Based Techniques for Filter Bank Creation

The concept of the filter bank implementation of the STFT has an extensive literature from the past ten years, with many contributions concentrating on the real-time structures needed to efficiently implement them. These techniques make use of the fact that, as the output of each filter is band-limited, it is oversampled and can be decimated without fear of aliasing. Coupled with developments pertaining to quadrature mirror filter banks, the structures of filter-decimator pairs may be efficiently realised by use of polyphase decomposition. This technique permits all of the computations in applying the filter bank to the signal occur at the lowest rate achievable within the given context. This results in dramatic improvements in computational efficiency. Vaidyanathan [30] demonstrates that given certain constraints on the design of the filters, perfect reconstruction systems may be achieved with quite low complexity implementations.

Recently, it has been recognised that these multi-rate filter bank systems have a close connection with the wavelet transform, a transform with a different set of basis functions than the traditional Fourier transform. These ideas are presented by Vaidyanathan and demonstrate direct relationships between the basis properties of the transform and the desirable properties for perfect reconstruction filterbanks. In particular, the *orthonormality* of the basis formed by a group of wavelet functions is implied by an implementation as a QMF filterbank with the *paraunitary* property [30]. Paraunitary filterbanks may be designed to have the property of perfect reconstruction, and this translates to the *completeness* of the wavelet basis.

For some time, the music application field has held that the best sub-band decomposition of a sound is given by non-uniformly spaced filters. The octave-spaced filterbank closely mirrors the decreasing resolution of the human ear with increasing frequency. This non-uniform nature of human hearing is also responsible for the development of the musical scale. Notes are logarithmically spaced in terms of frequency, and two different pitches, one double the other in frequency, are perceived as

the same note. In effect, this means that notes are sparser and intervals, the gaps between notes, are longer with increasing frequency.

For the non-uniform filterbank, the individual bandpass filters are created by frequency scaling a single prototype filter. In the uniform case, the filters were obtained by frequency shifting a single prototype. This arrangement of bandpass filters is also known in the analog world as a constant Q system. The ratio of the bandwidth of each filter to the centre frequency of that filter is constant. So unlike equation (5.5), the bandwidth does depend on k,

$$\frac{BW_k}{f_k} = K \qquad (5.8)$$

where K is constant. This so-called octave spacing has been found to be very useful in the analysis of sound signals because it mirrors the decreasing frequency resolution of the ear with increasing frequency.

In a wavelet transform, however, the resolution is not only non-uniform in the frequency domain, it is also non-uniform with respect to time. To see this, consider the outputs of the STFT (or an equivalent bank of filters) as samples in time-frequency space, illustrated in Figure 5-8 below.

The time and frequency axes are uniformly divided in contrast to the samples for the wavelet transform, shown in Figure 5-9. Here we see frequency samples are closer together at lower frequencies and the corresponding time samples are spaced further apart. The system cannot be considered as a moving window (or uniform segmentation-in-time) but rather as a family of windows. When represented as a filterbank, the unequal time-sampling is achieved by the decimation in each sub-band. The decimation factors are non-uniform due to the differing bandwidths.

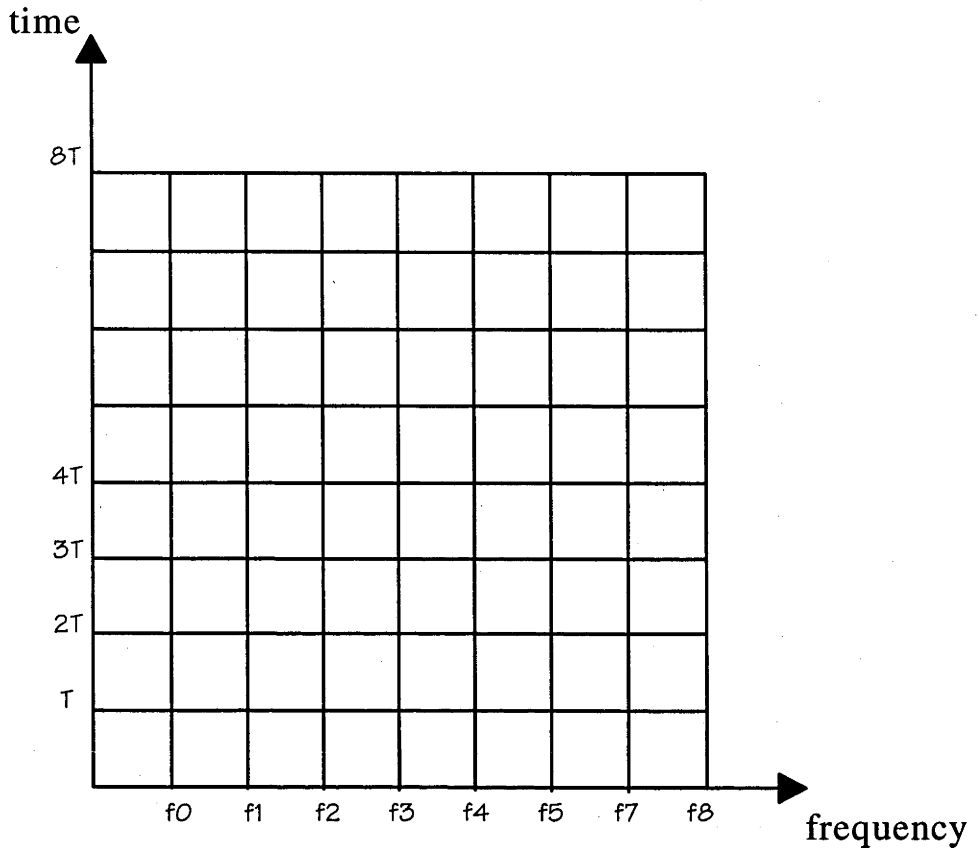

Figure 5-8: Time-Frequency Representation of STFT. The joint time frequency domain is sampled at regular intervals in both domains.

As suggested by Riley [25], the time-frequency resolution of the system is localised for the application at hand. When the non-uniform filterbank is used in a pitch modification system, we will still be affected by the shift error for each source harmonic if the harmonics are not centrally located in each band. If we wish to achieve a natural sounding result, we will need to place the filters on the source harmonics by an explicit pitch extraction. This would require recalculation of the analysis and synthesis filters at the rate of pitch evolution. Although the implementations of QMF-based filterbanks may be very computationally efficient, the *generation* of such a set of filters is still complex.

For example, using QMF techniques with sixteen filters for adequate frequency resolution at low frequencies (the lowest filter would cover the range 0-30Hz) would require four convolutions of the prototype filter per analysis or synthesis filter. If we use sharp cutoff prototypes with, say 64 taps, then we require $16x4x64^2$ multiplies per

filter bank generation. We have an analysis filterbank and a synthesis filterbank to generate if the source and target are time-varying and we need to generate these every 20ms. This translates to 50x2x16x4x642 FLOPs or 26 MFLOPs to generate the filters. As some new DSP devices can perform in the range of 1600 MFLOPs to GFLOPs, this technique is worth further investigation. The device must also perform the pitch estimation, signal filtering and modulation calculations in conjunction with the filter generation, but this presents the opportunity to use parallel processing.

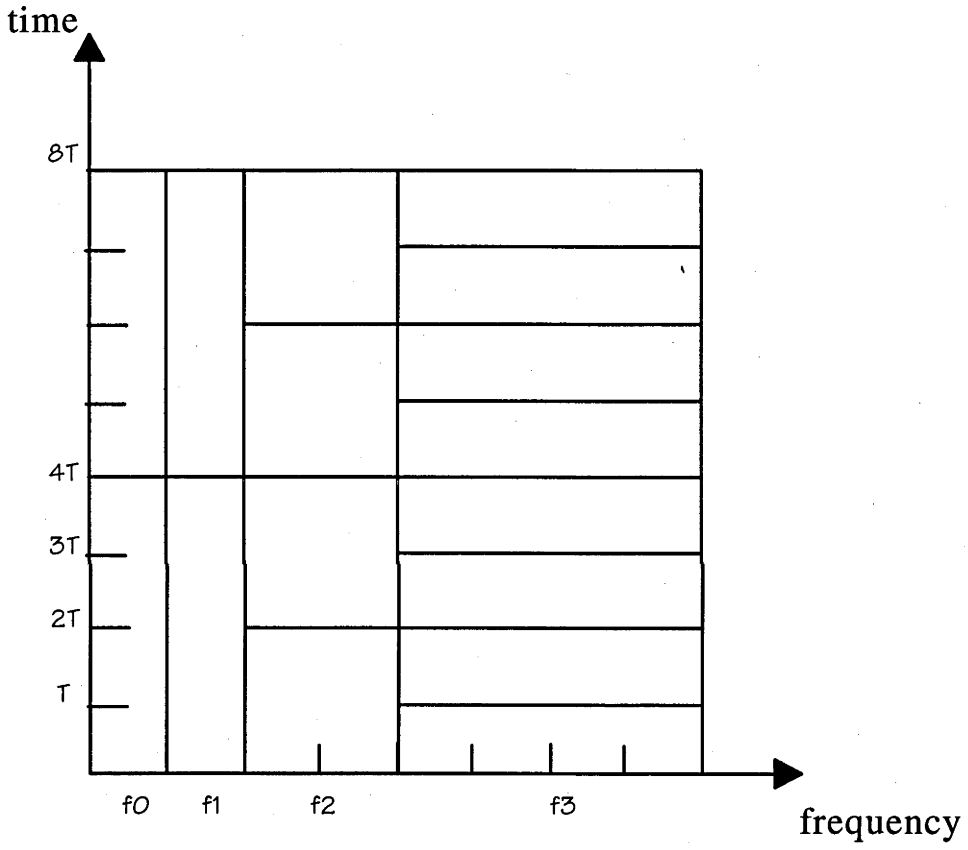

Figure 5-9: Time-Frequency Representation of wavelet transform. The resolution of the transform varies with both time and frequency. This particular arrangement makes frequency resolution decrease with increase frequency which is known to mirror the characteristics of human hearing.

5.2.4. Lowpass Modulation Implementation

We now investigate an alternate technique to the standard array of parallel bandpass filters. The motivation for this is twofold. First, if we need to track the pitch, the centre-frequencies of the analysis and synthesis filters will need to be recalculated as the source fundamental pitch (and perhaps the scaling factor) change with time. We seek a system where a whole array of filters do not need to be computed, instead we use a single low-pass filter for all of the anti-aliasing work.

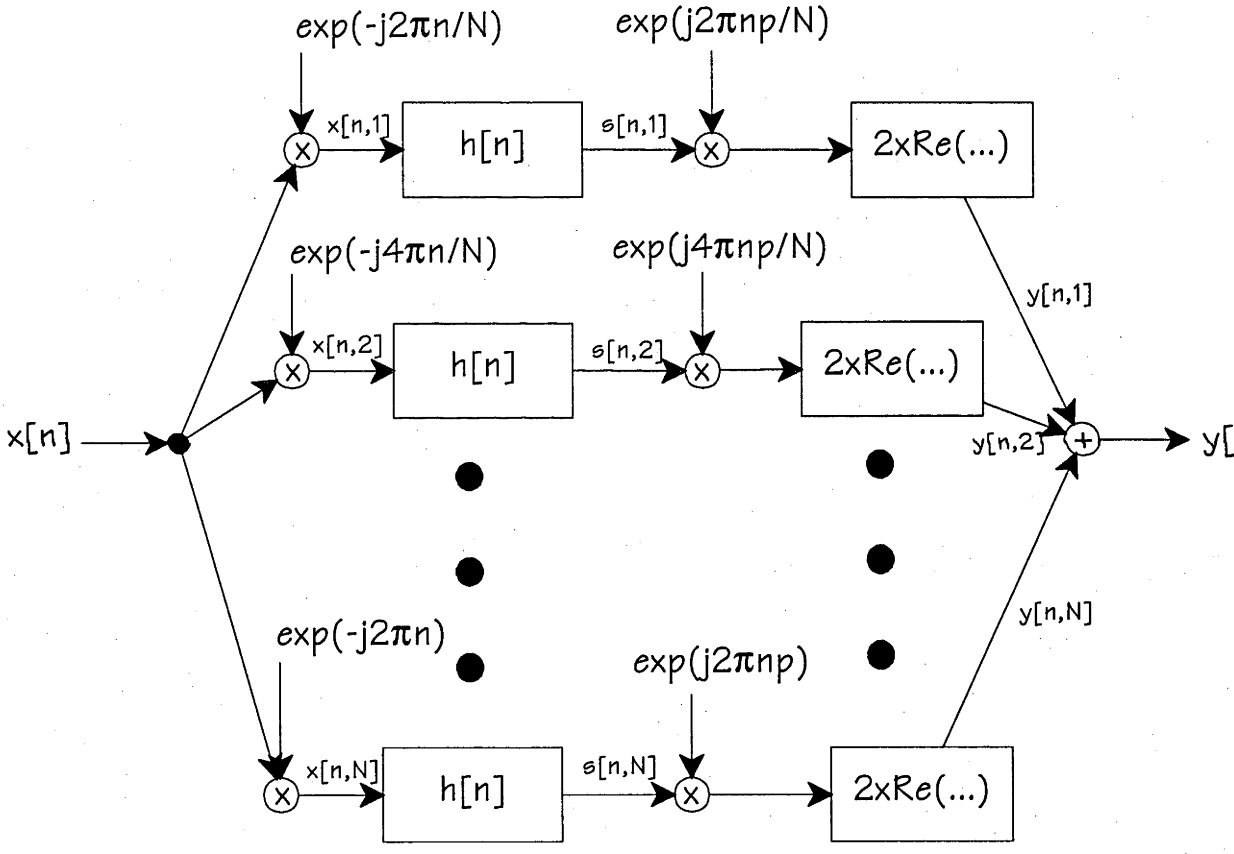

Figure 5-10: Complex modulators with a single low-pass filter. The baseband equivalent of each frequency segment is filtered, and then modulated to a new target frequency. The complex values in the resulting signal are discarded in the synthesis stage.

Secondly, we wish to compare a full sidelobe modulation technique to the SSB method described above. In other words, we shall modulate the individual frequency patches by

complex sinusoids rather than merely their real part as in the cosine modulated case. Such a scheme is presented in Figure 5-10.

Each individual spectral segment is modulated to baseband and low-pass filtered to isolate it. The signal is then modulated to the target pitch. As the re-modulated signals are now complex, a real output is achieved by summing the real parts. This system is heuristically just a relocation of the *Re(.)* operator from about the modulation carrier, to acting on the output signal. It is analogous to the attempts by Griffin and Lim [2] to make modified STFTs valid. A previous discussion of this technique appears in Seneff [33] and in Crochiere and Rabiner [31]. A different justification for discarding the imaginary component follows.

One advantage of this scheme is that in order to alter pitch scaling rate we only have to alter the complex sinusoidal modulator frequencies, and not recalculate two entire banks of filters. We can alter the modulation factors smoothly between successive pitch estimates rather than jumps in analysis spectral segments at temporally segmented moments. This may be achieved by fitting a piece-wise continuous curve to the pitch contour (Milenkovic [40]).

The use of the *Re(.)* operator is essential in order to make the output signal non-complex. We would hope that this technique, which is simple to implement, may result in a signal which contains the desired spectral modifications and sounds natural to the human ear. We show in Appendix B, that if p is unity the system will output the original signal scaled only in magnitude. That is, the modification system exhibits *perfect reconstruction.*

5.2.5. Future Direction

We have looked at constant bandwidth filterbanks and constant Q filterbanks. The placement of each filter with respect to the harmonic excitation of the source has also been examined. We have treated the pitch shift factor as being constant for the purposes of our analysis and experimentation. In practical situations, it is likely that both the source fundamental and the target pitch are varying with time. We wish in this section to examine some of the issues in extending the pitch modification scheme.

From Jafari et al. [38], it is known that the pitch rise speed decreases with increasing initial pitch to a maximum of 160 semitones/s. Pitch lowering speed increases with increasing initial pitch to a max of 230 semitones/s. This is equivalent to maintaining that pitch rise speed decreases with a reduction in the interval being traversed. These maximum pitch shift rates represent the performance of a professional singer with 5 years of training. Untrained singers can produce only produce maximum rates of 100 semitones/s in either case. These results indicate that pitch extraction and modification at the standard quasi-stationary period of 20 milliseconds is sufficient for all amateur and most professional applications.

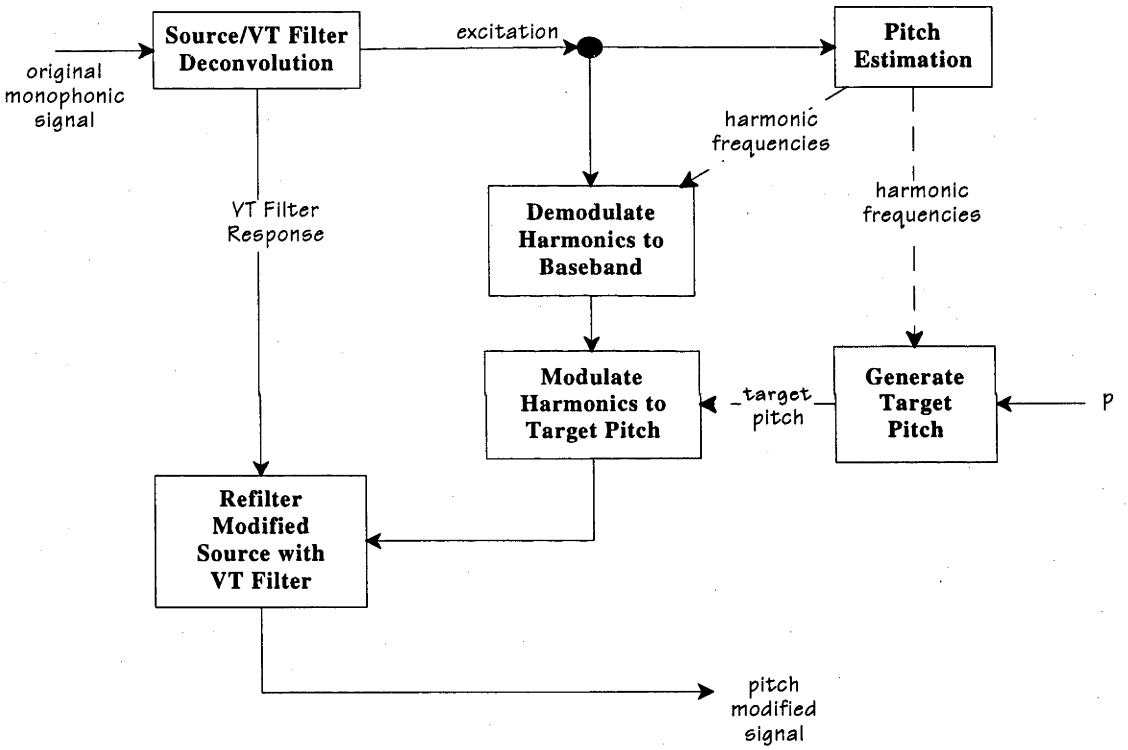

Figure 5-11: Proposed adaptive pitch modification system. After deconvolution of the vocal tract response, the fundamental pitch of the excitation is determined. This information is used to demodulate the frequency segments centred on the source harmonics. The target frequencies for the new harmonics are determined externally, perhaps with reference to a musical score, and the modulation factors are calculated from the ratio of these targets to the original pitch frequencies.

Figure 5-11 illustrates a proposed system to take all of these factors into account. It makes use of the modulating filter technique investigated in Section 5.2.4 but explicitly incorporates a pitch estimation stage and would recalculate the modulation factor every 20ms.

Explicit pitch extraction is common to all the high-quality time-domain methods of pitch scaling in one form or another. From the autocorrelation used in the SOLA (Roucos and Wilgus [32]) to counting zero-crossings as employed by Lent [18]. A particularly onerous method is used by the PSOLA technique (Valbret et al. [11]) where the impulse maxima had to be located. All of these techniques estimate the pitch at discrete moments in time, either at a pitch synchronous rate or at a constant 20 millisecond period.

When lowering the pitch of a signal, although the squishing problem is overcome by the proposed method, the problem of the missing high-frequency portion of the spectrum remains. This may be overcome by repeating the highest down-shifted harmonic at octave spaced frequency until the spectrum is "filled". This suggests that the proposed solution is a parametric method albeit with simple parameter extraction performed in the time domain.

5.3. Adaptive Codebooks

The use of a correlation-based technique in the SOLA method is quite reminiscent of the methods of pitch extraction used in speech coding algorithms, specifically the CELP algorithms. The adaptive codebook technique may be used to simplify or supplant the correlation search techniques set out by Atkinson et al. [12], and Yim and Pawate [34]. The global search used by Yim and Pawate seems specifically inspired by complexity reduction work in the speech coding field.

In the GSM and CELP coding systems, the basic structure is a Linear Prediction Filtering operation cascaded with a long term pitch predictor. The pitch predictor operates as shown in Figure 5-12.

The current segment of the linear prediction residual, from time *0* to *L*, is cross-correlated with a similar length segment from the past. This cross-correlation also involves a search for the optimum gain factor for each segment. All segments within the range 2ms to 18ms in the past are so correlated, and the past segment at time *–M* with maximal correlation, is subtracted from the LPC residual. The operation is reversed for synthesis; that is, past segments are added back into the signal at the appropriate shift, *M*, and gain *g*.

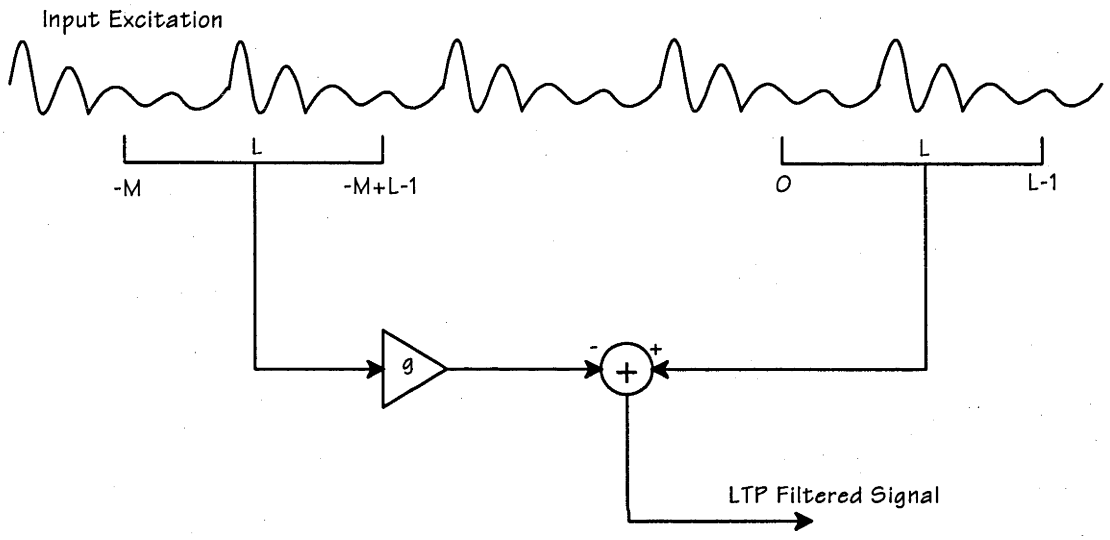

Figure 5-12: Standard Long Term Predictor pitch extraction filter. Similarly to the SOLA method, a cross-correlation maximisation is used as a defacto pitch picker. Frames are then subtracted from the original at the maximally correlating time, M, to reduce the dynamic range and information content of the signal.

A system for pitch modification was implemented based on this LPC/LTP reduction system. Modification of source harmonics by a factor *p* was attempted in the synthesis stage. When the excitation pitch information is being reinstated, instead of adding the past segment at time *M*, the section is added at time *Mp*.

The results of this technique are quite poor. Please refer to the Appendix C for locations of the MATLAB M-files and WAV file data for this method. Considerable artefacts are introduced by the overlapping of segments with different gain values and

by the modified residual being shifted into linear prediction frames with vastly different characteristics.

These effects may be somewhat ameliorated by introducing windowing and overlapping analysis frames. This would make the system similar to the SOLA system of Roucos and Wilgus [32], but with an LPC front end. The failure of this technique brings up some useful points about making use of the LPC residual for pitch modification and treating it as though it were the source signal however.

The LPC and LTP operations are primarily used in systems for bit-rate reduction in communications applications. A principal objective of these systems is a reduction in dynamic range, for the sake of a couple of parameters such as shift and gain. The result is that the excitation signal is devoid of much amplitude to work with when performing the modification step.

Another potential trap in systems consisting of concatenated LPC and pitch modification systems is demonstrated by the artefacts found in this system. If the excitation obtained by filtering with a set of LP coefficients, a_k, is used to excite the a linear prediction filter next in time, a_{k+1}, very noisy spikes and attacks are produced. This is best visualised by imagining the onset of a speech segment where room noise is generated as an excitation in the segment prior to vocalisation. This random excitation, if used in whole or part, as input to the formant structured and amplifying LP synthesis filter is the cause of the unpleasant and unnatural noise bursts apparent in the above implementation.

This also raises a point for the concatenation of filter bank modification methods with LP filters. If the filters introduce any signal delay, the LP coefficient frames must be delayed by a similar amount. Another approach may be to interpolate between successive LP coefficient sets for the instantaneous LP coefficients.

Chapter 6 – Summary and Conclusions

In this final chapter, we present a summary of the key findings of this research and present some conclusions drawn from the author's work.

Whether scaling a voice signal in time or altering its perceived pitch, the most common analysis methods involve the segmentation of the waveform in the time domain. If we wish to avoid this, we must accept a complementary segmentation in the frequency domain. This spectral segmentation can only work when each patch is centred on a feature of interest. In the case of the human voice, these features are the harmonic frequencies of the source.

We have proposed a new method of altering the perceived pitch of human speech and singing, and found that it performs comparably to existing high-quality techniques. This performance is achieved at the cost of tracking the fundamental pitch period of the excitation signal.

The form of this method when considered as a joint time-frequency representation after the style of Riley [25], is that the transform resolution is fixed in the time domain by the length of the filter implementations. In the frequency sense, the resolution is "tuned" to the areas of interest by the pitch tracking and filter recalculation.

This research has also demonstrated that we cannot divorce the analysis process from the physical production model of the sound if we wish the modification process to produce a signal lacking in artefacts. The act of source/filter decomposition is fundamental to high quality modification techniques, and greatly affects the perceived quality of the output. A suitable source extraction method is necessary for the proposed modification implementation. An appropriate method would leave sufficient amplitude in the extracted signal for the modification to be successful. Problems in equalising phase differences between successively modified time segments are avoided by performing all of the operations in the time-domain in a near-continuous fashion. Each band in the proposed scheme is treated by the same filtering operations.

An issue raised by Sundberg [17] is the assumption of vocal tract invariancy. Most high quality methods benefit from deconvolving the source excitation from the vocal tract response (or spectral envelope). The excitation pitch is then independently modified. Moulines and Laroche [6] claim, however, that annoying effects on the timbre of speech occur at moderate pitch scale factors. This is perhaps due to the invariancy assumption; it is not unreasonable to assume that the same voiced sounds at different pitch require slightly different vocal tract responses. A mapping then may be defined to perform a transform on the vocal tract parameters based on the pitch scale factor in addition to the modification of the excitation signal.

Effects of vibrato, trills and other embellishments to singing (and other instruments) make the pitch contour P(n) harder to track in time. One way around this would be to estimate the fundamental pitch over different analysis segments of some appropriate length and then extrapolate from the evolving pitch envelope the sample-by-sample instantaneous pitch. This is analogous to a scheme for estimating instantaneous pitch harmonic frequencies proposed by McAulay and Quatiere [35].

The nature of the entire speech (and arbitrary waveform) modification is entirely subjective. We have achieved a high quality result if we can fool most of the people most of the time. (Coding and CELP make use of subjective anomalies all the time). Several issues present themselves here.

As can be seen from the literature survey, very few psycho-acoustically valid tests have been performed on the techniques expounded. A more thorough evaluation and comparison of existing techniques would no doubt be valuable in providing feedback at this point into which methods are appropriate for which type of signals. Confusion has already been seen in Quatieri and McAulay [36], where a technique based entirely on the assumptions of the human speech production model in fact achieved similar results on polyphonic voices and music. (Adaptation of whale speech is also mentioned but how the quality was assessed in this case remains a mystery).

Several other subjective phenomenon related to human perception of pitch are known from the field of psycho-acoustics. Well known effects such as Shepard tones derive

from constantly varying tones, but other effects, such as those examined by Schroeder [26] may bear some fruit in finding ways of tricking the human ear, and mind, into believing a pitch modification has taken place. The use of a perceptual weighting filter is widely used in speech coding, and has also been applied to speech modification by Abe et al. [21]. This technique may be also of use in improving the quality of low-complexity algorithms.

The step of separating the vocal tract filter response phase from that of the excitation phase leads to implementation difficulties in many cases from the literature. The solution most recently shown to produce high quality results is to position the analysis frames in a pitch synchronous fashion. In this thesis, a time domain system has been proposed which remains pitch synchronous by frequency-domain segmentation and pitch estimation guiding this segmentation. It seems to be self-evident that in performing pitch modification, the source frequencies must be found in order to scale them to the correct target pitch.

To sum up, three main issues are of paramount importance in the analysis of the human voice for pitch modification purposes. These are the time frequency resolution of the representation used in the modification step, the separation of source waveform from the vocal tract response and the estimation of the local pitch period. In the synthesis of a pitch modified signal the primary consideration should be the subjective response of the listener. Commercial devices have demonstrated that heuristic, seemingly under-analysed techniques can perform both time and pitch modifications with high quality.

Areas of future work include:

- Investigation of source/filter deconvolution methods suitable for use with pitch scaling techniques to be applied to the excitation residual
- Investigation of the assumption of vocal tract invariancy with changes in excitation fundamental pitch
- Application of existing techniques to cases of time-varying fundamental frequencies and time-varying modification factors
- A rigorous subjective evaluation of competing modification implementations
- Implementation of the adaptive system proposed in Section 5.2.5

References

1. C. d'Alessandro, "Time-frequency speech transformation based on an elementary waveform representation", *Speech Communication (Netherlands)*, vol. 9, no. 5-6, pp. 419-431, 1990.

2. D. Griffin, J. Lim, "Signal Estimation from Modified Short-Time Fourier Transform", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 2, pp. 236-242, 1984.

3. D. Malah, "Time-Domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Speech", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-27, pp. 311-323, 1979.

4. E. Burns, "Circularity in relative pitch judgements for inharmonic complex tones: The Shepard demonstration revisited, again", *Perception and Psychophysics*, vol. 30, no. 5, pp. 467-472, 1981.

5. E. Moulines, F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones", *Speech Communication (Netherlands)*, vol. 9, no. 5-6, pp. 453-467, 1990.

6. E. Moulines, J. Laroche, "Non-Parametric Techniques for Pitch-Scale and Time-Scale Modification of Speech", *Speech Communication (Netherlands)*, vol. 16, pp. 175-205, 1995.

7. H. Fujisaki, K. Hirose, K. Shimizu, "A New System for Reliable Pitch Extraction of Speech", *Proc. IEEE 1987 International Conference on Acoustics, Speech and Signal Processing (ICASSP-87)*, pp. 2422-5, 1987.

8. H. Kuwabara, K. Ohgushi, "Spectral manipulation by analysis-synthesis method and the perception of speaker", *NHK Tech. Journal*, vol. 39, no. 1, pp. 25-33, 1987.

9. H. Kuwabara, T. Takagi, "Acoustic Parameters of Voice Individuality and Voice Quality Control by Analysis-Synthesis Method", *Speech Communication (Netherlands)*, vol. 10, no. 5-6, p. 491-495, 1991.

10. H. Martinez-Alfaro, J. Contreras-Vidal, "A Robust Real-Time Pitch Detector Based on Neural Networks", *Proc. IEEE 1991 International Conference on Acoustics, Speech and Signal Processing (ICASSP-91)*, pp. 521-523, 1991

11. H. Valbret, E. Moulines, J. Tubach, "Voice Transformation Using PSOLA Technique", *Speech Communication (Netherlands)*, vol. 11, pp. 175-187, 1992.

12. I.Atkinson, A. Kondoz, B. Evans, "Pitch detection of speech signals using segmented autocorrelation", *Electronics Letters*, vol. 31, no. 7, pp 533-535, 1995

13. I.Pollack, "Continuation of auditory frequency gradients across temporal breaks: The auditory Poggendorff", *Perception and Psychophysics*, vol. 21, no. 6, pp. 563-568, 1977.

14. J. Chen, "Toll Quality 16 kb/s CELP Speech Coding with Very Low Complexity", *Proc. IEEE 1995 International Conference on Acoustics, Speech and Signal Processing (ICASSP-95)*, pp. 9-12, 1995

15. J. Ford, "Pitch Shifting and Time Scaling of Sampled Speech", BE Thesis, Australian National University, 1994.

16. J. Proakis, C. Rader, F. Ling, C. Nikias, *Advanced Digital Signal Processing*, Macmillan,1992.

17. J. Sundberg, "How Can Music be Expressive?", *Speech Communication (Netherlands)*, vol. 13, no. 1-2, pp. 239-253, 1993.

18. K. Lent, "An Efficient Method for Pitch Shifting Digitally Sampled Sounds", *Computer Music Journal*, vol. 13, no. 4, pp. 65-71, 1989.

19. L. Thorpe, B. Shelton, "Subjective Test Methodology: MOS vs. DMOS in Evaluation of Speech Coding Algorithms", *Proc. 1993 IEEE Workshop on Speech Coding for Telecommunications*, 1993.

20. M. Abe, S. Tamura, H. Kuwabara, "A New Speech Modification Method by Signal Reconstruction", *Proc. IEEE 1989 International Conference on Acoustics, Speech and Signal Processing (ICASSP-89)*, pp. 592-5, 1989.

21. M. Abe, S. Tamura, H. Kuwabara, "A Speech Modification Method by Signal Reconstruction Using Short-Term Fourier Transform", *Systems and Computers in Japan*, vol. 21, no. 10, pp. 26-34, 1990.

22. M. Benolke, C. Swanson, "The effect of pitch-related changes on the perception of sung vowels", *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1781-1785, 1990

23. M. Portnoff, "Short-Time Fourier Analysis of Sampled Speech", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-29, no. 3, pp. 364-373, 1981.

24. M. Portnoff, "Time-Scale Modification of Speech Based on Short-Time Fourier analysis", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-29, no. 3, pp. 374-390, 1981.

25. M. Riley, *Speech Time-Frequency Representations*, Kluwer Academic Publishers, 1989.

26. M. Schroeder, "Auditory paradox based on fractal waveform", *Journal of the Acoustical Society of America*, vol. 79, no. 1, pp. 186-189, 1986

27. N. de Bruijn, "Uncertainty principle in Fourier analysis", *Inequalities*, O. Sisha (Ed.), Academic Press, New York, pp. 57-71, 1965.

28. N. Seiyama, T. Takagi, T. Umeda, E. Miyasaka, "A New Method of Pitch Modification by Complex Cepstrum Analysis-Synthesis", *Electronics and Communications in Japan, Part 3 (Fundamental Electronic Science)*, vol. 75, no. 11, pp. 102-113, 1992

29. A.Oppenheim, R. Schafer, *Discrete Time Signal Processing*, Prentice Hall,1989.

30. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, 1993.

31. R. Crochiere, L. Rabiner, *Multirate Digital Signal Processing*, Prentice-Hall, 1983.

32. S. Roucos, A. M. Wilgus, "High Quality Time-Scale Modification for Speech", *Proc. IEEE 1985 International Conference on Acoustics, Speech and Signal Processing (ICASSP-85)*, pp. 493-496, 1985.

33. S. Seneff, "System to Independently Modify Excitation and/or Spectrum of Speech Waveform Without Explicit Pitch Extraction", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-30, no. 4, 1982.

34. S. Yim, B. Pawate, "Computationally Efficient Algorithm for Time Scale Modification (GLS-TSM)", *IEEE Transactions on Signal Processing*, pp. 1009-12, 1996.

35. T. Quatieri, R. McAulay, "Shape Invariant Time-Scale and Pitch Modification of Speech", *IEEE Transactions on Signal Processing*, vol. 40, no. 3, pp. 497-510, 1992.

36. T. Quatieri, R. McAulay, "Speech Transformations Based on a Sinusoidal Representation", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-34, no. 6, pp. 1449-1464, 1986.

37. W. Kleijn, "On the Periodicity of Speech Coded with Linear-Prediction Based Analysis by Synthesis Coders", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 539-542, 1994.

38. M. Jafari, K. Wong, K. Behbani, G.Kondraske, "Performance Characterisation of human pitch control system: An acoustic approach", *Journal of the Acoustical Society of America*, vol.85, no. 3, pp. 1322-1328, 1989.

39. R. Meddis, M. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification", *Journal of the Acoustical Society of America*, vol.89, no. 6, pp. 2866-2882, 1991.

40. P. Milenkovic, "Voice source model for continuous control of pitch period", *Journal of the Acoustical Society of America*, vol.93, no. 2, pp. 1087-1096, 1993.

41. R. Bristow-Johnson, "A detailed analysis of a time-domain formant-corrected pitch-shifting algorithm", *Journal of the Audio Engineering Society*, vol.43, no. 5, pp. 340-352, 1995.

42. L. Rabiner, R. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

43. J. Johnston, "A filter family designed for use in quadrature mirror filter banks *Proc. IEEE 1980 International Conference on Acoustics, Speech and Signal Processing (ICASSP-80)*, pp. 291-294, 1980.

Appendix A: Development Account

During the course of this research, several published techniques were implemented and several new techniques were developed for time and pitch scaling sampled voice. This Appendix relates the course of this development work and provides some pseudo-code illustrations of the techniques implemented.

Due to initial constraints of hardware (a 386DX-33), the Lent algorithm (Section 4.3.1.1) and the Griffin-Lim algorithm (Section 4.2.2.2) were initially implemented with a C++ program written for Windows 3.11 and making use of the Borland C++ compiler V4.1 and accompanying Object Windows Library. The code for this is of historical interest only and is included on the accompanying disk in the /misc subdirectory.

The following is a pseudo-code translation of the Lent algorithm (Section 4.3.1.1) which is capable of both time and pitch scaling. This code was also implemented as a Matlab M-file which is available on the accompanying disk. The input data array, inData, has length inlen and is real-valued, as is the output array outData. The time scaling factor is denoted by tf and the pitch scaling factor by pf. These conventions are maintained for all of the pseudo-code in this Appendix.

```
x = 0
xold = 0
oldzero = 0
outcount = 0
outlen = tf * inlen

for incount = 0 to inlen {
        xold = x
        x = lowpass_filter( inData )

        if ( x <= 0 ) AND ( xold > 0 ) {
          periodlength = incount – oldzero
          oldzero = incount
```

```
            while ( outcount / outlen < incount / inlen ) {
              outcount = outcount + pf * periodlength
              for framecount = -periodlength to +periodlength {
                outData[outcount + framecount] += window( framecount ) *
                                                  inData[incount + framecount]
              } end for framecount ...
            } end while ...
          } end if ...
} end for incount ...
```

The low pass filter was implemented as a simple running average as suggested by Lent's paper. Each period was windowed with a Hamming window of length 2*periodlength. The above algorithm pays no attention to the boundaries of the signal arrays.

The following pseudo-code is adapted from the C++ program implementation of the Griffin and Lim technique for time scaling (Section 4.2.2.2).

```
numframes = outlen / framelen
numiters = 100
Ss = Sa / tf
outlen = inlen / tf
fill_gauss( outData )
for currentframe = 0 to numframes {
  targetfreq = fft( window( inData[ currentframe*Sa...currentframe*Sa + framelen ] ) )
  outwindow = outData( currentframe*Ss...currentframe*Ss + framelen )
  for iter = 0 to numiters {
    outfreq = fft( window( outwindow ) )
    estfreq = polar( mag( targetfreq ) , arg( outfreq ) )
    outwindow = window( RE[ ifft( estfreq ) ] )
  } end for iter ...
  outData[currentframe*Ss...currentframe*Ss + framelen] += gain*outwindow
} end for currentframe ...
```

Best results are obtained if the synthesis window shift Ss is related to the frame length by Ss/framelen <= ¼. In the implementation developed, the FFT has a length of 1024 points and each frame is of length 256 samples (or 32ms at 8kHz). The fill_gauss

function fills the output array with gaussian distributed values of mean 0 and variance 100. The windowing function used is the modified Hanning window introduced by Griffin and Lim to avoid scaling problems in the iteration steps.

Note the gain factor in the overlap add step after the iterations have finished. This gain term is necessary to avoid the amplitude of the signal from exceeding the dynamic range and causing clipping in the 16 bit WAV files used for sample storage.

The initial investigations of filterbank approaches to voice modification revolved around the use of quadrature mirror filter pairs (see Johnston [43]). These efforts are included in the files qmfsplit.m, qmfcomb.m, qmf16.m and qmf16f.m. This development was based around the QMF filters described in [43], with the split and comb files providing sub-band decomposition and recombination respectively. The file qmf16f.m contains a sample-by-sample implementation of this method incorporating pitch modification by means of complex modulation. These experiments proved cumbersome and of poor quality.

The filterbank system investigated next involved designing a set of bandpass filters using the Remez filter design technique (Sections 5.2.1 and 5.2.2). Each filter was designed to have 3dB of passband ripple and 40dB of stopband rejection. The band 0 filter was a low pass filter with a 200Hz cutoff. Bands 1 to 18 were band pass designs with centre frequencies, Fc, of 300, 500, 700, ... 3700 Hz. The band 19 filter was a high pass design with lower cutoff frequency of 3800 Hz.

The modification step for pitch shifting the decomposed sub-bands involved a single side band modulation of each sub-band signal. The modulated sub-bands are then filtered with manually created synthesis filters to eliminate aliasing effects and the sub-bands added to produce the output. The following pseudo-code illustrates this method, denoted ALFB (Arbitrarily Located Filter Banks). The filters are created in an array h[0 to 19], and for a particular pitch scaling factor the synthesis filters hmod[0 to 19] are created with shifted centre frequencies given by fc[bank]*pf.

```
for count = 0 to inlen {
  for band = 0 to 19 {
    in[band] = filter( h[band], inData( count ) )
    inshift[band] = cos( 2*pi*(pf-1)*fc[band]*count )
    mod[band] = in[band] * inshift[band]
    out[band] = filter( hmod[band], mod[band] )
    outData(count) += gain*out[band]
  } end for band ...
} end for count ...
```

Note the gain term again necessary to avoid saturation problems in limited range representations. It may be necessary to remove some bands from the synthesis step to avoid the "squishing" effect noted in Section 5.2.1.

Another method then tried was to manually place the filter centre frequencies over the source harmonics (Section 5.2.2). This method, denoted HRFB (Harmonically Related FilterBank) is quite similar to the ALFB technique and involved the manual placement of the analysis filters at the source harmonic frequencies. This was done for a single signal where the source was a male singer singing a vowel at a 200 Hz fundamental. The bandpass centre frequencies were set at 200, 400, 600, ..., 3800 Hz and the synthesis filters were at centre frequencies again scaled by the pitch factor. The HRFB algorithm is otherwise identical to that of ALFB.

A related technique was implemented involved modulating the sub-bands to baseband, where a single low pass filter is used to isolate the sub-band (Section 5.2.4). The sub-band is then shifted to the target frequency and the real parts recombined to form the output signal. The lowpass filter used was designed as a 6^{th} order elliptical filter with 3dB of passband ripple, 50dB of stopband rejection and a 250Hz cutoff. This method, implemented in the modfilt.m Matlab script, is illustrated in the following panel.

Note that some bands may have to be discarded before adding to the output array as in the ALFB and HRFB cases. In particular the apparent amplification of DC "hum" may become annoying when scaling the pitch upwards, and this band may be omitted in the synthesis step.

```
for count = 0 to inlen {
  for band = 0 to numbands {
    inshift[band] = cos(-2*pi*250*(band-1)*count) + j*sin(-2*pi*250*(band-1)*count)
    outshift[band] = cos(2*pi*250*pf*(band-1)*count) + j*sin(2*pi*250*pf*(band-1)*count)
    in[band] = inData[count] * inshift[band]
    in[band] = filter( h, in[band] )
    out[band] = in[band] * outshift[band]
    outData[count] += 2 * RE[ out[band] ]
  } end for band ...
}end for count ...
```

The Adaptive CodeBook (Section 5.3) technique was also implemented with Matlab, and the implementations may be found in the /acb/src subdirectory on the disk. This involved implementing the Levinson-Durbin recursion for linear prediction and the pitch estimation method used in CELP (Code Excited Linear Prediction). This method proved to be of poor quality suffering from gain synchronisation problems.

In the file sfblpc.m (in the /misc subdirectory on the accompanying disk) the sample-by-sample QMF pitch modification technique was attempted on an LPC residual signal. This also proved to be of poor quality suggesting that the linear prediction is not a good source/filter decomposition method for us in pitch modification algorithms.

Appendix B: Sketch Proof of Perfect Reconstruction in Modfilt Pitch System

Each isolated spectrum band, frequency shifted to base band, may be expressed as

$$x[n,k] = x[n]e^{-j\omega_f nk} \tag{B.1}$$

The convolution of these bands with the low-pass filter, *h[n]*, is given by

$$\begin{aligned} s[n,k] &= \sum_{m=-\infty}^{n} x_k[m]h[n-m] \\ &= \sum_{m=-\infty}^{n} x[m]e^{-j\omega_f mk}h[n-m] \end{aligned} \tag{B.2}$$

The remodulated bands may then be expressed as

$$\begin{aligned} y[n,k] &= e^{j\omega_f nkp}S[n,k] \\ &= e^{j\omega_f nkp}\sum_{m=-\infty}^{n} x[m]e^{-j\omega_f mk}h[n-m] \end{aligned} \tag{B.3}$$

And the final synthesis signal is then

$$\begin{aligned} y[n] &= 2\sum_{k=0}^{N} \mathrm{Re}\{y[n,k]\} \\ &= 2\sum_{k=0}^{N} \mathrm{Re}\left\{e^{j\omega_f nkp}\sum_{m=-\infty}^{n} x[m]e^{-j\omega_f mk}h[n-m]\right\} \end{aligned} \tag{B.4}$$

To prevent aliasing in the reconstruction, we require the DFT to be one-sided, that is

$$X(e^{j\omega}) = 0 \quad \textbf{for} \quad -\pi \le \omega < 0 \tag{B.5}$$

This implies, by the Hilbert relations, that

$$X_R(e^{j\omega}) = \frac{1}{2}\left[X(e^{j\omega}) + X^*(e^{-j\omega})\right] \tag{B.6}$$

Thus to ensure a one sided DFT, we require

$$\begin{aligned} X(e^{j\omega}) &= 2X_R(e^{j\omega}) & 0 \le \omega < \pi \\ &= 0 & -\pi \le \omega < 0 \end{aligned} \tag{B.7}$$

The low pass filter has finite length L; that is

$$h[n] = 0 \quad \textbf{for} \quad n < 0, n > L-1 \tag{B.8}$$

then

$$s[n,k] = \sum_{m=n-L}^{n} x[m] e^{-j\omega_f mk} h[n-m] \tag{B.9}$$

Which is bandlimited by the impulse response, *h[n],* which by design is low-pass.

With the filters equally spaced in the frequency dimension,

$$\begin{aligned}\omega_k &= \frac{2\pi k}{N} \\ &= \omega_f k\end{aligned} \qquad \textbf{(B.10)}$$

If we define the DFT of the windowed sequence to be

$$X[n,k] = \sum_{m=0}^{L-1} x[n+m]h[m]e^{-j\left(\frac{2\pi}{L}\right)km} \qquad \textbf{(B.11)}$$

then

$$\begin{aligned}s[n,k] &= e^{j\omega_f L} X[n,k] \\ &= X[n,k]\end{aligned} \qquad \textbf{(B.12)}$$

and the output bands simplify to

$$y[n,k] = e^{j\omega_f nkp} X[n,k] \qquad \textbf{(B.13)}$$

This equation simply represents the fact that the output is the shifted DFT.

If p is unity, each $s[n,k]$ is the instantaneous kth DFT point. We want only the one-sided DFT to avoid aliasing in the reconstruction so we take twice the real part of this DFT in the summation of the y_k.

Using the IDFT on (B.11) we have

$$x[n+m] = \frac{1}{Nh[m]} \sum_{k=0}^{N-1} X[n,k]e^{j\left(\frac{2\pi k}{L}\right)m} \qquad \textbf{(B.14)}$$

and since, by (B.13), the system output is

$$y[n] = \sum_{k=0}^{N} X[n,k] e^{j\left(\frac{2\pi k}{L}\right)np} \qquad \textbf{(B.15)}$$

then

$$x[n+m] = \frac{1}{Lh[m]} y[n] e^{m} \qquad \textbf{(B.16)}$$

If we set m to zero, without loss of generality as the signal is causal, we have

$$x[n] = \frac{1}{h[0]} y[n] \qquad \textbf{(B.17)}$$

Thus

$$y[n] = h[0]x[n] \qquad \textbf{(B.18)}$$

The output signal consists of each input sample, scaled by the first coefficient of the low pass filter impulse response.

Appendix C: Disk Index

The accompanying 3.5" diskette is in IBM format and contains two types of files. The first are Matlab scripts, or M-files, denoted by the use of a ".m" suffix. These were written for MATLAB version 5.1 for Windows and are ASCII format with Intel byte ordering.

The second type of file on the diskette are sample files, containing example modifications and stored in the WAV format for audio files. All files are mono with 16 bit samples obtained at a rate of 8kHz. These files are denoted by the ".wav" suffix.

Each modification method which was implemented as part of this thesis has a directory on the disk. Two published techniques are represented, one by Griffin and Lim, the other by Lent. Examples of modified speech and singing are included with the MATLAB scripts to be found in the "/src" subdirectory.

The following table indexes all of the files on the disk.

Technique	Sub-directory	Filename	Description
grifflim	speech	glt05.wav	"Oak is strong and also gives shade". Time compressed using the phase vocoder frequency domain method with signal reconstruction by Griffin & Lim's method, t=0.5
		glt10.wav	"Oak..." sentence, t=1.0
		glt20.wav	"Oak..." sentence, t=2.0
		glthov05.wav	"My hovercraft is full of eels". Time compressed as above, t=0.5
		glthov10.wav	"... hovercraft ..." sentence, t=1.0
		glthov20.wav	"... hovercraft ..." sentence, t=2.0
	src	grifflimt.m	MATLAB M-file implementing Griffin & Lim style time scale modification
lent	speech	p05.wav	"Oak..." sentence. Pitch scaled using

Technique	Sub-directory	Filename	Description
			the Lent algorithm, p=0.5
		p20.wav	"Oak..." sentence, p=2.0
		t05.wav	"Oak..." sentence. Time scaled using the Lent algorithm, t=0.5
		t20.wav	"Oak..." sentence, t=2.0
	src	lent.m	MATLAB M-file implementing Lents algorithm for time and pitch scale modification
al-fb	speech	bp05.wav	"Oak..." sentence. Pitch scaled using the arbitrarily located filterbank algorithm, p=0.5
		bp10.wav	"Oak..." sentence, p=1.0
		bp20.wav	"Oak..." sentence, p=2.0
	singing	a200_1up	Human male voice singing an "A" at 200 Hz, pitch scaled to 400Hz, p=2.0
	src	alfb.m	MATLAB M-file implementing analysis and synthesis filterbanks with 20 sub-bands
		do_Alfbu.m	MATLAB M-file for doubling the pitch of a signal
		do_Alfbd.m	MATLAB M-file for halving the pitch of a signal
	squish	bp05sqsh.wav	"Oak..." sentence pitch scaled with p=0.5, but synthesis filters allowed to overlap
hr-fb	speech	bp05.wav	"Oak..." sentence. Pitch scaled using the harmonically related filterbank algorithm, p=0.5
		bp20.wav	"Oak..." sentence, p=2.0
	singing	a200_1up.wav	Human male voice singing an "A" at 200 Hz, pitch scaled to 400Hz, p=2.0
		a200_2up.wav	Sung "A" scaled from 200 to 400 Hz, p=2.0
		a400_1dn.wav	Sung "A" scaled from 400 to 200 Hz, p=0.5
		a400_1dn-regen.wav	Sung "A" scaled from 400 to 200 Hz, with high frequency harmonics regenerated, p=0.5
	src	hrfb.m	MATLAB M-file implementing analysis and synthesis filterbanks with 21 harmonically related sub-bands

Technique	Sub-directory	Filename	Description
		do_hrfbu.m	MATLAB M-file for doubling the pitch of a signal
		do_hrfbd.m	MATLAB M-file for halving the pitch of a signal
modfilt	speech	mf05.wav	"Oak..." sentence. Pitch scaled using the complex modulation algorithm, p=0.5
		mf10.wav	"Oak..." sentence, p=1.0
		mf20.wav	"Oak..." sentence, p=2.0
	singing	a200_1up.wav	Human male voice singing an "A" at 200 Hz, pitch scaled to 400Hz, p=2.0
		a200_1dn.wav	Sung "A" scaled from 200 to 100 Hz, p=0.5
		a400_1dn.wav	Sung "A" scaled from 400 to 200 Hz, p=0.5
	src	modfilt.m	MATLAB M-file implementing complex modulation with single low pass filter technique for pitch scaling
acb	speech	acb050.wav	"Oak..." sentence. Pitch scaled using the adaptive codebook algorithm, p=0.5
		acb075.wav	"Oak..." sentence, p=0.75
		acb100.wav	"Oak..." sentence, p=1.0
		acb150.wav	"Oak..." sentence, p=1.5
		acb200.wav	"Oak..." sentence, p=2.0
	src	do_ltp.m	MATLAB M-file implementing adaptive codebook technique for pitch scaling
misc		repitch.prj	Borland project file for C++ code
		qmf*.m	Matlab M-files investigating Quadrature Mirror Filterbank approaches to pitch scaling