Concatenative Speech Synthesis: A Framework for Reducing Perceived Distortion when using the TD-PSOLA Algorithm

Jennifer Ann Longster

A thesis submitted in partial fulfilment of the requirements of Bournemouth University for the degree of Doctor of Philosophy

May 2003

Bournemouth University

BLANK IN ORIGINAL

Abstract

Concatenative Speech Synthesis: A Framework for Reducing Perceived Distortion when using the TD-PSOLA Algorithm

Author: Jennifer Ann Longster

This thesis presents the design and evaluation of an approach to concatenative speech synthesis using the Time-Domain Pitch-Synchronous OverLap-Add (TD-PSOLA) signal processing algorithm. Concatenative synthesis systems make use of pre-recorded speech segments stored in a speech corpus. At synthesis time, the 'best' segments available to synthesise the new utterances are chosen from the corpus using a process known as *unit selection*. During the synthesis process, the pitch and duration of these segments may be modified to generate the desired prosody. The TD-PSOLA algorithm provides an efficient and essentially successful solution to perform these modifications, although some perceptible distortion, in the form of 'buzzyness', may be introduced into the speech signal.

Despite the popularity of the TD-PSOLA algorithm, little formal research has been undertaken to address this recognised problem of distortion. The approach in the thesis has been developed towards reducing the perceived distortion that is introduced when TD-PSOLA is applied to speech.

To investigate the occurrence of this distortion, a psychoacoustic evaluation of the effect of pitch modification using the TD-PSOLA algorithm is presented. Subjective experiments in the form of a set of listening tests were undertaken using word-level stimuli that had been manipulated using TD-PSOLA. The data collected from these experiments were analysed for patterns of co-occurrence or correlations to investigate where this distortion may occur.

From this, parameters were identified which may have contributed to increased distortion. These parameters were concerned with the relationship between the spectral content of individual phonemes, the extent of pitch manipulation, and aspects of the original recordings.

Based on these results, a framework was designed for use in conjunction with TD-PSOLA to minimise the possible causes of distortion. The framework consisted of a novel speech corpus design, a signal processing distortion measure, and a selection process for especially problematic phonemes. Rather than phonetically balanced, the corpus is balanced to the needs of the signal processing algorithm, containing more of the adversely affected phonemes. The aim is to reduce the potential extent of pitch modification of such segments, and hence produce synthetic speech with less perceptible distortion.

The signal processing distortion measure was developed to allow the prediction of perceptible distortion in pitch-modified speech. Different weightings were estimated for individual phonemes, trained using the experimental data collected during the listening tests. The potential

benefit of such a measure for existing unit selection processes in a corpus-based system using TD-PSOLA is illustrated. Finally, the special-case selection process was developed for highly problematic voiced fricative phonemes to minimise the occurrence of perceived distortion in these segments.

The success of the framework, in terms of generating synthetic speech with reduced distortion, was evaluated. A listening test showed that the TD-PSOLA balanced speech corpus may be capable of generating pitch-modified synthetic sentences with significantly less distortion than those generated using a typical phonetically balanced corpus. The voiced fricative selection process was also shown to produce pitch-modified versions of these phonemes with less perceived distortion than a standard selection process. The listening test then indicated that the signal processing distortion measure was able to predict the resulting amount of distortion at the sentence-level after the application of TD-PSOLA, suggesting that it may be beneficial to include such a measure in existing unit selection processes.

The framework was found to be capable of producing speech with reduced perceptible distortion in certain situations, although the effects seen at the sentence-level were less than those seen in the previous investigative experiments that made use of word-level stimuli. This suggests that the effect of the TD-PSOLA algorithm cannot always be easily anticipated due to the highly dynamic nature of speech, and that the reduction of perceptible distortion in TD-PSOLA-modified speech remains a challenge to the speech community.

TABLE OF CONTENTS

BLANK IN ORIGINAL

LIST OF FIGURES

LIST OF TABLES

BLANK IN ORIGINAL

Acknowledgements

This research was funded by Bournemouth University.

First thanks go to my supervisors, Dr Martin Lefley and Dr Chris Cowley who have supported and advised me throughout the course of the research. Without them this would not have been possible.

I would also like to thank my colleagues for their friendship and technical support, and especially for their participation in many of my experiments.

Finally, thanks to my family for the continuing support of my path through life.

BLANK IN ORIGINAL

Author's Declaration

I hereby declare that the research documented in this thesis was carried out by myself in the Design, Engineering and Computing Department at Bournemouth University.

Jenny Longster

The following conference and journal papers were published during the course of this work:

Longster, J., Sahandi, R. & Vine, D.S.G. (1999). Prosody Generation in the Time Domain. In: *Proc. SPECOM '99*, Moscow, 170-173.

Vine, D.S.G., Sahandi, R. & Longster, J. (1999). Recording Concatenative Units for Speech Synthesis using a Reference Pitch Prompt. In: *Proc. SPECOM '99*, Moscow, 174-177.

Vine, D.S.G., Longster, J. & Sahandi, R. (1999). Reference Pitch Prompting: A Recording Method for Concatenative Speech Synthesis. In: *Proc. IEE Colloquium on Interactive Spoken Dialogue Systems for Telephony Applications*, London.

Longster, J., Sahandi, R. & Vine, D.S.G. (1998). Facial Animation to Support Bi-Modal Communication. In: *KT'98 International Conference on Knowledge Transfer through Multimedia & Virtual Reality*, Cairo, 13-14 April 1998, 137-144.

Sahandi, R., Longster, J. & Vine, D.S.G. (1998). Text-to-Speech Animation. *Informatica*, 22, 445-450.

BLANK IN ORIGINAL

Chapter 1. Introduction

1.1 Area of Research and Motivation

1.1.1 Area of Research

Text-to-Speech (TTS) is the art of designing talking machines, whereby arbitrary sentences in a textual format are automatically transformed into the spoken word. Speech technology potentially provides an efficient mode of communication between human and computer, and has a wide range of applications from reading machines for the visually impaired, to hands and eyes free operations of controls in avionics. In recent years, many text-to-speech systems have been developed that are able to provide intelligible, unlimited vocabulary output, however it is still possible to distinguish the resulting synthetic speech from natural speech (Sproat *et al.* 1999, Black 2002). Text-to-speech synthesis is a complex interaction between two very different fields of research, namely Natural Language Processing (NLP) and Digital Signal Processing (DSP). Syrdal *et al.* (1998b) suggest that text-to-speech systems may be improved in terms of naturalness by addressing the three areas of linguistic analysis, prosody modeling, and speech synthesis models. There remain aspects needing attention in all stages of the text-to-speech process but it is the domain of speech synthesis that is the focus of this work. A major issue in speech synthesis research is concerned with maintaining the resulting speech quality at the digital signal processing stage, and it is this challenge that provides the motivation for the thesis.

1.1.2 Statement of Problem

There are three main speech synthesis approaches: articulatory, formant, and concatenative synthesis. Concatenative synthesis is currently the most promising approach, providing intelligible speech output in an efficient manner. Its main drawback is in the use of pre-recorded speech segments stored in an inventory, which makes this approach somewhat inflexible in terms of spectral modifications. When synthesising a new utterance, such modifications are often necessary if segments with suitable prosody cannot be found to exist in the inventory. To increase flexibility, corpus-based approaches to concatenative synthesis store multiple versions of speech segments. These segments are extracted from different phonetic and prosodic contexts

and hence have varying voice qualities and prosody (pitch and durations). During synthesis, the 'best' segment is selected from the corpus, in terms of criteria such as phonetic context, position in syllable, word and phrase, and pitch and duration using a process known as *unit selection.* However, the prosody of these segments may still not be suitable when synthesising arbitrary sentences; it is not possible to store every combination of pitch and duration due to the variability of speech. Whilst the corpus-based approach reduces the distance (in Hz and seconds) of the candidate values of the segments in the corpus to the target values of pitch and duration, a signal processing algorithm may still be necessary to perform small modifications. The Time-Domain Pitch-Synchronous OverLap-Add (TD-PSOLA) algorithm provides an efficient and generally successful solution, although certain modifications introduce distortion in the form of 'buzzyness' into the speech. During the thesis, the perceptible distortion that occurs in speech after the application of the TD-PSOLA algorithm will be termed 'buzzyness'. Unless otherwise stated, the term distortion is defined as subjective perceived distortion, as opposed to objective signal distortion.

In a corpus-based approach to concatenative synthesis, the design of the speech corpus is usually phonetically balanced, not balanced to the additional needs of the signal processing algorithm. This lack of consideration may lead to the introduction of increased perceptible distortion in certain circumstances when signal processing algorithms, such as TD-PSOLA, are used for prosodic modifications. Indeed, research often reports results for TD-PSOLA modified speech as 'success in the main' or as giving 'the best and worst' results.

For corpus-based systems, unit selection algorithms are used to select the best candidate segment from the corpus during synthesis. Current unit selection processes that take into account the cost of signal processing in terms of the distortion that may be introduced, often do so by calculating the distance of the candidate values available in the corpus to the target values of the construct to be synthesised. The distance can be defined as the amount in Hz or seconds that the signal processing algorithm must modify the pitch or duration of the speech segments. This cost is estimated as an absolute distance for all segments, rather than weighting the distance according to the peculiarities of the particular algorithm. This neglect may lead to increased perceived distortion.

The thesis is concerned with maintaining the speech quality at the digital signal processing stage of the speech synthesis process, by minimising the perceptible distortion in the output.

1.1.3 Research Aims & Objectives

The thesis attempts to address the specific problem of the introduction of distortion, perceived as 'buzzyness', when speech is modified in pitch using the TD-PSOLA algorithm. The aim of the research is to develop a framework, which will facilitate the synthesis of speech using TD-PSOLA with reduced distortion. The framework will consist of a speech corpus design, tailored to the needs of the TD-PSOLA algorithm, and a signal processing distortion measure, weighted according to the effect of the algorithm on individual speech segments. The design of the corpus and signal processing measure will be guided by the results of investigative experiments undertaken to determine the effect of TD-PSOLA on speech. This will allow the development of an approach to speech synthesis which best minimises perceived distortion. The implementation of the speech corpus design and signal processing measure will be formally evaluated to determine whether the framework reduces distortion when speech is modified using the TD-PSOLA algorithm. To this end, the work has the following research objectives:

1. Identify extant speech synthesis models: The first objective is to identify and critically evaluate the popular models for the generation of synthetic speech. The remainder of this chapter reviews the entire speech synthesis process to set the work in context, provides the terminology and background required, and ends by identifying the main speech synthesis models in existence to meet this objective. The model considered most potentially successful for the future of speech synthesis is identified and justified as the choice for further investigation during this thesis.
2. Analyse the effect of the TD-PSOLA algorithm: The second objective is to identify some of the potential distortions associated with the algorithm. Chapter 2 begins by presenting the operation of the TD-PSOLA algorithm mathematically and then describes the implementation of the algorithm to be used during the research. Throughout the thesis, the Praat speech software (Boersma & Weenink, 1999) is used to analyse speech samples and provides the implementation of TD-PSOLA under investigation (see Section 2.3 for an introduction to this software). The basic distortions TD-PSOLA introduces into abstract signals, such as pure sine waves and single formant signals, are then investigated.

The distortion referred to here is in the form of objective signal distortion, rather than subjective perceptible distortion. The chapter then attempts to determine whether the objective distortions observed are perceptible, and whether they may be perceptible in more complex signals such as natural speech. To this end, extant research concerning TD-PSOLA is reviewed, identifying parameters that may lead to the occurrence of perceptible distortions.

3. Review speech assessment techniques: Chapter 3 presents a review of current, popular subjective techniques and practices for the assessment of intelligibility and quality of speech. This is used to inform the design and procedure of experiments carried out during this work.
4. Investigate the effect of TD-PSOLA on natural speech: Chapter 4 documents a series of subjective listening experiments undertaken to investigate the effect of the TD-PSOLA algorithm on resulting distortion levels, when used for pitch-modification of natural speech. The results of these experiments are used to suggest parameters that may contribute to perceptible distortion.
5. Develop a framework for producing synthetic speech with less perceived distortion: Chapter 5 documents how the results of the investigative experiments were analysed to inform the development of a novel speech corpus, tailored to the needs of the TD-PSOLA algorithm. The data were also analysed to develop a signal processing distortion measure. The measure is weighted according to the phonetic identity of the individual speech sounds to reflect how each segment responds to the algorithm in terms of perceived distortion levels. Finally, a special-case selection process was developed for highly problematic voiced fricative phonemes.
6. Evaluate the framework: Chapter 6 describes an experiment to determine the success of the speech corpus at producing synthetic speech, with less perceived distortion than a standard approach. It also evaluates the validity of the signal processing distortion measure to justify the need for such a measure in standard unit selection procedures. Finally, the special-case voiced fricative selection process is evaluated in terms of its ability to produce TD-PSOLA-modified versions of these phonemes with less perceived distortion.

The following introductory material in Chapter 1 explains the context for the work and provides the necessary conceptual underpinning. Initially, an overview of human speech production and its physical and phonetic representation is given, followed by an introduction to the overall text-to-speech process. Following this, a detailed and critical examination of some current, more popular techniques used at the speech synthesis stage is presented. The chapter then discusses the selection of concatenative synthesis, using a corpus-based approach in conjunction with the TD-PSOLA algorithm, as a promising direction for speech synthesis and for further investigation during this research. Finally, the chapter reiterates the structure of the remainder of the thesis.

1.2 Overview of Speech Production and its Representation

Speech synthesis is a complex research field. To fully appreciate this complexity and the challenges of speech synthesis research, knowledge of the human speech production process is essential. It is also necessary to understand the physical representation of the resulting speech signals and be conversant with phonetics to be able to describe speech in an abstract symbolic representation.

1.2.1 Speech Production

Phonation or human speech is produced by the vocal organs, which are depicted in Figure 1.1. The respiratory organs of the lungs and the diaphragm produce and force air up the trachea and through the vocal cords (or folds) to the main cavities of the vocal tract: the pharynx, and the oral and nasal cavities.

The opening between the vocal cords is called the glottis. Air flows freely through the glottis during breathing or unvoiced speech such as /s/ or /f/, but during voiced speech, such as /I/ or /E/, the cavity containing the vocal cords (the larynx) is obstructed.

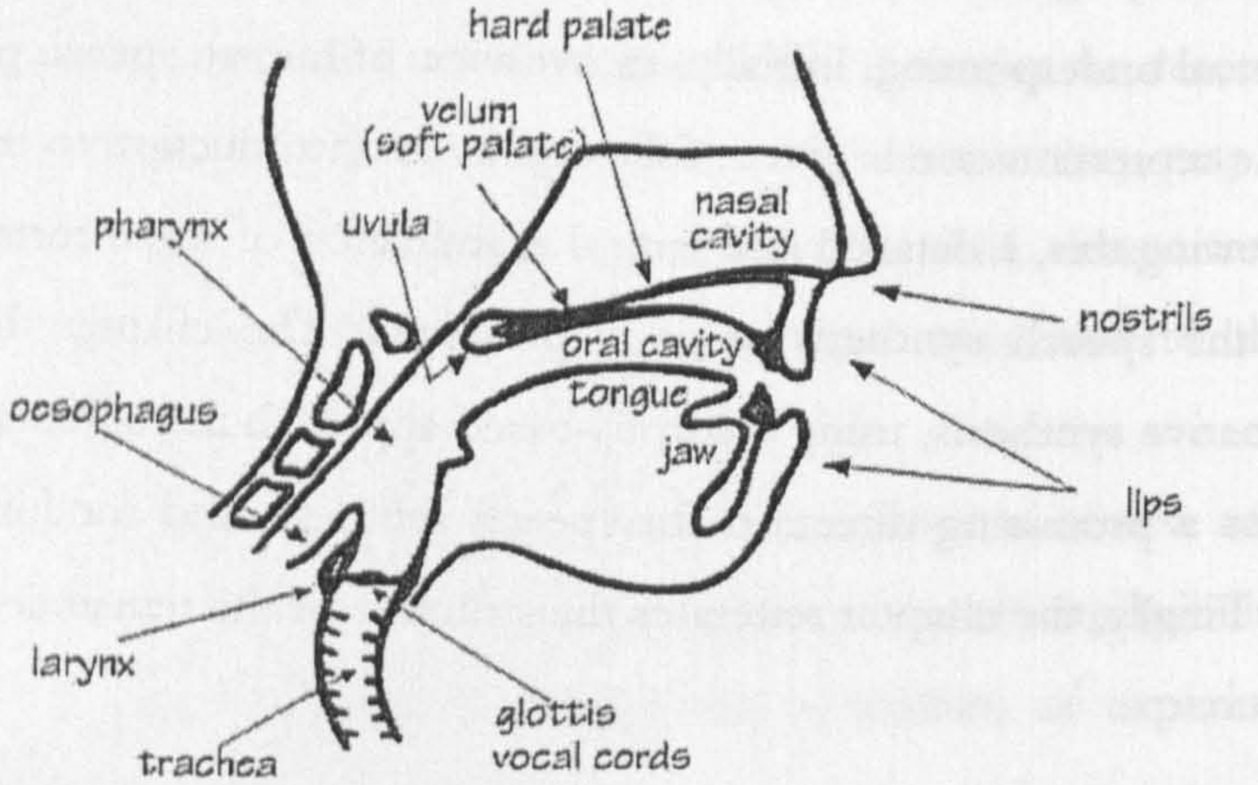


FIGURE 1.1. THE SPEECH PRODUCTION ORGANS (FROM DUTOIT, 1997)

Increasing air pressure forced upwards causes the vocal cords to open to release this air, leading to a pressure drop and the closure of the glottis. In this way, the vocal cords modulate the airflow by rapidly opening and closing, causing a vibrating sound. This is in the form of a *glottal waveform* or a sequence of pulses that are fed into the vocal cavities, from which voiced speech is produced. The frequency of this sound depends on the mass and tension of the vocal cords, and is known as the *fundamental frequency*. Average fundamental frequencies (f0) range from 70 to 200, 150 to 400, and 200 to 600Hz for men, women and children respectively (Dutoit, 1997).

During unvoiced speech, the airflow in the vocal cavities is turbulent due to several constrictions in the vocal tract, which may occur anywhere between the glottis and the mouth. Some speech sound production requires both this turbulent noise and a glottal waveform to be present at the same time, for example during the production of the voiced fricative /v/. Alternatively, the air from the lungs may be stopped totally by the closure of the vocal cords, called a *glottal stop* or by a closure somewhere in the vocal tract, such as the lips.

The pharynx and oral cavity are used for most sounds, although nasal sounds (/m/ or /n/) require the nasal cavity to be shunted with the oral cavity by lowering the velum. The size and shape of the oral cavity are altered by movements of the palate, tongue, cheeks, lips and teeth, which determine the timbre of the sounds produced.

1.2.2 Physical Representation of Speech

The perceptual aspects of speech such as pitch, rhythm, loudness, and timbre have acoustic correlates of fundamental frequency, duration, intensity, and spectral energy distribution that can be represented diagrammatically in the time and frequency domain.

In the time domain, the basic representation of a speech signal is the *waveform*, which depicts the speech signal as a series of pressure changes in air as a function of time. Figure 1.2 shows a time-domain waveform of the word "kit" recorded by a female voice.

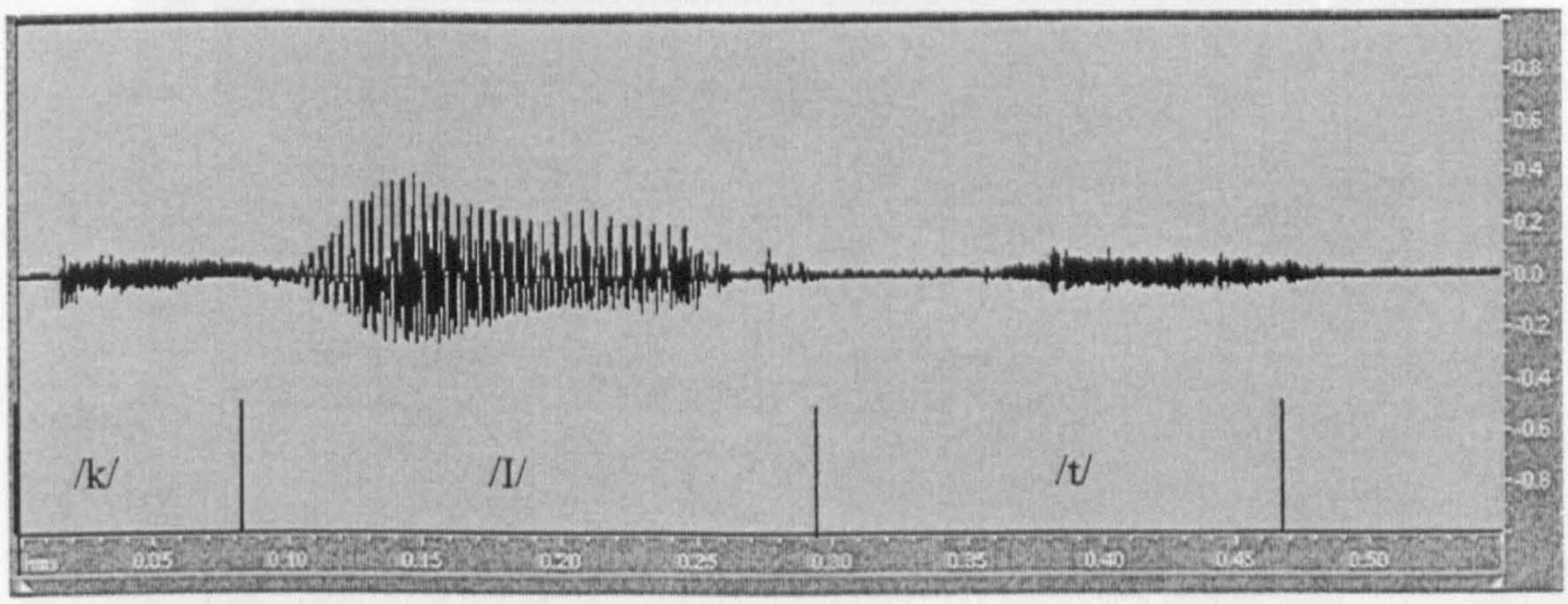


FIGURE 1.2 TIME-DOMAIN WAVEFORM OF THE WORD "KIT"

The acoustics of the speech originate from the production. Voiced speech has a fundamental frequency (f0), produced by the opening and closing of the vocal cords, and the harmonic components of this frequency. Unvoiced speech has no fundamental frequency and no harmonic structure; it may be viewed as white noise caused by air forced through the constricted vocal tract. The waveform in Figure 1.2 shows the periodic, or voiced parts of speech (/I/), and the noisy, or unvoiced parts of speech (/k/ and /t/) in the word "kit". It also shows the loudness or intensity of the speech as the amplitude of the pressure changes from the resting value.

The frequency domain representation of a speech signal, or *spectrum*, can be generated by calculating the Discrete Fourier Transform (DFT) of the time-domain signal. As most speech sounds are bounded in time, or are only *quasiperiodic*, meaning they vary slightly from one period

to the next, the spectrum of a speech signal is usually calculated at one particular point in time. This is achieved by applying an analysis window, such as a Hamming or Hanning window, which makes the small portion of sound of interest fade in and out and renders the rest of the signal zero. The window is usually of 10ms to 30ms duration, over which time the signal is assumed to be stationary.

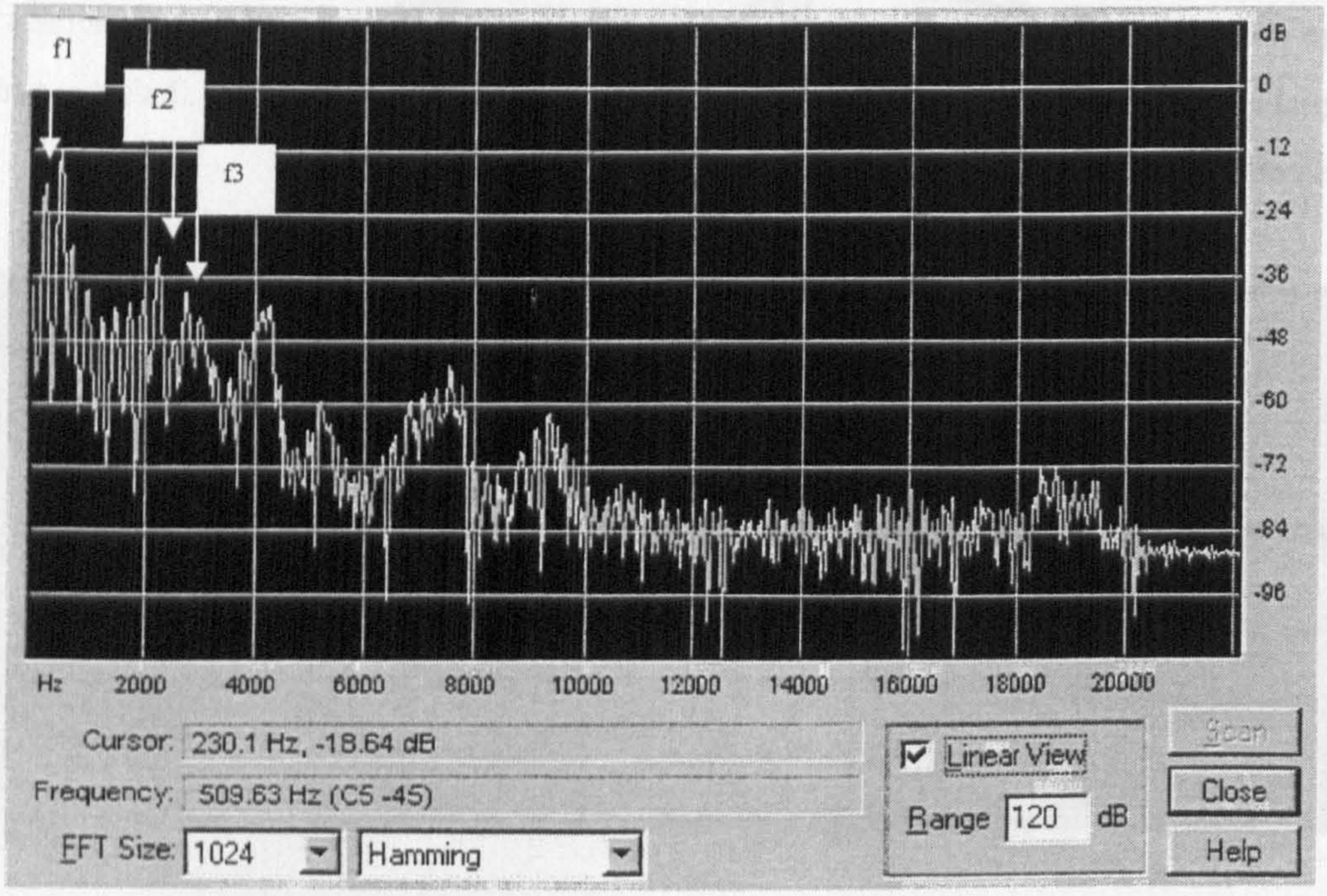


FIGURE 1.3. SPECTRUM OF PHONEME /I/

In Figure 1.3, the spectrum of the phoneme /I/ clearly shows the many different frequency components, and the intensity of these components, that make up a complex signal at a particular instant. The first spike or narrow peak at 240 Hz represents the fundamental frequency, with the other spikes representing the harmonics of this frequency. The timbre of speech depends on the overall spectral shape, called the *spectral envelope,* which appears as a series of broad peaks showing higher energy levels. The peaks and troughs in the spectral envelope are determined by how the vocal tract modifies the excitation signal due to its resonant frequencies, causing formants (poles) and sometimes antiformants (zeros). The first, second and third formants (f1, f2 and f3) are visible, with their bandwidths and amplitudes, at approximately 500Hz, 2400Hz and 3000Hz respectively. These values were confirmed using the Praat software (Boersma & Weenink, 1999)

and found to lie within the typical ranges for the female production of /I/. The formant frequencies depend on where and to what extent the vocal tract is constricted (Flanagan, 1972).

For speech applications, it is not always convenient to view only one particular instant of the signal in the frequency domain. *Spectrograms* (Koenig *et al.*, 1946) provide both a time and frequency domain representation of speech and are composed of a collection of spectra. Frequency is shown on the vertical axis and time on the horizontal, with a third dimension of amplitude represented by shade of grey. The speech sounds may be shown as the temporal evolution of the spectral components and their varying intensity.

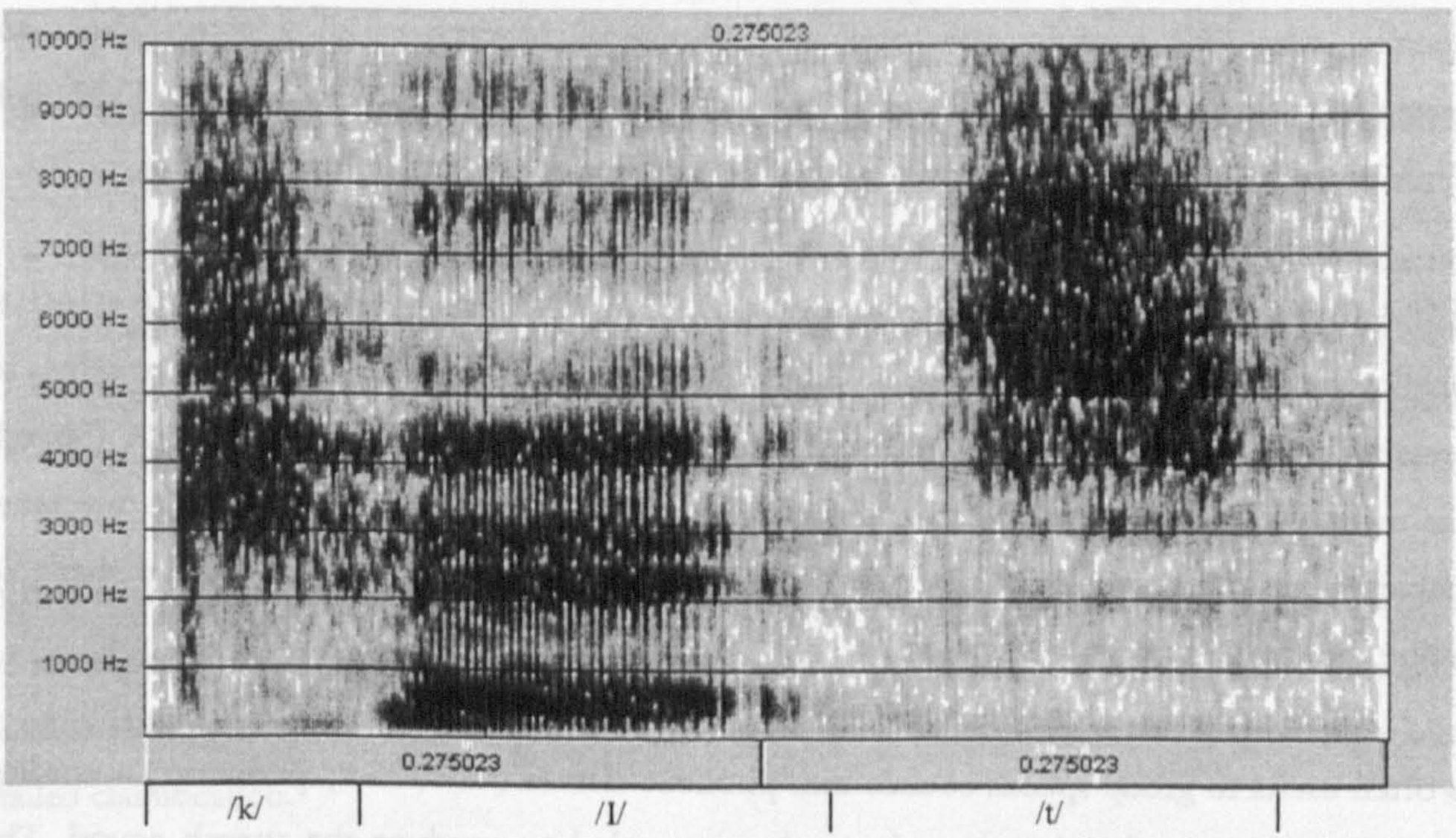


FIGURE 1.4 SPECTROGRAM OF THE WORD "KIT"

Figure 1.4 shows a spectrogram of the word "kit". Voiced sounds can be seen to have more energy focused at lower frequencies with each formant centre frequency, bandwidth and amplitude evolving over time. Unvoiced consonants are more silent, having lower energy levels usually focused at higher frequencies. Unvoiced sounds are also less steady involving rapid changes.

1.2.3 Phonetics

Phonetics is concerned with transcribing written text into the correct pronunciation of the spoken word using a symbolic representation. Each language has its own phonetic alphabet that describes every possible phoneme. A *phoneme* is an abstract unit that may be defined as the smallest contrastive unit in a language (Crystal, 1987) or alternatively as a group of sounds classified as the same by native speakers of that language. An example given by Gelfand (1998) illustrates the phoneme /p/ in "pipe" is recognised as a /p/ in both positions in the word, although in fact they are produced differently (the first is accompanied by a burst of air, the second is not) leading to two *allophones* of that phoneme, or dissimilar members of the same phonemic class. The acoustic realisation of a phoneme is often called a *phone*. There are approximately 40 phonemes in the English language (Breen *et al.* 1996, Donovan 1996) although the number cannot be determined easily due to the complexity and variability of speech.

The IPA (International Phonetic Alphabet) notation (IPA, 1949) has been developed to associate phonetic symbols to sounds using Greek letters, which unfortunately do not lend themselves to processing with computers not in possession of correct character sets. The SAMPA (Speech Assessment Methods Phonetic Alphabet) notation (Wells *et al.*, 1992) provides a machine-readable phonetic transcription. The SAMPA notation will be used throughout the thesis to describe the speech under investigation.

It is often useful to group speech sounds into phonetic classes (articulatory phonetics) according to *manner of articulation* i.e. the type of articulation needed to produce the speech sound. The English language is comprised of two main classes: vowels and consonants.

English vowels can be grouped as either checked (of short duration) or free (of longer duration). The free vowels are made up of monothongs (one vowel sound in a single syllable, such as /i:/ in "ease") and diphthongs (two vowel sounds in a syllable, such as /U@/ in "cures") although it is often difficult to classify them as such. There is also one unstressed vowel /@/, occurring for example at the beginning of the word "another". A review of vowel perception may be found in Kent & Read (1992).

Consonants may be described as fricatives, affricatives, plosives or stops, nasals, and semi-vowels (glides and liquids). Their manner of articulation is described below and examples are given in Table 1.1.

- Fricatives and Affricatives (Hughes & Halle, 1956): during the production of fricatives, the vocal tract is constricted at various places such as the glottis, hard palate, teeth or lips making the airflow turbulent. Affricatives begin as plosives, but when the vocal tract is released, a fricative sound emerges.
- Plosives (Halle *et al.*, 1957): the vocal tract is closed causing a build-up of pressure. When it reopens, a burst of sound is released.
- Nasals (Fujimura, 1962): when the vocal tract is closed and the velum is lowered, air flows out through the nasal cavity.
- Semivowels (O'Connor *et al.*, 1957) consist of two groups: glides and liquids. Production of glides involves a fast transition from a vowel-like open position, producing a frication. Liquids involve vowel-like articulations, which are produced in conjunction with partial closure of the vocal tract with the tongue.

English consonants can be further grouped into obstruents (plosives, affricatives and fricatives) and sonorants (nasals, liquids and glides). The obstruents may be loosely classified as voiced or voiceless although this depends heavily on their context. O'Shaughnessy (1987) provides a more detailed classification.

Table 1.1 describes the speech sounds in the English language, grouped according to manner of articulation. The SAMPA notation, an example word, and the transcription of this word are given. In the final column, consonants are classified as either voiced or unvoiced, and vowels (which are all voiced) are classified as either monothongs or diphthongs.

Manner of Articulation	SAMPA Symbol	Example Word	Transcription	Additional Information
Plosives	p	pin	pIn	unvoiced
	b	bin	bIn	voiced
	t	tin	tIn	unvoiced
	d	din	dIn	voiced
	k	kin	kIn	unvoiced
	g	give	gIv	voiced
Affricatives	tS	chin	tSIn	unvoiced
	dZ	gin	dZIn	voiced
Fricatives	f	fin	fIn	unvoiced
	v	vim	vIm	voiced
	T	thin	TIn	unvoiced
	D	this	DIs	voiced
	s	sin	sIn	unvoiced
	z	zing	zIN	voiced
	S	shin	SIn	unvoiced
	Z	measure	meZ@	voiced
	h	hit	hIt	unvoiced
Nasals	m	mock	mQk	voiced
	n	knock	nQk	voiced
	N	thing	TIN	voiced
Liquids	r	wrong	rQN	voiced
	l	long	lQN	voiced
Glides	w	wasp	wQsp	voiced
	j	yacht	jQt	voiced
Checked vowels	I	pit	pIt	monothong

	E	pet	pEt	monothong
	{	pat	p{t	monothong
	Q	pot	pQt	monothong
	V	cut	kVt	monothong
	U	put	pUt	monothong
Unstressed	@	another	@nVD@	monothong
Free vowels	i:	ease	i:z	monothong
	eI	raise	reIz	monothong
	aI	rise	raIz	diphthong
	OI	noise	nOIz	diphthong
	u:	lose	lu:z	monothong
	@U	nose	n@Uz	monothong
	aU	rouse	raUz	diphthong
	3:	furs	f3:z	monothong
	A:	stars	stA:z	monothong
	O:	cause	kO:z	monothong
	I@	fears	fI@z	diphthong
	e@	stairs	ste@z	diphthong
	U@	cures	kU@z	diphthong

Table 1.1. Phonemes of the English Language

Speech sounds may also be described by *place of articulation* i.e. the location of primary constriction needed to produce the speech sound, depending on whether the sounds are produced at the front or back, with an open or closed mouth etc. The more common places of articulation are listed here:

- Labial: lips e.g. /p, b, m/
- Dental: teeth e.g. /T/ as in "thin"
- Labio-dental: lower lip and upper teeth e.g. /f, v/
- Alveolar: blade/ tip of tongue with dental ridge e.g. /t/

- Palatal: tongue and roof of mouth e.g. /dZ/ as in "jam"
- Palato-alveolar: as alveolar but tongue tip is lowered e.g. /S/ as in "shoe"
- Velar: back of tongue and roof e.g. /k, g/
- Glottal: momentary closure of throat e.g. "go' a lo' o' bo'lle"

Vowels may be described from front to back in terms of tongue elevation and lip rounding. Front vowels (/i:, I, eI, E, {/) are produced with the lips retracted while back vowels (/u:, U, @U, O:, Q, A:/) require rounded lips. Middle vowels (/V, @, 3:/) are produced when the tongue elevation is in the vicinity of the hard palate. Vowels may also be described as tense or lax depending on the degree of muscle contraction e.g. /i:/ ("peat") is tense and /I / ("pit") is lax.

1.3 The Text-to-Speech Process

A Text-to-Speech (TTS) system automatically converts textual input into audible speech. Figure 1.5 shows a general TTS system, consisting of a Natural Language Processing (NLP) module and a Digital Signal Processing (DSP) module.

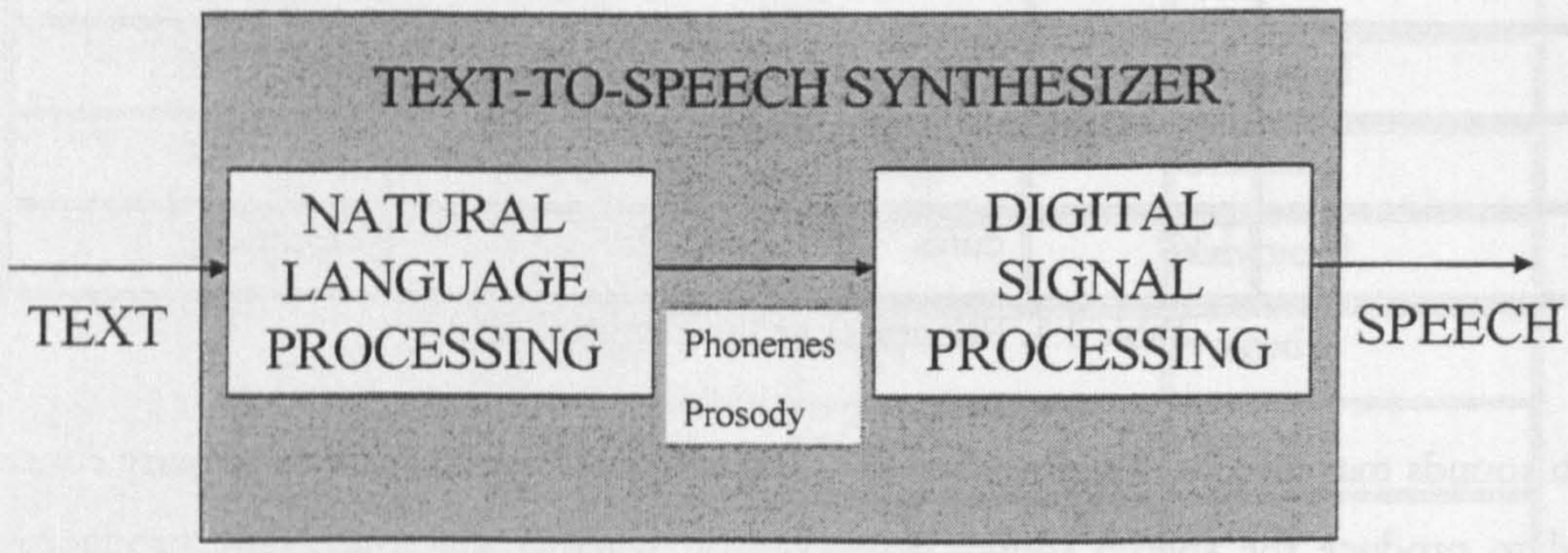


FIGURE 1.5 GENERAL TTS SYSTEM

The NLP module takes the textual input and produces a phonetic transcription of the sounds that are to be produced. It also predicts prosodic information from the text, describing how the

sounds are to be produced in terms of rhythm and intonation. The DSP module synthesises the required speech by transforming this symbolic information into a physical waveform.

In the following sections, a brief description of each process that occurs in these two modules is given. This illustrates the complexity of text-to-speech and sets the topic of this research of *speech synthesis* at the DSP stage in context.

1.3.1 The Natural Language Processing Module

Figure 1.6 shows a more detailed diagram of a general Natural Language Processing (NLP) module. The NLP stage is extremely difficult, as mere text does not contain all of the information needed to produce speech.

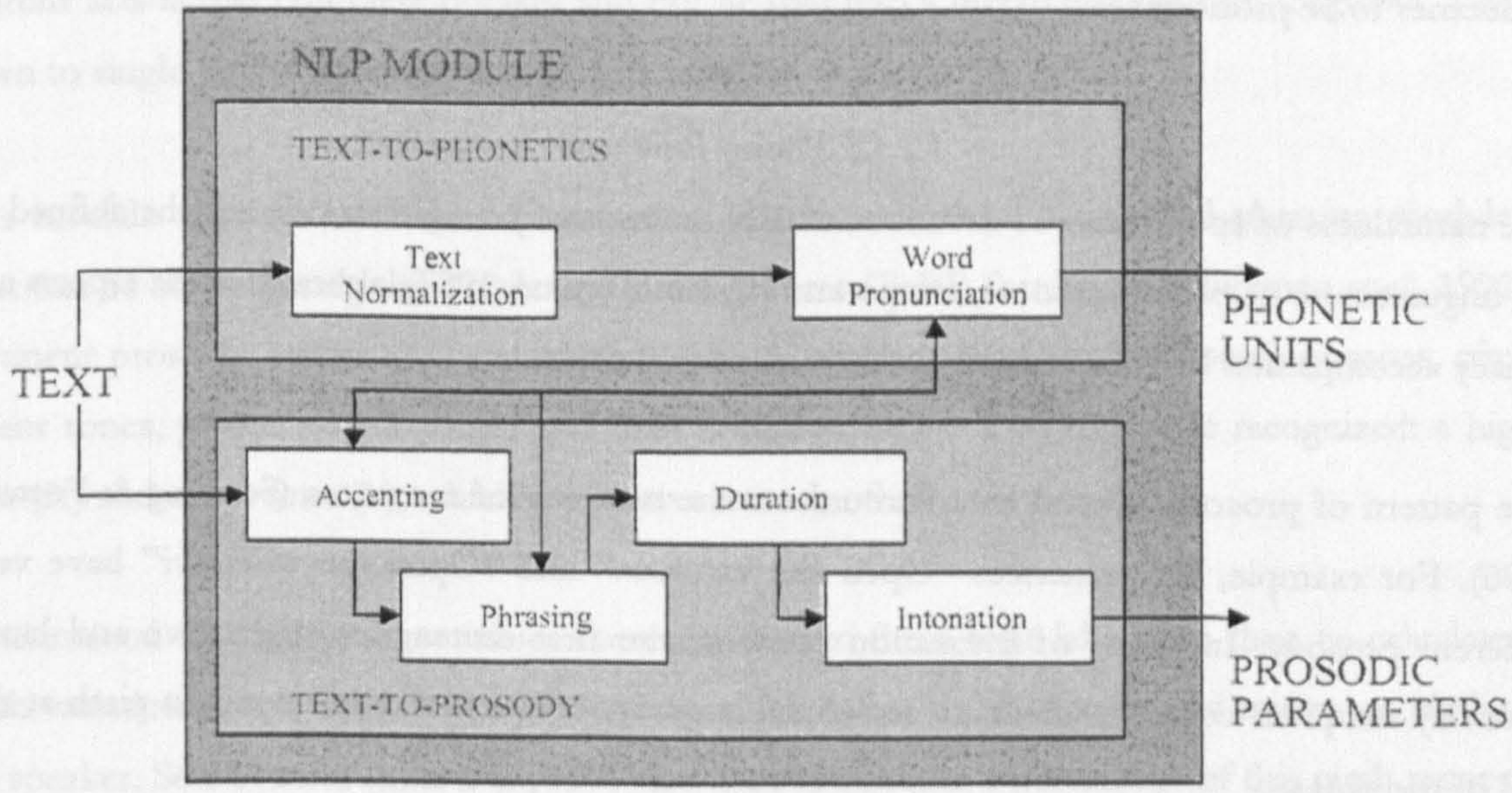


FIGURE 1.6. NATURAL LANGUAGE PROCESSING MODULE

The first block of the NLP module converts the text to phonetic information and may be further broken down into a *text normalization* process and a w*ord pronunciation* process. The second block produces prosody information from the text and from the output of the word pronunciation process. It is broken down into smaller processes that determine *accenting*, *phrasing*, *duration*, and *intonation*. For a review, see Edgington *et al.* (1996a).

1.3.1.1 Text Normalization

A text normalization module, or text preprocessor, allows any ambiguous text, such as numbers, dates, abbreviations, acronyms and idiomatics, in any format to be resolved. It parses the text into sentences and organizes these into lists of smaller units such as words.

1.3.1.2 Word Pronunciation

Once the sequence of words has been generated, their pronunciation can be determined. Where words are pronounced as they are written, a simple set of letter-to-sound rules may be applied. Where this is not the case, a morpho-syntactic analyser may be used to tag the speech with various identities, such as prefixes, roots and suffixes, and organizes the sentences into syntactically related groups of words, such as nouns, verbs, and adjectives. The pronunciation of these can then be determined using a lexicon. Finally, a phonetizer provides the sequence of phonemes to be pronounced.

1.3.1.3 Prosody Prediction

The naturalness of speech can be described mainly in terms of *prosody*. Prosody may be defined as the linguistic use of pitch, loudness, tempo and rhythm (Crystal, 1987), although these aspects are usually accompanied by variations in phonation and voice quality.

The pattern of prosody is used to communicate the meaning of sentences (Sonntag & Portele, 1996). For example, the sentences "Open the window." and "Open the window?" have very different prosody. In terms of intonation contour, the first sentence is declarative and has a relatively flat pitch contour, whereas the second is questioning and exhibits a rise in pitch at the end of the phrase.

The naturalness of a TTS system is a function of prosody (Dutoit, 1997). Prosody prediction is performed by the Text-to-Prosody module, which determines the accenting, phrasing, intonation and duration for each sentence. Intensity variations are perceptually the least important aspect of prosody (Howell, 1993), and are often ignored.

Accenting. Accent or stress assignment is based on the category of the word e.g. context words (nouns, adjectives and verbs) are typically accented and function words (prepositions and

auxiliary verbs) are usually not. This information is used to help predict the intonation and duration.

Phrasing: Sentences are broken down into phrasal units and phrase boundaries are assigned to the text. These boundaries indicate pauses and the resetting of intonation contours.

Intonation: Intonation clarifies sentence type and hence its meaning such as questioning, declarative etc. In addition, pitch variations convey information about stress, emphasis, gender and emotion.

The intonation module generates a pitch contour for the sentences. Pitch contours may be stylized; not all variations of pitch seen in natural speech are perceptible ('t Hart *et al.*, 1990). The contour acts across complete phrases and can be split into a hierarchical pattern (Sagisaka, 1990) down to single pitch targets associated with syllables or parts of syllables.

This module requires information from the phonetic, accent, duration and phrasing modules. Text can be labeled using the Tones and Break Indices (ToBI) formalism (Silverman *et al.*, 1992). Pertinent prosodic events are marked with one of four tone labels: initial boundary tones, pitch accent tones, phrase accent tones, and final boundary tones. Two tones are recognized: a high tone (H) and a low tone (L), which are relative to each other.

Fundamental frequency target values corresponding to these tone labels can then be calculated. Pitch variations occur between *declination lines*, which define the maximum and minimum pitch of the speaker. Sets of rules (Jilka *et al.*, 1999) are used to calculate a percentage of this pitch range to give physical f0 values, which are then assigned to the voiced parts of speech. Transitions between target values are specified as either linear interpolations or more complex transitions such as exponentials, to provide a stylized pitch contour. The pitch movements can be characterized by direction (rise and fall), rate of change (slow or fast) and size (half or full) and timing (early, late, very late in the syllable).

Duration: Segmental duration is an essential aspect of prosody (Carlson *et al.*, 1979) that affects the overall rhythm of the speech, stress and emphasis, the syntactic structure of the sentence, and the

speaking rate (Klatt, 1979). Many factors contribute to the duration of a speech segment, such as the identity of the phone itself, the identity and characteristics of neighbouring phones, the accent status of the syllable containing the phone, its phrase position and the speaking rate and dialect of the speaker. Duration prediction is usually achieved using a rule-based model, which takes these factors into account (Klatt 1979, Bartkova & Sorin 1987). For a complete review of prosody prediction, see Edgington *et al.* (1996b).

1.3.2 Digital Signal Processing Module

Once the phoneme list has been generated from the text-to-phonetics stage, and the prosody has been predicted in terms of duration and frequency values, the physical speech may be synthesised. The required speech sounds are extracted from an inventory and joined together. The extracted speech segments may already have the desired prosody or it may be imposed on the segments using signal processing techniques to fit the new utterance.

The following section describes the existing approaches to speech synthesis, which forms the broad area of this research. The advantages and disadvantages of each approach are discussed to determine the currently most promising synthesis strategy for further investigation.

1.4 Speech Synthesis Strategies

There are currently three main approaches to synthetic speech production: articulatory synthesis, synthesis by rule (also known as formant synthesis), and concatenative synthesis. Before discussing each of these, the *source-filter* model of speech is introduced, upon which articulatory synthesis, formant synthesis, and Linear Predictive (LP) synthesis (a form of concatenative synthesis) are based.

1.4.1 Source-Filter Model of Speech

The source-filter theory of speech production (Fant 1960, Velhuis 1998) is based on the assumption that human speech can be modeled as an excitation source and a vocal tract response

that are independent of each other. During synthesis applications, the excitation signal is modeled by two sound sources; one to model the vibration of the vocal folds that occurs during voiced speech, and one to model the turbulent noise caused by air pushed through the vocal tract during unvoiced speech. These consist of a quasiperiodic train of pulses and a noise signal respectively.

A filter models the frequency response of the vocal tract and the radiation characteristics of the lips and nostrils. The resonance characteristics of the vocal tract are caused by many factors but the most important are the length of the vocal tract and the cross-sectional area profile.

1.4.2 Articulatory Synthesis

Articulatory synthesis (Kröger 1992, Rahim *et al.* 1993) models the movement of the speech organs themselves based upon the source-filter model of speech described above. Articulatory control parameters may be lip aperture, lip protrusion, tongue-tip height and position, tongue height and position, and velic aperture. Excitation parameters may be glottal aperture, cord tension and lung pressure. Articulators are modeled as a set of mathematical functions between glottis and mouth for each phonetic segment.

Whilst modeling the speech organs provides intelligible synthesis (Klatt, 1987), its main drawback is the difficulty in determining the control parameters. The parameter data are historically derived from X-ray analysis of the production of natural speech. Unfortunately this does not provide sufficient data for the complex articulatory movements. The second drawback is that it is computationally expensive (Kröger, 1992) and hence remains essentially a research tool rather than finding applications in commercial speech synthesis systems. As analysis methods progress and computational power increases, articulatory synthesis may eventually provide the way forward to more natural synthetic speech.

1.4.3 Synthesis by Rule

Synthesis by rule, or *formant synthesis* (Holmes 1983, Allen *et al.* 1987), models the speech signal itself based on the source-filter model of speech.

Speech is synthesised using a data table of up to 60 continually varying acoustic parameters for each speech sound (Stevens, 1990). Examples of such parameters are voicing f0, degree of voicing in excitation, formant frequencies, antiformant frequencies, bandwidths, and amplitudes etc. It is difficult to determine these parameters and the rules governing their dynamic evolution, which are found by laborious analysis of natural speech. At synthesis time, these rules are matched to the phonetic input and a parametric speech signal is generated, which is fed into a bank of filters, representing each formant frequency.

A fundamental frequency control determines the frequency of the pulses generated and a mixer controls the amount of voiced/unvoiced excitation signal. An amplitude control is used to vary the loudness at the input to the filters.

Formant synthesis is infinitely flexible in terms of prosody generation and speaker independence. Formant synthesisers e.g. JSRU (Holmes *et al.*, 1964), Klattalk (Klatt, 1982) (the predecessor to the Digital Equipment Corporation's DECtalk) and MITALK (Allen *et al.*, 1987), provide intelligible speech, although the resulting speech has an inherent buzzy sound that makes it sound synthetic (Edgington *et al.*, 1996b). A more detailed description of rule-based synthesisers can be found in Holmes (1983) and Allen *et al.* (1987).

1.4.4 Concatenative Synthesis

Concatenative synthesis has been in existence since the late 1970's and is capable of producing highly intelligible speech (Dutoit & Leich, 1994). It is the synthesis strategy chosen for this research because it gives rise to significant advances in terms of simplicity and lack of inherent buzzyness when compared to articulatory and formant synthesis respectively. This is mainly because concatenative synthesis makes use of pre-recorded segments of speech and hence does not model either the way humans generate speech or the speech signal itself. A typical concatenative synthesis system is shown in Figure 1.7.

Concatenative synthesis takes pre-recorded segments of natural speech and joins them together, or *concatenates* them, to produce new utterances. A concatenative speech synthesis system uses small speech segments extracted from natural speech, which are stored in either a parametric

form, as waveforms in an inventory, or as continuous speech in a corpus. To provide the required synthetic output, the appropriate segments are selected from the inventory at run-time and concatenated.

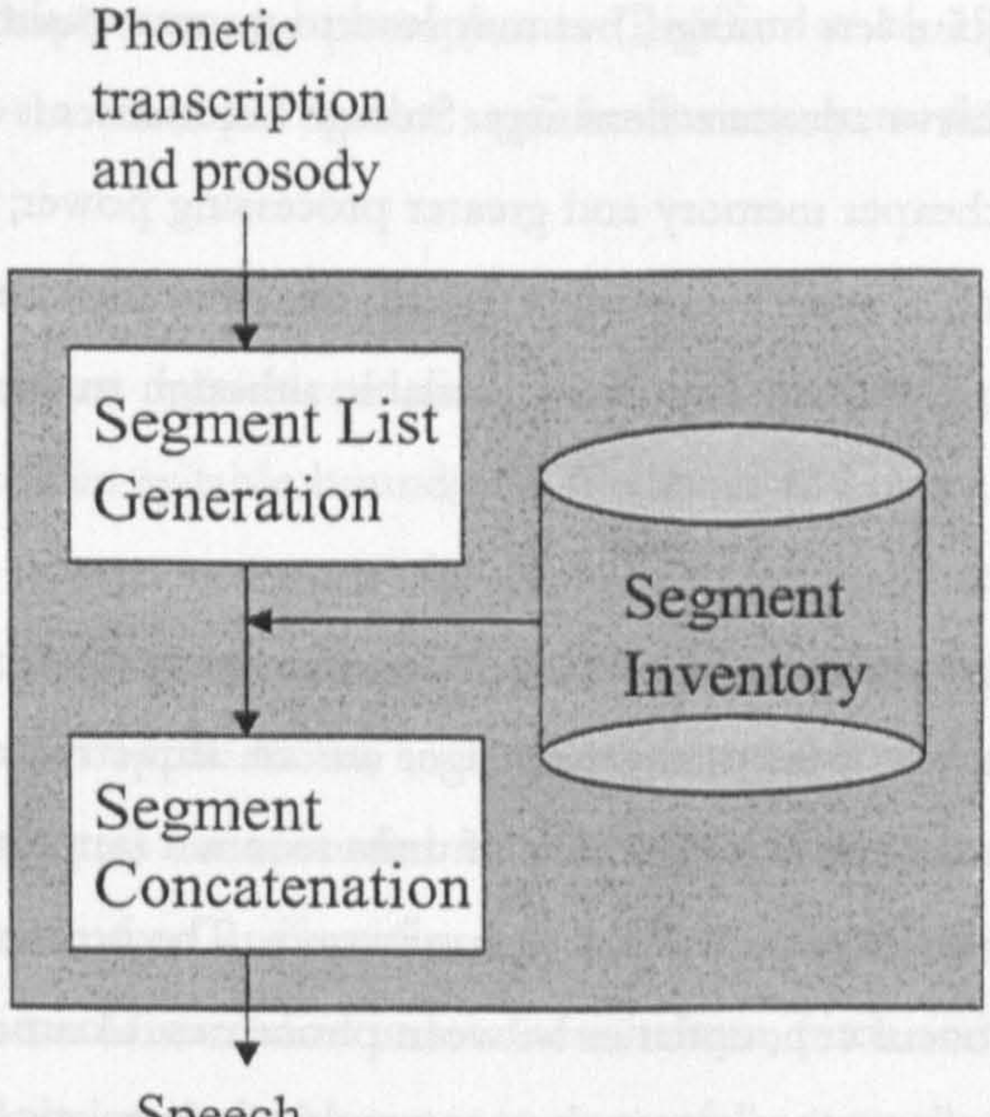


FIGURE 1.7 A TYPICAL CONCATENATIVE SYNTHESIS SYSTEM

The major drawback of concatenative synthesis is its limited flexibility due to the use of pre-recorded speech, which may not allow necessary prosody variations when synthesising novel constructs. The synthetic voice is also restricted to the voice of the speaker used for the recording. Concatenation at segment boundaries may be a problem when segments are extracted from different contexts due to spectral differences, and prosody modification is more difficult. Such issues can be addressed by choice and size of speech segments to be stored in the inventory, and the careful design and creation of the inventory. These issues and possible solutions are discussed in the following sections.

1.4.4.1 Choice of Speech Segment

The type of speech segment stored in the segment inventory has great bearing on the flexibility and quality of the resulting synthetic speech. Such decisions often involve trade-offs between storage requirements, performance, and the extent to which signal processing is required. For example, fewer segments require less storage, but may lead to poorer quality speech and require more signal processing to achieve adequate flexibility. Storage requirements are becoming less of an issue with the advent of cheaper memory and greater processing power; however the time to record, segment and annotate a speech inventory is still an issue. Semi-automatic annotation systems (Wightman & Talkin, 1996) are becoming available although manual correction is still a necessity.

Various types of speech segment have been used in concatenative speech synthesis systems. *Phonemes*, the smallest contrasting unit in the language, are an attractive choice (Witten 1982, Chappell & Hansen 1997) due to the small number of units required (approximately forty for the English language), keeping storage requirements to a minimum. The greatest disadvantage is the coarticulation problems that occur at boundaries between phonemes. Coarticulation describes the way in which humans produce *continuous* speech as a result of articulating a series of isolated words. The articulatory movements are adjusted for different contexts to minimise the effort needed to produce the speech. Coarticulation occurs as each articulator moves continuously from the production of one phoneme to the next and appears in even the most careful speech. Complex rules are needed to deal with this (Linggard, 1985). An additional problem of the use of phonemes is that all joins occur at the least stable part of the waveform where one phone changes to the next, which may cause audible discontinuities.

The use of the *diphone* provides a solution (Lenzo & Black, 2000). A diphone consists of the transition from the centre of one phoneme to the centre of the following one (Dixon & Maxey, 1968). In this structure, the transitional information between phonemes is captured. A set of diphones for the English language numbers approximately 1600, since there are 40^2 possible combinations of phoneme pairs. Simple diphone speech synthesis requires slightly more storage than phoneme synthesis and still has the disadvantage of a high density of concatenation points (one per phoneme). A large number of concatenation points produces the perception of unnaturalness (Donovan & Woodland, 1999) as spectral discontinuities may occur when

segments are selected from different contexts. This places heavy reliance upon smoothing algorithms, which may also degrade quality (Chappell & Hansen 1998, Wouters & Macon 2000).

A variation of the basic diphone system makes use of polyphones such as triphones, half syllables, or even quadraphones and pentaphones (Boëffard *et al.*, 1993). Multiphone constructs of varying length may be included in the segment inventory to deal with highly coarticulated speech.

Larger constructs such as *syllables* or *words* deal with problems of coarticulation as most coarticulation occurs within syllable boundaries (Fujimura & Lovins, 1978). The disadvantage is the large amount of storage necessary for such inventories; there are approximately 10,000 syllables for the English language. These approaches are usually valid for limited vocabulary systems, or closed domain applications, such as train timetable systems, talking clocks etc. Here the necessary word variations are recorded and the required word may be slotted into a standard sentence. This approach often suffers from a lack of coarticulation at the word boundaries, resulting in unnaturalness. Mismatches in loudness, tempo, pitch and voice quality may also lead to disfluent speech if the recordings have not been carefully controlled.

More recent research has extended the simple diphone approach to n-diphone synthesis (Klabbers & Veldhuis, 2001). In n-diphone synthesis, more than one example of each diphone may be stored. This takes into account various coarticulation effects that can occur over a syllable or further over several syllables. The 'best-fit' diphone is selected at synthesis time to minimise spectral differences between adjoining segments. Klabbers & Veldhuis (2001) extend the diphone inventory with additional context-sensitive diphones to reduce the occurrence of audible discontinuities.

The latest developments to improve the flexibility of concatenative synthesis involve corpus-based speech synthesis systems. During natural speech, prosody varies for speech sounds in terms of f0 and duration and also in terms of voice quality. Acoustics of speech sounds vary due to their position in the phrase and the context of the utterance. If all segments are uttered in a neutral manner as in typical diphone systems, these variations are lost. Corpus-based synthesis overcomes these limitations by storing hours of continuous natural speech. This approach also

minimises the large number of concatenations; segments of longer length are selected if they exist together in the corpus. Large prosody manipulations are also reduced by storing and selecting segments having acoustic characteristics that are closer to the target values. The corpus-based approach has been chosen for further research due to its ability to increase the flexibility of concatenative speech synthesis. Corpus-based synthesis introduces its own problems and is discussed further in Section 1.5.5.

1.4.4.2. Creating a Speech Inventory

Creating a large speech inventory is a long process. It involves choosing the phone set, designing the carrier material, generating prompts, recording, segmentation, labeling, pruning and quality control. Designing the optimal speech inventory is one of the most important research issues (Möbius, 2000); it has a huge impact on quality.

The most important criterion is that the inventory must provide adequate phonetic coverage; all segments should be represented. Phonetically rich inventories, used for diphone systems, contain every possible diphone transition, where even the rarest combinations are represented.

Inventories for corpus-based synthesis are often designed to contain sets of phonetically balanced sentences e.g. CHATR (Black & Campbell, 1995), where the phones appear in the same distribution as they appear in normal language. This may be achieved by recording radio news sentences (Black & Campbell, 1995) or a short story of the speaker's choice (Campbell, 1999) for example. This is perhaps contrary to the opinion that it is important to design a database including all relevant realizations of phonemes.

Corpora can be designed specifically for limited-domain applications, such as a speaking clock (Black & Lenzo, 2000b), or it may be enough to record only the required words for applications using a 'slot filler' approach, where the relevant word is inserted into a standard sentence.

Speech segment inventories require carrier material for the segments to be recorded. The choice of carrier material should reflect the application of the synthesiser. It may be important to keep a fixed speaking rate, durations, and perhaps monotone speech for a simple diphone inventory. To achieve this, segments can be extracted from nonsense syllabic sequences (logatoms), or isolated

words. The use of logatoms ensures coverage although there may be a loss of naturalness due to the abstract nature of the material and boredom of the speaker. The recording of segments only in word stress positions in carrier sentences, not providing any reduced segments, may lead to over-articulated speech.

For greater prosodic coverage, natural sentences, of possibly longer passages are used. The use of such text means the speaker is more relaxed; overall quality may be less consistent failing to produce exactly the desired speech segments. It also gives rise to greater variability of the speech, which may cause greater spectral discontinuities between joining segments during synthesis.

The recording of the segments should be performed in a quiet room, preferably an anechoic chamber. Audio settings for the recording process must be fixed, and the speaker should be at a fixed distance from microphone, which can be achieved using a head-worn microphone. Problems occur due to the time gaps between successive recordings; the emotional state and health of the speaker may vary, leading to changes in speaking style and voice quality. Inter-session variations arise due to the increasing boredom of the speaker over a long session.

Once recorded, speech units are segmented from the carrier material. This is performed either manually, or semi-automatically (usually involving manual corrections). The units are stored as segments extracted from logatoms or natural speech, or as part of a speech corpus from which they are extracted at runtime. The segments are tagged with the information required for segment selection, prosody modifications and synthesis. Such labels may be the segment identity, duration, pitch, internal phoneme boundaries, position in phrase etc. The segments may be coded or given a parametric form as a temporal sequence of vectors of parameters. Coding reduces memory loads and may be required for speech models that use a parametric form to allow concatenation and prosody matching. The speech model used in the synthesis system obviously determines the storage and tagging formats.

1.4.5 Summary

The three main existing approaches to synthesis have been presented. Articulatory synthesis provides high quality speech but is extremely complex. It may well be the preferred approach in

the future as processing power increases, but is essentially a research tool at the present time. Formant synthesis has the advantage of great flexibility, but suffers from an inherent 'buzzyness' that cannot be avoided due to its parametric nature. The concatenative approach currently provides better quality synthetic speech than formant synthesis in terms of buzzyness, although it does not have the flexibility of parametric speech models. Additionally, concatenative synthesis does not suffer from the complexity of articulatory synthesis, making it an attractive choice for many successful commercial systems such as BT's Laureate (Page & Breen, 1996) and the AT&T Next-Gen system (Beutnagel *et al.*, 1999a).

The aim of this thesis is to design a framework that provides high quality speech for open domain applications. To this end, it should provide adequate phonetic coverage and also be capable of producing a prosodically rich output. As a result of its high quality output and efficiency, concatenative synthesis was chosen for further investigation. As previously stated, concatenative synthesis does not possess the flexibility of parametric speech models, in terms of the ability to synthesise various types of phonation and with a wide range of prosody. With the advent of cheaper memory and processing power, increasing the size of the inventory from a simple diphone approach can reduce this inflexibility. The development of the corpus-based approach retains details such as variations in phonation and also provides greater ability to achieve the desired prosody when creating novel utterances by storing multiple versions of segments. However, even the largest inventory cannot contain every possible combination of segments in every prosodic context due to the high variability of speech. In order to increase flexibility further and ensure a robust output, whereby all prosodic targets can be met, it may be advantageous to employ a speech synthesis model to concatenate the segments and allow prosodic modifications. The following section describes and analyses the current, popular concatenative synthesis speech models used to concatenate segments and facilitate prosody modifications.

1.5 Speech Models for Concatenative Synthesis

Concatenative synthesis speech models must be able to concatenate a sequence of segments, and adjust them for new prosody when creating arbitrary sentences. The model must maintain high

speech quality with minimal introduction of artifacts or reduction in naturalness. There are many models in use and the more popular ones are discussed in turn in the following sections, leading to a discussion on the choice of the model chosen for this research that provides the most promising approach to concatenative speech synthesis.

1.5.1 Linear Prediction

The Linear Prediction (LP) model (Markel & Gray, 1976) was originally designed for speech coding, providing accurate estimates of speech parameters, but can be used successfully for both coding and synthesis (Sproat & Olive, 1995). It is based on the source-filter model of speech; human speech is modeled as the response of a time-varying digital filter to a periodic or random excitation signal.

For Linear Prediction coding purposes, the natural speech signal is separated into the response of the vocal tract and the excitation signal. The response of the vocal tract, in terms of its formant frequencies, is removed from the signal to be stored as digital filter coefficients i.e. the digital filter coefficients are estimated automatically from frames of natural speech. LP theory assumes the current speech sample $y(n)$ can be approximated or predicted from a linear combination of p previous samples $y(n-1)$ to $y(n-k)$ with an error term $e(n)$, called the residual signal.

$$y(n) = e(n) + \sum_{k-1}^{p} a(k)y(n-k) \qquad \text{Eqn 1.1}$$

$$\text{and } e(n) = y(n) - \sum a(k)y(n-k) = y(n) - \tilde{y}(n) \qquad \text{Eqn 1.2}$$

where $\tilde{y}(n)$ is a predicted value, p is the linear predictor order, and $a(k)$ are the linear coefficients. The coefficients are found using an adaptive algorithm such as the Least Mean Square (LMS) algorithm which minimises the mean-square error between the predicted signal and the actual signal. Autocorrelation or covariance methods are often used for this (Markel & Gray, 1976).

These effects are then removed from the speech signal and if the predictor coefficients are accurate, only the pure excitation signal remains (a harmonic structure and/or white noise), the

intensity and frequency of which can be calculated. This process is known as *inverse filtering* and the remaining excitation signal is called the *residue*. The values of the formants and the residue are stored in an inventory.

LP synthesis reverses this process using a synthetic excitation or the residue to create the source signal and the formant values to create the filter. During synthesis, the speech information stored in frames (usually representing 25 ms of speech and characterized by 10 or 12 LP parameters) is fed to the synthesiser every 25ms. The frame parameters are used to update the digital filter coefficients and select the excitation source and amplitude.

The excitation, which is filtered with the digital filter having the coefficients $a(k)$, may have a different fundamental frequency, therefore providing a new harmonic structure. The filter requires an order (number of coefficients) of 10 to 12 at 8kHz sampling rate and 20 to 24 at 22kHz sampling rate. To obtain intelligible speech and smooth spectral transitions, the coefficients are updated every 5-10ms by interpolating between the previous and current frame parameters.

The main drawback of LP synthesis is that it is inherently buzzy due to its parametric nature, and this degrades speech quality (Klatt, 1987). LP is an all-pole model; phonemes such as nasals and nasal vowels that contain antiformants are not modeled sufficiently thereby decreasing intelligibility. A more detailed explanation may be found in Markel & Gray (1976).

Variations of the basic LP model have been developed to improve the quality although these are computationally more expensive. The excitation signal may be more complex and the source and filter may not be treated as separate. Multipulse LPC (MLPC) (Moulines & Charpentier, 1988) uses a complex excitation constructed from a set of several pulses. Residual Excited LP (RELP) (used in Lernout & Hauspie's commercial TTS system) uses the error signal as an excitation signal. Code Excited LP (CELP) (Campos & Gouvea, 1996) uses a number of excitations that are stored in a code-book.

1.5.2 Sinusoidal Models

Sinusoidal models are based on the assumption that all signals can be composed of a sum of sine waves with various phases, amplitudes and frequencies (McAulay & Quatieri, 1986). This is expressed in Equation 1.3.

$$s(n) = \sum_{l=1}^{L} A_l \cos(w_l n + \phi_l) \qquad \text{Eqn 1.3}$$

where L is the total number of sinusoids, $A_l(n)$ and $\phi_l(n)$ are the amplitudes and phase of each component with frequencies w_l. Parameters $A_l(n)$ and $\phi_l(n)$ are found by taking the Discrete Fourier Transform (DFT) of the windowed signal. Frequencies w_l are estimated by peak picking of the DFT magnitude.

Synthesis reverses this process. Duration modifications are achieved by modifying the parameters corresponding to the vocal tract, so they evolve faster or slower, and the excitation can be stretched or compressed whilst maintaining the same pitch. Pitch modification is achieved by scaling the frequency of the excitation function. The vocal tract is unmodified, which may not model actual human speech production where the vocal tract characteristics alter during higher or lower pitch speech.

Sinusoidal models (Macon & Clements 1996, Crespo *et al.* 1996) perform well for periodic signals; they are particularly adept at synthesising singing speech, which is characterized by elongated vowels. They do not perform so well for unvoiced speech. Sinusoidal approaches make use of glottal closure instants, which does not always provide successful concatenation and may lead to poor quality due to phase mismatch at segment boundaries. Similar models have been developed, such as the hybrid Harmonic plus Noise model (Laroche *et al.*, 1993), which propose a different noise model for unvoiced speech whilst maintaining the harmonic model for speech. These models are discussed in the following section.

1.5.3 Harmonic plus Noise Models

The Harmonic plus Noise Model (HNM) (Stylianou, 1998) assumes that speech is composed of deterministic and stochastic components. The harmonic part models the periodic parts, or voiced speech, and the noise models the stochastic, or unvoiced, parts of speech. The deterministic component is modeled by sums of harmonically related sinusoidal components with various amplitudes. The stochastic part comprises the residual signal when the sinusoidal components have been extracted from the original signal i.e. rather than modeled purely in the frequency domain, the stochastic part is obtained from real parts of speech, for example from plosives or fricatives etc.

The synthesis signal $\hat{s}(t)$ is modeled as the sum of harmonic components and a noise signal as shown in Equation 1.4.

$$\hat{s}(t) = \sum_{k=-K(t)}^{K(t)} A_k(t) \exp(jktw_o(t) + e(t)) \qquad \text{Eqn 1.4}$$

where $A_k(t)$ is the complex harmonic amplitude at time t, $w_o(t)$ is the fundamental frequency and $e(t)$ is the stochastic component. These parameters are updated at specific time-instants. The parameters are estimated as follows:

- Fundamental frequency $w_o(t)$ is estimated using a standard time-domain pitch detection algorithm, such as Hess (1983).
- On voiced parts of speech, the values of the amplitudes and slopes of pitch harmonics are estimated using a weighted least-squares method (Laroche, 1989).
- The residual signal is obtained by subtracting the deterministic part from the original signal in the time-domain.

At synthesis time, the deterministic and stochastic components are synthesised separately then added together. The deterministic part is synthesised by overlap-adding a stream of ST-signals $s_i(t)$ at time-instants t_i in a PSOLA synthesis manner (Moulines & Charpentier, 1990). The ST-signals are obtained from the harmonic parameters by applying a Hamming window centered at t_i. The stochastic component is obtained by filtering Gaussian noise. Time-scale modifications

are achieved by determining the number of synthetic pitch-periods that need to be generated from the parameters at time t_i.

The mixed voice segments are free from any buzzy quality as the stochastic and deterministic components are dealt with separately. As it is a parametric model it is very flexible allowing modification of speaker voice qualities and timbral aspects of speech quality (Syrdal *et al.*, 1998b).

1.5.4 Pitch-Synchronous OverLap-Add

The family of Pitch-Synchronous Overlap-Add (PSOLA) techniques (Charpentier & Stella, 1986) was developed by France Telecom. The PSOLA algorithm and its variants do not synthesise speech themselves, but allow pre-recorded segments of speech to be concatenated and can modify the prosody (pitch and duration) of the speech signal, which may be necessary when creating novel utterances. This technique avoids parameterization of the speech; parameterization inherently degrades the segmental quality. This is opposed to LP and sinusoidal models, which decompose the signal into separate source and vocal tract models.

The PSOLA algorithm involves three stages: analysis, modification and synthesis. The three stages are illustrated in Figure 1.8. Speech waveforms are first *analysed*; the speech signal is broken down into a sequence of Short-Term (ST) signals by windowing it at successive intervals with a sequence of pitch-synchronous windows, such as Hanning windows. The Hanning window is a symmetrical window that restricts the analysis to the section of waveform under the window by rendering the rest of the signal zero. The Hanning windows are centred pitch-synchronously on pitch markers (which are placed at the glottal closure instant during voiced portions of the signal, and at a constant rate for unvoiced parts). The length of the Hanning window is set so that adjacent ST-signals overlap. Generally, each window's length is set to be twice the local pitch period. If the window is short enough, the signal under the window can be considered stationary.

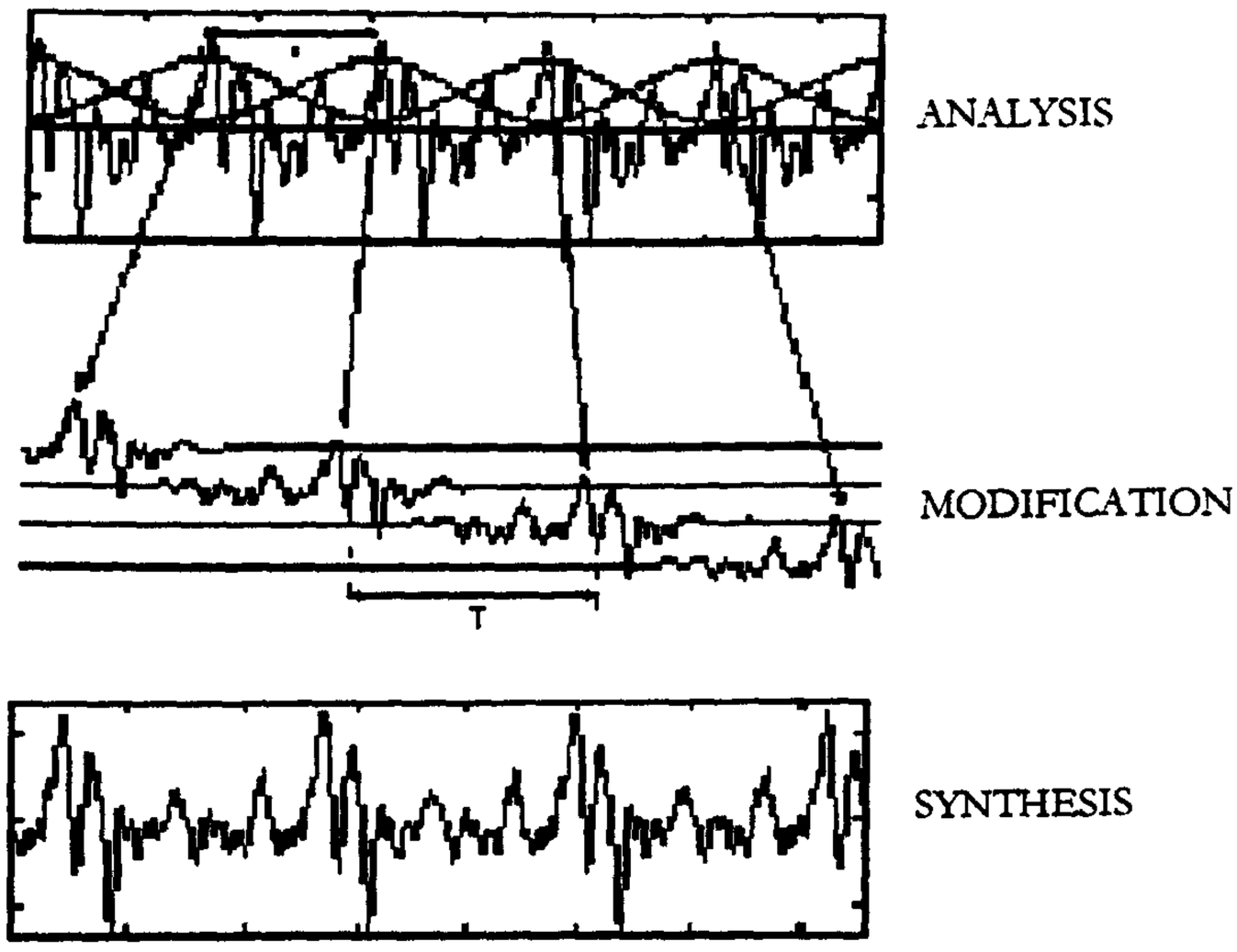


FIGURE 1.8 THE PSOLA OPERATION

Prosodic *modifications* can then be imposed on this intermediate representation. Pitch modifications are produced by altering the spacing between the ST-signals; the original pitch period is altered to a new period. Duration modifications are achieved by repeating or deleting the ST-signals.

Synthesis is achieved by the recombination of the modified intermediate representations to produce the final synthetic signal using an overlap-add (OLA) operation (Moulines & Charpentier, 1990) that adds the new ST-signal sequence together.

The family of PSOLA algorithms for manipulating the prosody of speech waveforms all use this three stage technique, although the above explicitly describes Time-Domain PSOLA. The PSOLA variations are described in the following section.

1.5.4.1 PSOLA variations

Several variations of the PSOLA operations are available, such as Time-Domain (TD), Frequency-Domain (FD), Linear-Prediction (LP), and Multi-Band Resynthesis (MBR) PSOLA.

Time-Domain (TD) signal processing algorithms, such as the Time-Domain Pitch-Synchronous OverLap-Add (TD-PSOLA) algorithm (Hamon *et al.* 1989, Moulines & Charpentier 1990), are of particular interest as they are computationally inexpensive.

TD-PSOLA generates natural speech with minimal effect on segmental quality (Bigorgne *et al.* 1993, Moulines *et al.* 1990), although some perceptible distortion is introduced for certain prosodic modifications. TD-PSOLA provides limited smoothing capabilities between concatenated speech segments; pitch and spectral mismatches at segment boundaries are not minimised, which may lead to audible discontinuities. It would be ideal to eliminate such mismatches before applying the TD-PSOLA algorithm.

One such solution involves resynthesising the voiced parts of the segment inventory with a standard pitch, as performed in the Multi-band Resynthesis PSOLA (MBR-PSOLA or MBROLA for short) algorithm (Dutoit & Leich, 1993). Additionally, all of the harmonics in the voiced instances of speech are given fixed initial phases for each period. Elimination of pitch mismatches is achieved inherently by the resynthesis process. As the two segments to be concatenated have the same pitch and identical harmonic phases, spectral mismatches can be eliminated by simple time-domain interpolation.

MBROLA provides a speech signal with minimal mismatches between segments upon which the TD-PSOLA algorithm can be applied to generate the required prosody. MBROLA has good segment smoothing capabilities and good prosody matching, producing a natural speech output. This is at the expense of much greater distortion, in the form of buzzyness, introduced during the resynthesis process.

A Linear Predictive Pitch-Synchronous OverLap Add (LP-PSOLA) approach has also been developed (Moulines & Charpentier 1990, Edgington & Lowry 1996). LP analysis is performed on the speech signal to separate the signal's source and filter components; at each time instant the spectral envelope is estimated and used to extract the excitation. TD-PSOLA pitch and duration modifications can then be applied directly to the LP filter's excitation or *residual* signal. After modifications, the new signal is produced by recombining the modified source with the spectral

envelopes. Moulines & Charpentier (1990) report that LP-PSOLA produces perceptibly less distortion when applied to the residual waveform than conventional TD-PSOLA, although this is at the expense of far greater complexity.

A frequency-domain version of PSOLA has also been proposed (Moulines & Charpentier, 1990). In FD-PSOLA, the Short Term Fourier Transform (STFT) for each analysis window is calculated to provide a frequency-domain representation of the short-term signals; the signal is separated into source and filter components. The pitch and duration modifications are performed in the frequency domain, allowing better control over the spectral envelope, by altering the spacing between the pitch harmonics. The modified representation of the signal is then converted back into the time-domain by taking the inverse Fourier Transform. FD-PSOLA is more flexible than TD-PSOLA and also supports modifications of voice quality (Valbret *et al.*, 1992), by alteration of the speech signal's spectral characteristics. The main drawback is that FD-PSOLA is much more complex.

1.5.4.2 Drawbacks of PSOLA

The PSOLA method requires the use of small Hanning windows (Linggard, 1985), which must contain only a single pitch pulse for the signal to be considered stationary. If this is not the case, a mismatch occurs during synthesis between the new synthesis frequency and the frequency inherent to each short-term signal. Conversely, when too small a window is used, formant bandwidths are broadened due to the poor estimation of the spectral envelope for each ST-signal. This results in some alteration of the amplitude of pitch harmonics for voiced speech, and can cause a reverberation to be heard. It is therefore important that the implementation of PSOLA uses a robust pitch detection algorithm.

Even when the correct sized window is used, PSOLA can introduce artifacts, described as 'hoarseness and roughness' by Kortekaas & Kohlrausch (1997a) and 'buzzyness' during this thesis, into the signal. Kortekaas and Kohlrausch state that these artifacts are often difficult to predict, and even if modification with PSOLA does not lead to the perception of artifacts, it does affect its spectral content.

PSOLA maintains good quality for moderate modifications. Problems can occur when increasing the durations of unvoiced sounds. Slowing speech involves repeating the unvoiced short-term signals, which for modifications of greater than a factor of two can result in tonal noise or 'buzzyness'. This can be avoided for unvoiced sounds by removing this local periodicity through reversing the time axis for repeated short-term signals. Voiced fricatives also suffer similar problems for increasing duration and increasing pitch, but this solution cannot be applied here due to the voiced component. Buzzyness qualities may also appear for large pitch modifications especially for female and children's voices.

1.5.5 Corpus-based Techniques

Simple diphone speech synthesis systems store only one segment per phonetic type in a waveform inventory, which have usually been excised from prosodically neutral speech and so fail to model the dynamic characteristics of prosody. Prosodic parameters, such as fundamental frequency, duration and intensity, can be manipulated by algorithms such as TD-PSOLA, but these parameters may not be the only important aspects of prosody (Campbell & Black, 1996). Spectral effects, which cannot be modeled by TD-PSOLA, are responsible for other additional aspects, such as variations in phonation or changes in voice quality (breathy voice, pressed voice etc). The corpus-based approach aims to preserve these aspects by storing one segment for each phonetic and prosodic context (Campbell & Black, 1996) in a corpus of prosodically rich, continuous speech, leading to tens or hundreds of instances of each segment. The key issue then, is to select the longest available string of segments that will sound natural in a given phonetic context. This non-uniform approach was first investigated by Sagisaka (1988) and Takeda *et al.* (1990) for rule-based synthesis. Later parallel research was carried out, known as *unit selection*, by Black & Campbell (1995) for concatenative synthesis. Ideally the whole of the required utterance would be found and simply played back requiring no concatenation points and no prosodic modifications. This is extremely unlikely, although constructs longer than diphones may be found and less signal processing may be required. Automatic segment or unit selection (Hunt & Black, 1996) from the corpus is achieved by minimising distance measures (Gray & Markel, 1975) between target segments (predicted by preceding modules of the synthesis system) and segments in the corpus.

Storing such diverse segments may give larger potential pitch and spectral mismatches between joining segments, leading to possible audible discontinuities. A synthesis specification is generated giving phoneme identities, target durations and fundamental frequencies. All phonemes with the same name in the inventory are selected. A trade off between good prosody matching and good segment concatenation is necessary. Campbell and Black (1996) achieve this by balancing two sets of distance measurements. The first is the objective distance between features of the selected segment and the target segment, and the second is a measure of the quality of the join between a selected segment and its previously adjoining segment. To compute the target vector, only features that can be computed by analysing the input text are available. To compute the continuity or join vector, all features are available that have been computed during the offline annotation of the speech segments. Their distance measures are calculated using weighted vectors of features such as phonetic context (neighbouring phonemes, position in phrase, direction of pitch/power change etc.), duration, log power, and mean fundamental frequency. A network is constructed and the costs are assigned to each unit and to the links between each unit. The lowest cost path through the network is then selected.

Labelling such a large speech corpus with so many features is labour intensive, but the main drawback of this approach lies in finding the relevant features for the distance measures and the correct weighting between them (Möbius, 2000). Research by Boëffard *et al.* (1992) and Campbell & Black (1996) indicates that minimising certain distances between segments does not necessarily lead to perceptually better speech output; little is known about the perceptual relevance of such distance measures (Wouters & Macon, 1998). Problems arise due to the fact that human perception cannot be measured objectively and additionally, acoustic properties that can be measured are sometimes imperceptible (Dutoit, 1997).

Distance measures used in synthetic speech remain difficult to assess in terms of human perception. The most common way to test them is to measure participants' perceptions of the synthesised speech using a particular speech corpus. Unfortunately, such conclusions about a certain measure may not be valid when applied to a different corpus.

Finally, the speech corpus cannot be infinitely large, implying that even the best segment selection technique will not provide an exact prosodic match for the desired utterance. The

CHATR TTS system (Black & Taylor 1994, Campbell *et al.*, 1998) uses no signal modification. It performs better for closed-domain synthesis as increasing quality is only achieved by substantially increasing the amount of source data. Conkie (1999) concludes that signal processing may be applied selectively alongside the corpus-based strategy. Though some signal processing may be required, the manipulation of pitch and duration of the segments is minor, introducing minimal distortion.

Deketelaere *et al.* (2001) state that the corpus-based approach is a promising technique and is possibly the future of speech research, although quality is achieved with extremely high storage requirements. They provide an example of the AT&T NextGen system, which requires several hours of speech and runs over a server via the Internet, making such systems unusable for current low-cost or hand-held electronic devices.

1.6 Choice of Synthesis Model

Current speech models available for concatenating speech segments and imposing prosody have been critically evaluated.

TD-PSOLA (Moulines & Charpentier, 1990) is currently one of the most popular concatenation methods (Syrdal *et al.*, 1998b). It involves very low computational loads and provides high quality synthesis in the main. TD-PSOLA does not suffer from the inherent buzzyness of LP synthesis as it is not a parametric model. TD-PSOLA also retains fine spectral details without the smoothing or distorting effects of formant or LP approaches. Moulines & Charpentier (1990) claim all PSOLA algorithms (TD-, LP-, FD-) are comparable in quality of output, and all are better than LP synthesis in formal listening tests.

TD-PSOLA has limitations being a non-parametric method; only the voice of the speaker may be synthesised and spectral mismatches may occur at concatenation boundaries unless units are chosen carefully. Finally, a buzzy quality may be perceived when some prosodic modifications are applied.

Sinusoidal approaches (Macon & Clements, 1996) and hybrid stochastic approaches (Stylianou, 2001) are more flexible than TD-PSOLA and MBROLA for compression, modification and smoothing but are ten times more computationally expensive.

Many alternative methods to PSOLA have been developed (Syrdal *et al.* 1998b, Violaro & Boëffard 1998). Laroche *et al.* (1993) use a Harmonic and Noise Model (HNM), which they found to eliminate many of the artifacts that occur during duration modification with PSOLA. HNMs also eliminate the buzzyness that occurs during modification of mixed-voice segments. Breen (1998) and Stylianou *et al.* (1995) conclude that although these models perform better on voiced fricatives and unvoiced frames, and provide better spectral control, they are computationally more expensive, not as robust, and produce speech of a slightly lower quality than PSOLA overall.

Dutoit & Leich (1994) compared an LP model, TD-PSOLA, a pitch-asynchronous hybrid Harmonic and Noise model and the MultiBand Resynthesis PSOLA (MBROLA) in terms of naturalness and intelligibility. MBROLA was considered comparable with TD-PSOLA, with the HNM and LP model third and fourth. Violaro and Boëffard (1998) compared TD-PSOLA and a HNM and found their quality to be similar but judged naturalness to be better for unvoiced speech using the HNM method. Charpentier & Stella (1986) compared PSOLA and LP and concluded that PSOLA provides a more natural output than both an LP and a multipulse coding system (Stella & Charpentier, 1985). Overall, TD-PSOLA compares comparably and often favourably in terms of intelligibility and naturalness, with other models.

The corpus-based approach provides concatenative synthesis with greater flexibility by storing more diverse speech units and threatens to eliminate the need for signal processing in the future. Currently, although it does reduce the amount of modification, signal processing may still be necessary as it is not possible to store every combination of segments in every prosodic context. Conkie (1999) using a corpus-based approach found that synthetic sentences were preferred when no prosodic modifications were performed, although limited prosodic modification did appear to be beneficial in terms of improved naturalness by smoothing mismatches at segment boundaries and allowing suitable prosodic modifications for new utterances.

Portele (1998) advocates that TD-PSOLA should be used in conjunction with a corpus-based approach, and currently this combination appears to provide the most promising solution to generating high quality, natural synthetic speech. The use of a speech corpus increases the flexibility of concatenative speech by retaining details, such as variations in phonation or changes in voice quality, that simple diphone approaches cannot model. Unfortunately, even the largest corpus cannot provide every prosodic combination of pitch and duration. The use of a signal processing algorithm such as TD-PSOLA, has the advantage of producing a less distortion output due to its non-parametric nature. Conversely, this lack of a parametric representation means it is inflexible when required to model spectral aspects of speech such as variations in phonation and voice quality if used with a simple diphone system. The use of a corpus and TD-PSOLA together provide the most flexible approach to concatenative synthesis; TD-PSOLA allows small prosodic modifications if such values are not present in the corpus, and the corpus retains spectral details of speech that TD-PSOLA cannot model.

The TD-PSOLA algorithm is of great interest to the speech synthesis community due to its overall simplicity and success for moderate modifications. Its main drawback is the introduction of perceptible distortion into speech, in the form of buzzyness, for some modifications, and it is this issue that is addressed in the thesis to improve the quality of the speech output.

The research investigates the effects of the application of the TD-PSOLA algorithm on natural speech when used for small prosody modifications that may be necessary in a corpus-based system. The aim is to develop a framework that minimises the introduction of perceived distortion and thus retains the quality of the speech output. This may be achieved by careful design and use of a speech corpus tailored to the needs of TD-PSOLA.

1.7 Summary

In this chapter, a brief description of human speech production, its physical representation and phonetics, used to describe the speech in an abstract format, was presented. During the thesis, the speech under investigation will be described using the SAMPA notation and waveforms, spectra and spectrograms will be used to depict and analyse the speech.

A description of a general TTS system was then given. Speech synthesis, the main area of this research, was set in context as part of the whole TTS process, and the complex interaction of the modules in a TTS system was illustrated.

The three main synthesis strategies were then described and their advantages and disadvantages discussed. Articulatory synthesis, although capable of producing high quality synthesis, is deemed too complex for current applications. Formant synthesis is inherently buzzy and relies on the careful choice of parameters, which is a notoriously difficult process. Concatenative synthesis performs well in terms of intelligibility and quality although it suffers from inflexibility due to the use of recorded speech.

Concatenative synthesis is currently a very popular approach and the issue of inflexibility may be overcome by the choice of synthesis strategy employed. The more common synthesis strategies used during concatenative synthesis were then examined. The corpus-based approach shows promise in terms of efficiency and flexibility and currently appears to be one of the more attractive strategies. The corpus stores multiple versions of segments in many contexts and hence retains details such as varying phonation and voice qualities. Due to the variability of natural speech, it is not possible to store all speech segments in every prosodic context, and so a signal processing algorithm may still necessary. LP models were found to be inherently buzzy, sinusoidal models were found to perform poorly for unvoiced speech, and some of the PSOLA family (FD- and LP-PSOLA) were found to be complex operations. TD-PSOLA does not suffer the inherent buzzyness of the parametric models although this leads to some inflexibility when spectral aspects need to be modelled if used with a simple diphone inventory. The strategy of employing TD-PSOLA in conjunction with a corpus-based approach provides greater flexibility; the corpus provides aspects of speech that TD-PSOLA cannot model, and TD-PSOLA ensures a robust output if segments in the corpus do not have suitable prosody for arbitrary sentences.

TD-PSOLA is a very popular and efficient algorithm, but its main drawback is the introduction of perceptible distortion, in the form of buzzyness, into the signal during some prosodic modifications. This research focuses on developing a framework to reduce such distortion and

retain the resulting speech quality when TD-PSOLA is used for moderate modifications in a corpus-based system.

The remainder of the thesis is structured as follows. Chapter 2 describes the TD-PSOLA algorithm in greater detail and reviews previous research into the effect of the algorithm on speech quality in terms of the introduction of perceptible distortion. The aim is to suggest parameters that may contribute to the occurrence of this distortion. These may then be investigated further to determine how to design a framework to remove or reduce such artifacts. Chapter 3 describes existing subjective listening tests and procedures used to evaluate the performance of various aspects of speech synthesis systems. These tests and practices are applied where possible in the listening tests documented in Chapter 4. The listening tests are undertaken to determine the perceptual effects of the TD-PSOLA algorithm when used for pitch modification of natural speech. Chapter 5 analyses data for patterns of co-occurrence and proposes a framework to reduce distortion. This is in the form of a novel corpus design tailored to the requirements of TD-PSOLA, a signal processing distortion measure that may be used to select a segment from the corpus that will result in less distortion, and a special selection process for highly problematic phonemes. Chapter 6 documents a listening experiment that evaluates the performance of such a framework. In Chapter 7 the findings of the research are discussed and possible future work is recommended.

BLANK IN ORIGINAL

Chapter 2. The TD-PSOLA algorithm and previous research

2.1 Introduction

Many commercial speech synthesis systems such as the ProVerbe TTS system (Elan Informatique) incorporate the TD-PSOLA algorithm with great overall success. Speech, which has been pitch and duration modified using the algorithm, is reportedly of high intelligibility and quality in the main (Moulines & Charpentier 1990, Donovan & Woodland 1999, Laroche *et al.* 1993, Dutoit & Leich 1994). The research does indicate though that PSOLA may introduce perceptible distortion into the speech signal in the form of buzzyness for certain modifications.

This chapter describes the operation of the algorithm in greater detail using a mathematical model and then illustrates the Praat software (Boersma & Weenink, 1999) implementation of the algorithm. The next section seeks to identify some potential problems associated with TD-PSOLA. To this end, existing research concerning the algorithm is documented, and some of the work is replicated to illustrate the basic objective signal distortions introduced by the algorithm on pure sine waves and then more complex single formant signals. Such signal distortions may not always be perceptible, especially for more complex signals such as natural speech. Possible causes of perceptible distortion are then discussed. These consist of incorrect pitch marking by the algorithm, the influence of the choice and size of the analysis window used by the algorithm, the extent of manipulation applied to the speech, and the speech type upon which the algorithm is acting. Finally, the chapter concludes with a discussion of these issues raised by existing research, and how they may be addressed to inform the design of a framework to reduce perceptible distortion.

2.2 The TD-PSOLA Algorithm

This section describes the operation of the algorithm based on Moulines & Charpentier (1990). The TD-PSOLA algorithm involves three steps; analysis, modification and synthesis.

2.2.1 Analysis

An original speech signal $x(n)$ is analysed to produce an intermediate representation. This representation is non-parametric and consists of a series of short-term signals $x_m(n)$. These are obtained by multiplying the original signal with a sequence of pitch-synchronous analysis windows $h_m(n)$:

$$x_m(n) = h_m(t_m - n)x(n) \qquad \text{Eqn 2.1}$$

The analysis windows are positioned at successive instants on pitch marks t_m, which are located at pitch-synchronous intervals on the voiced parts of speech and at a constant rate on unvoiced parts. The windows are Hanning windows and have a length determined by the local pitch period. They are longer than one pitch period so that there is some overlap between adjacent short-term signals and may range from twice the local period to four times, giving 50% and 75% overlap respectively.

2.2.2 Modification

This sequence of analysis short-term signals $x_m(n)$ is modified into a new sequence of synthesis short-term signals $\tilde{x}_q(n)$, which are repositioned on a new set of synthesis pitch marks $\tilde{t}_q$. Pitch modification requires modifying the delays between the short-term signals and duration modification requires the modification of the number of short-term signals; increased pitch requires decreasing the delays, and increased duration involves the repetition of some of the short-term signals.

2.2.3 Synthesis

Several overlap-add (OLA) methods exist that may be used to recombine the short-term signals to give the modified synthetic speech signal $\tilde{x}(n)$. Moulines & Charpentier (1990) describe the least-square overlap-add synthesis method (Griffin & Lim, 1984) and a more simple overlap-add

procedure (Allen, 1977). The overlap-add operation is at its simplest when the synthesis window is twice the local pitch period and may be reduced to a linear combination of the modified short-term signals:

$$\tilde{x}(n) = \sum_{q} \tilde{x}_q(n)$$

Eqn 2.2

The basic operation of the TD-PSOLA algorithm has been presented, and the following section describes the Praat software (Boersma & Weenink, 1999) implementation of TD-PSOLA that will be used during this work.

2.3 The Praat Software Implementation of the TD-PSOLA Algorithm

Praat (Boersma & Weenink, 1999) is a system for doing phonetics, developed by Paul Boersma and David Weenink in the Phonetic Science Department at the University of Amsterdam. It is a shareware program, which provides a flexible tool for speech research allowing pitch analysis, spectrographic analysis and speech synthesis amongst many other functions. For more information, see http:/www.fon.hum.uva.nl/praat/ or contact Dr Boersma: paul.boersma@hum.uva.nl.

The following sections describe how the stages of analysis, modification and synthesis for the TD-PSOLA algorithm are achieved using this software.

2.3.1 TD-PSOLA Analysis

Pitch analysis is performed on the time domain waveform using an acoustic periodicity detection algorithm based on an autocorrelation method (Boersma, 1993). This method is reported to be more accurate, robust and noise-resistant than methods using cepstrum or combs, or original autocorrelation methods. The minimum default pitch is set at 75Hz; any candidates below this level will be ignored.

A pitch contour results from this analysis, giving frequency values and voiced/unvoiced decisions. The pitch contour is converted to a frequency of points structure representing glottal

pulses which are positioned on the voiced intervals on the waveform. Figure 2.1 shows the Praat software editor window. In the top section, the time domain waveform is seen with glottal pulses (shown as vertical lines) positioned on the voiced parts of speech. The middle section shows the original pitch contour of the voiced parts of speech, which in this example, is flat or static and has a fundamental frequency of 223.8Hz. The bottom section facilitates the modification of the duration of the speech.

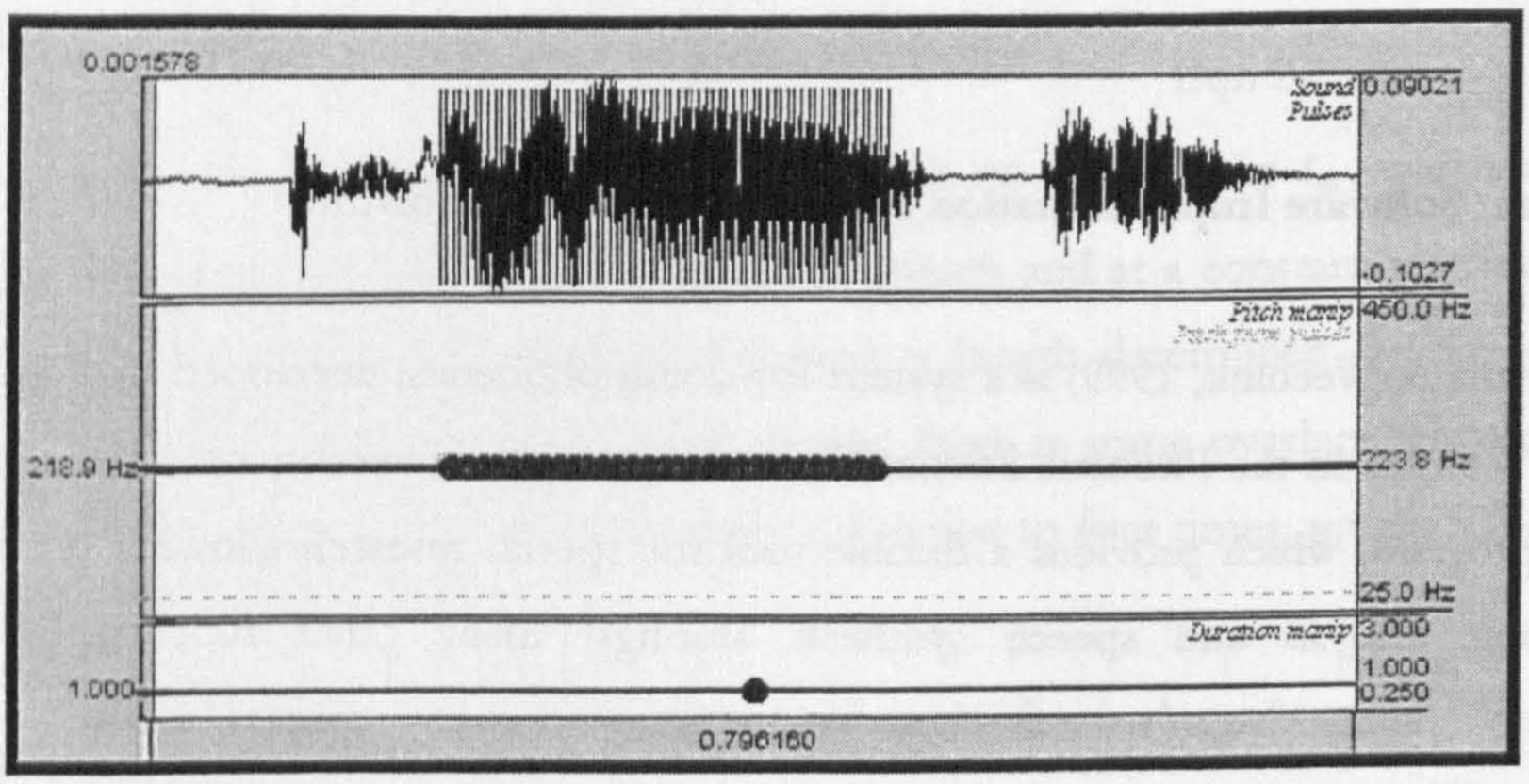


FIGURE 2.1 PRAAT SOFTWARE EDITOR WINDOW

2.3.2 TD-PSOLA Modification

The pitch contour when converted to the point structure may then be viewed or edited in the editor window. The editor in Figure 2.2 shows the original sound with the point structure representing the glottal pulses. The pitch points in the middle section may be simplified by removing certain points, and can be edited by moving them up or down to increase or decrease the resulting pitch. In the duration window points can be added and edited to manipulate the relative durations of parts of the waveform.

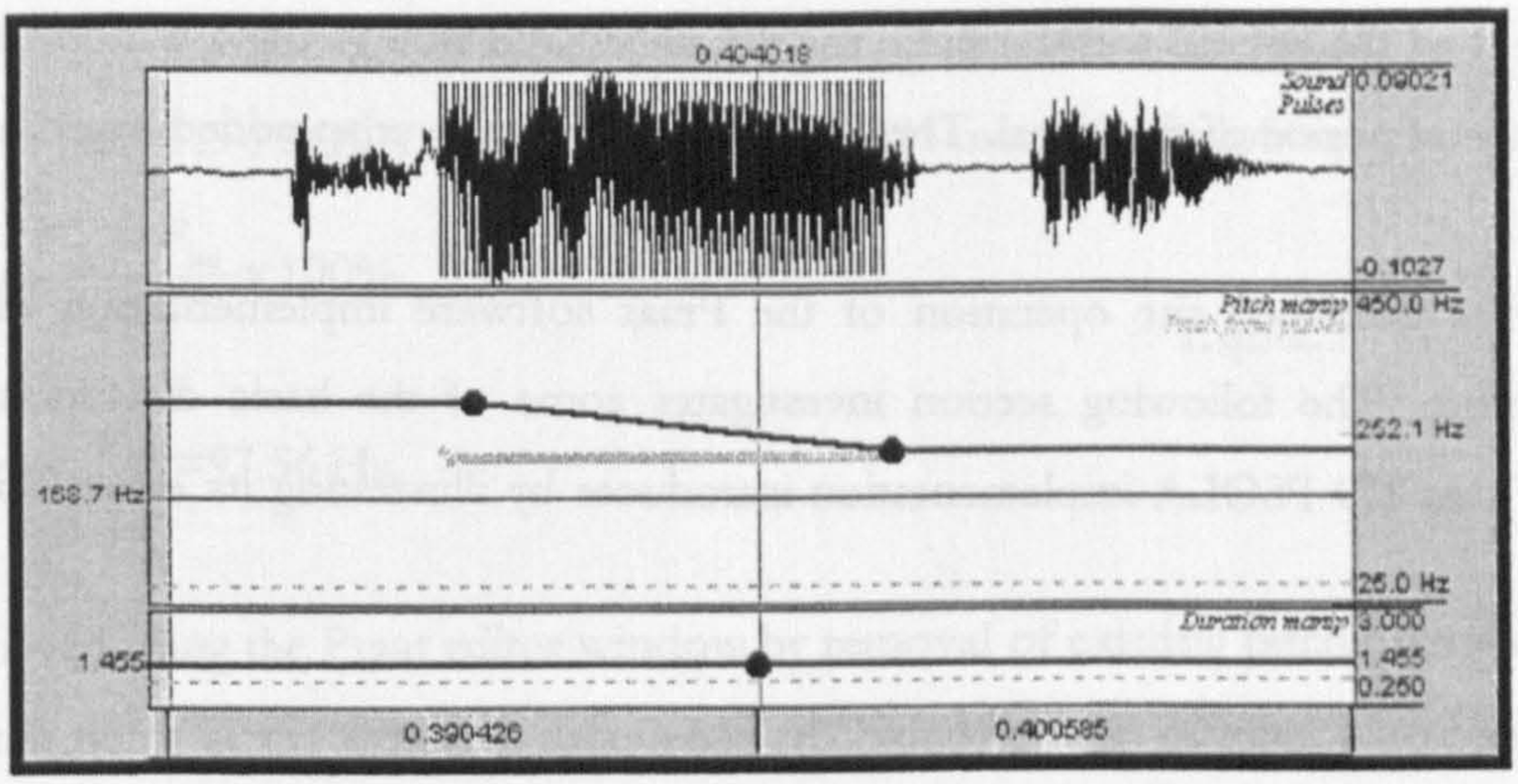


FIGURE 2.2. TD-PSOLA PITCH AND DURATION MODIFICATION IN PRAAT

2.3.3 TD-PSOLA Synthesis

The modified sound waveform is obtained from the analysis phase by taking the new pitch contour information (consisting of a time-stamped pitch contour without voiced/unvoiced information) and generating new points along the entire time-domain waveform.

The new acoustic pitch contour is interpreted as the frequency of occurrence of points representing the sequence of glottal closures during vocal fold vibration. The points are generated along the entire waveform as the voiced/unvoiced information is not taken into account yet.

The period information in the pulses is used to remove all points that lie in voiceless regions of the time-domain waveform. This is judged to be places where the distance between adjacent points in the original pulses is greater than 20ms.

The voiceless parts are then copied from the source waveform to the target waveform, repeating some ST-signals or deleting some ST-signals if the local duration is greater or less than 1.

For each new target point, the nearest source point is identified and the ST-signal centred on the source point is copied to the target sound and positioned at the target point. The window used is bell-shaped, called a Hanning window, whose left-hand length is the minimum of the left hand

periods adjacent to the source and the same for the right-hand side giving a window size of 2P, where P is the local period of the signal. The ST-signals are then overlap-added together.

This section has described the operation of the Praat software implementation of the TD-PSOLA algorithm. The following section investigates some of the basic distortions that the application of this TD-PSOLA implementation introduces by illustrating its effect on pure sine waves.

2.4 The Basic Distortions introduced by TD-PSOLA in Pure Sine Waves

Kortekaas & Kohlrausch (1997a) pitch manipulated a single pure tone to illustrate the basic signal distortions produced by the PSOLA operation. The pure tone may be thought of as a component of a harmonic spectrum e.g. a 1000Hz sine wave may be assumed to be the 10th harmonic of a 100Hz fundamental, or the 4th harmonic of a 250Hz fundamental.

Their work has been replicated here using the Praat software (Boersma & Weenink, 1999) to illustrate these distortions and to determine whether such distortions are perceptible. The following terminology will be used: T_a the analysis rate (ms) or the rate at which the signal is decomposed, T_s the synthesis rate or the rate at which the signal is recombined, F_{wa} the fundamental frequency of the original signal, and F_{ws} the fundamental frequency of the synthesised signal.

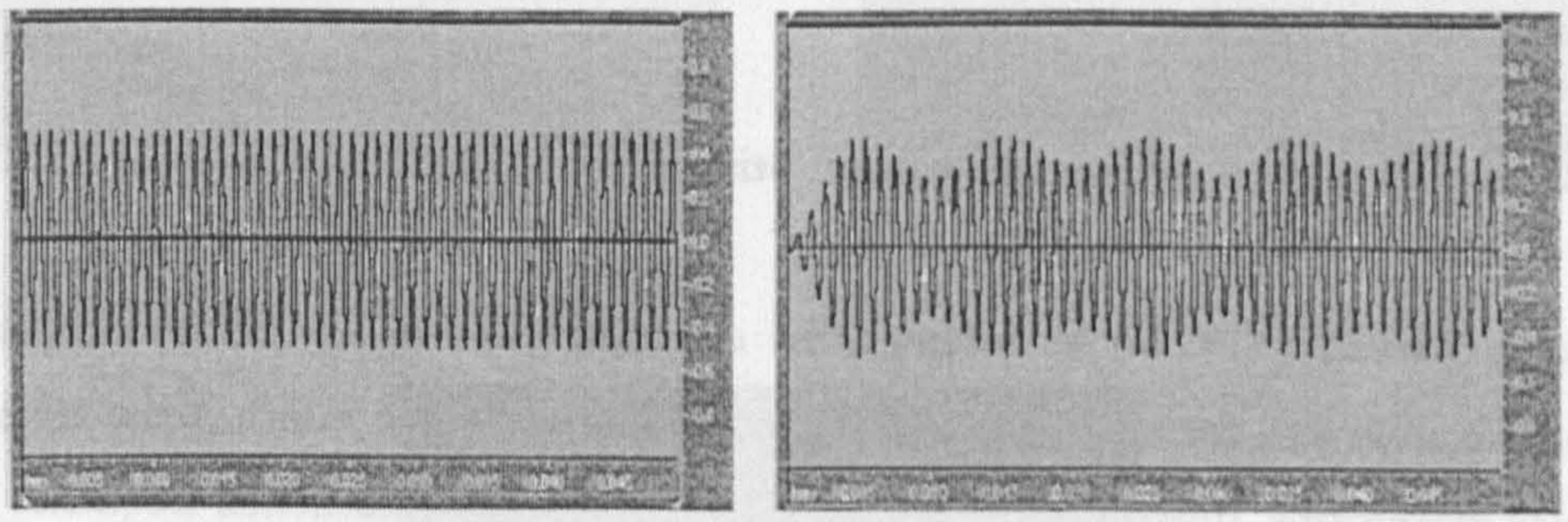
Initially, a 1000Hz sine wave was generated with the Praat software using the formula ½*sin(2*pi*1000*x). This signal was analysed as described in Section 2.2.1 to produce an intermediate representation of a series of Short-Term (ST) signals. In Praat this involved setting the maximum pitch to be considered in the analysis to 120Hz (found by trial and error) so that Praat pulses are positioned every 10th cycle of the 1000Hz waveform; assuming the pure tone is a harmonic of a 100Hz fundamental, pulses are positioned on each hypothetical cycle of the 100Hz fundamental. As a result, the signal will be windowed or decomposed every 10ms (T_a=10ms). The window length was set to 0.02 seconds, which represents 2*local pitch length of a 100Hz fundamental.

This signal was then TD-PSOLA modified by ΔF=-2.44% using the following relationship:

$$\Delta F = \frac{F_{ws} - F_{wa}}{F_{wa}} \times 100\% \qquad \text{Eqn 2.3}$$

giving a value of F_{wa} =97.56 Hz.

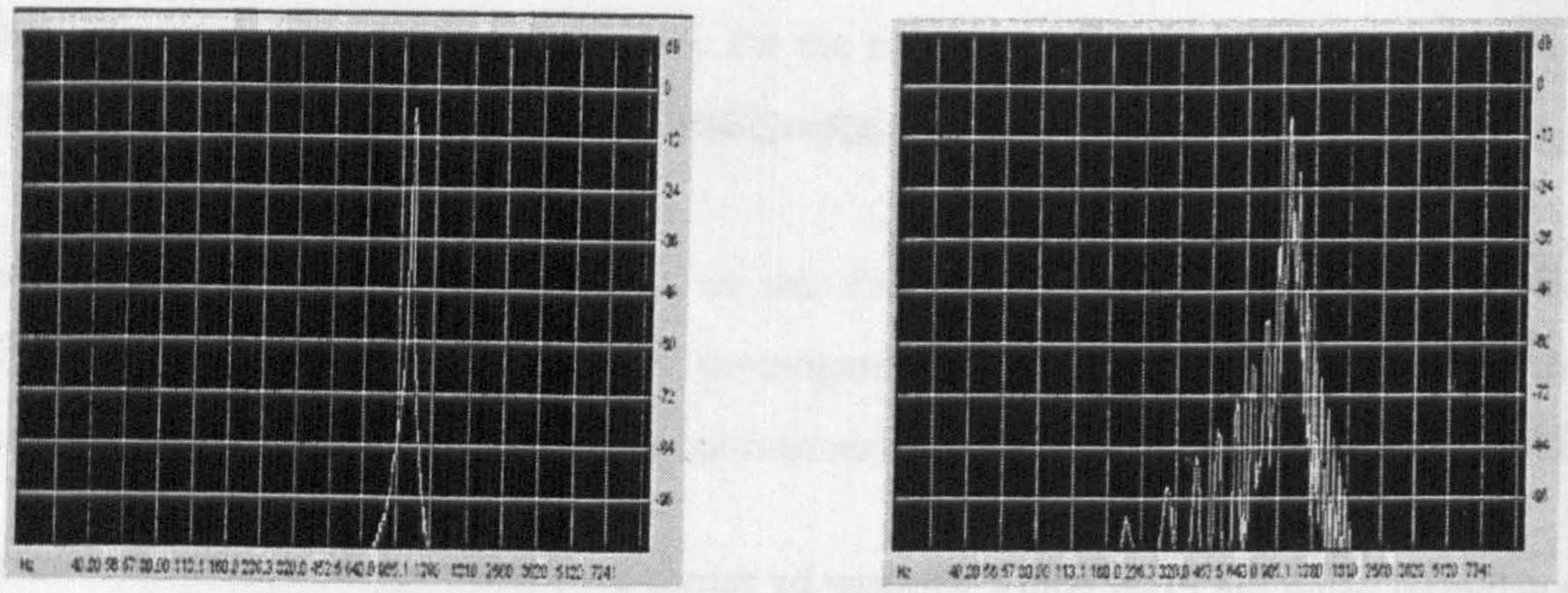
This was achieved using the Praat editor window by removal of existing pitch points and addition of a new pitch point at the synthesis frequency of 99.56Hz. The original waveform shown in Figure 2.3 (a) was decomposed at intervals of T_a=10ms, then TD-PSOLA modified to F_{wa} =97.56 Hz at T_s =10.25ms to produce the waveform in 2.3 (b). The TD-PSOLA modified waveform shows amplitude modulation (AM) in its envelope when compared to the original waveform.



(a) original waveform (b) TD-PSOLA modified waveform

FIGURE 2.3 TD-PSOLA DISTORTIONS: AMPLITUDE MODULATION

The distortion can also be seen in the frequency domain by calculating the Fast Fourier Transform (FFT) of the signal using a Hanning window (FFT size 4096). Figure 2.4 (a) shows the spectrum of the unmanipulated 1000Hz signal, and 2.4 (b) shows the TD-PSOLA modified version. The spectrum of the signal in Figure 2.4 (b) shows frequency modulation (FM) in the fine structure. Broadening of the spectral envelope due to the addition of side components in the form of harmonics of the assumed 100Hz fundamental is evident.



(a) spectrum of original signal

(b) spectrum of TD-PSOLA modified signal

FIGURE 2.4 TD-PSOLA DISTORTIONS: FREQUENCY MODULATION

These images illustrate some of the basic TD-PSOLA signal distortions on a pure sine wave. This distortion may be perceptible; the timbre of sound is determined by the shape of the spectral envelope and informal listening finds the modified version very rough or 'hoarse' sounding in comparison to the unmodified pure tone. The investigation in the following section extends to more complex single formant signals.

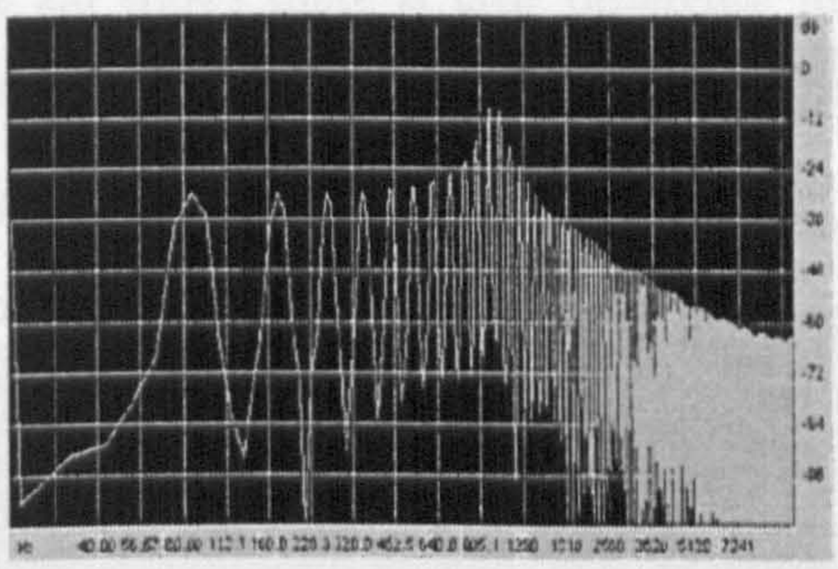
2.5 TD-PSOLA Distortions in Single Formant Stimuli

Kortekaas & Kohlrausch (1997a) investigated the effect of TD-PSOLA pitch manipulation on single formant stimuli and the work is recreated here to illustrate the effect of the algorithm on more complex waveforms. Initially, a signal was generated with a fundamental frequency of 87Hz and a formant frequency of 1000Hz with a bandwidth of 50Hz. A second signal was generated with a fundamental frequency of 100Hz and a formant frequency of 1000Hz, with a bandwidth of 50Hz. This signal was then TD-PSOLA modified to produce a signal with a fundamental frequency of 87Hz.

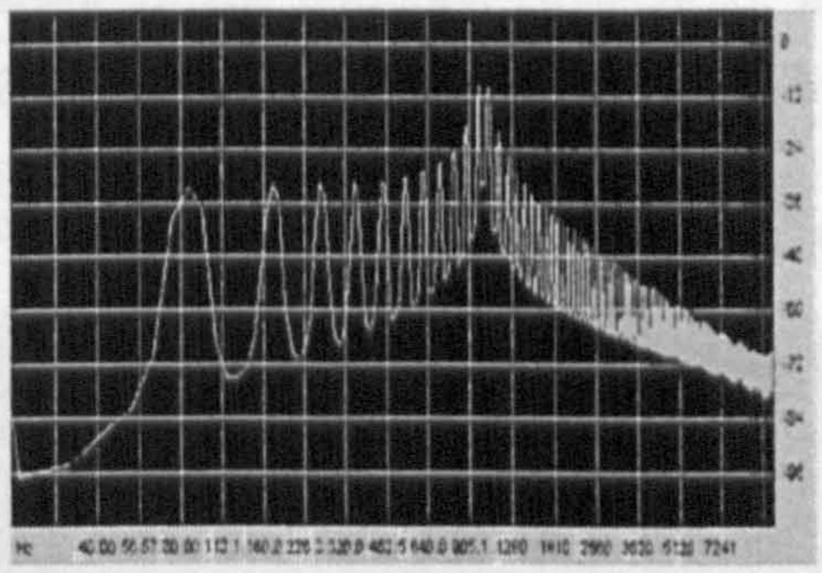
Kortekaas & Kohlrausch (1997a) also investigated the effect of incorrect marking of glottal closure instances. This was achieved by altering the position of the pitch markers and hence the position of the analysis windows on the 100Hz signal to an offset of 50%. This signal was then pitch-modified to 87Hz.

The single formant signal was created using the Praat software by generating a source (pulse train) at 87Hz and a filter, having a passband at 1000Hz with a 50Hz bandwidth. The single formant signal was created by filtering the pulse train.

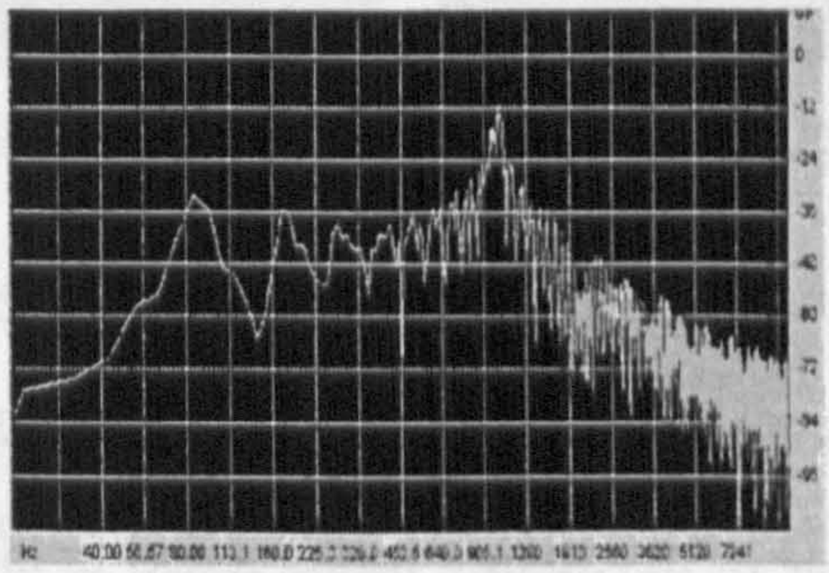
Figure 2.5 (a) shows the spectrum of the unmanipulated single formant signal composed of a F0 of 87Hz and a formant frequency of 1000Hz, formant bandwidth 50Hz, 2.5 (b) the TD-PSOLA manipulated signal generated with an F0 of 100Hz, formant frequency of 1000Hz, decomposed at F_{wa} =100Hz and resynthesised with a fundamental frequency of F_{ws} =87Hz, and 2.5 (c) a signal synthesised with the window offset by 50%.



(a) signal synthesised with f0 87Hz



(b) TD-PSOLA modified signal from f0 100Hz to 87Hz



(c) TD-PSOLA modified signal with 50% window offset

FIGURE 2.5 SPECTRA OF TD-PSOLA DISTORTIONS IN SINGLE FORMANT STIMULI: FM MODULATION

Figure 2.5(b) shows that the shape of the spectral envelope remains almost unaffected. The spectrum of the signal with the pitch markers set to a 50% offset in Figure 2.5(c) shows

pronounced notches in the spectral envelope, which is very discontinuous. These modifications may produce perceptible distortions; informal listening finds the timbre of the signal in 2.5(b) slightly modified, and the signal generated with the 50% analysis window offset in 2.5 (c) extremely 'hoarse'.

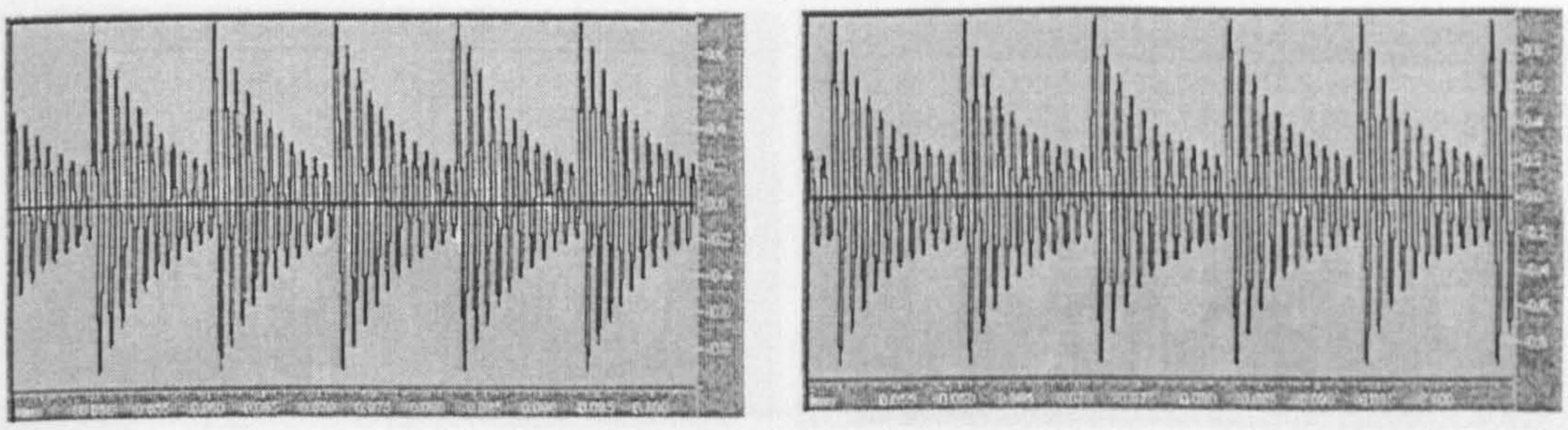
2.5.1 Thresholds for Discrimination of TD-PSOLA Modified Single Formant Stimuli

Kortekaas & Kohlrausch (1997a) performed experiments to determine the thresholds for TD-PSOLA discrimination of modified single-formant stimuli. F_{ws} was varied according to

$$F_{ws} = \frac{1}{T_a + (n/4)} \text{ Hz, where } n = \pm 1,2,3.....ms \qquad \text{Eqn 2.4}$$

Using a subjective experiment, a non-monotonic behaviour was found for TD-PSOLA discrimination. More precisely, participants were less likely to perceive any distortion for f0 manipulations when T_s was set to integer multiples of the formant frequency. For example, for a signal of F_{wa} =100Hz, f1=1000Hz, sub-threshold discrimination was found for T_s = 12, 11, 9ms (1ms=1/1000Hz) when T_a = 10ms (10ms=1/100Hz), giving F_{ws} values of 83.33, 90.9, and 111.11Hz. Similarly, for a signal of F_{wa} =250Hz, f1=500Hz, sub-threshold discrimination was found for T_s = 8, 6, 2ms (2ms=1/500) when T_a = 4ms (4ms=1/250Hz), giving F_{ws} values of 125, 166.7, and 500Hz etc.

There appeared to be a relationship between the degree of manipulation and the first formant value. Setting T_s to a multiple of the formant period results in in-phase addition of the fine structure of adjacent windows, resulting in minimal distortion of the temporal envelope. In spectral terms, a harmonic is produced which coincides with the centre of the formant frequency. The waveforms in Figure 2.6 show (a) a waveform synthesised with f0 90.90Hz, f1=1000Hz and (b) a waveform synthesised with f0 100Hz, f1=1000Hz which has been TD-PSOLA modified to give a signal with a fundamental frequency of 90.9Hz.

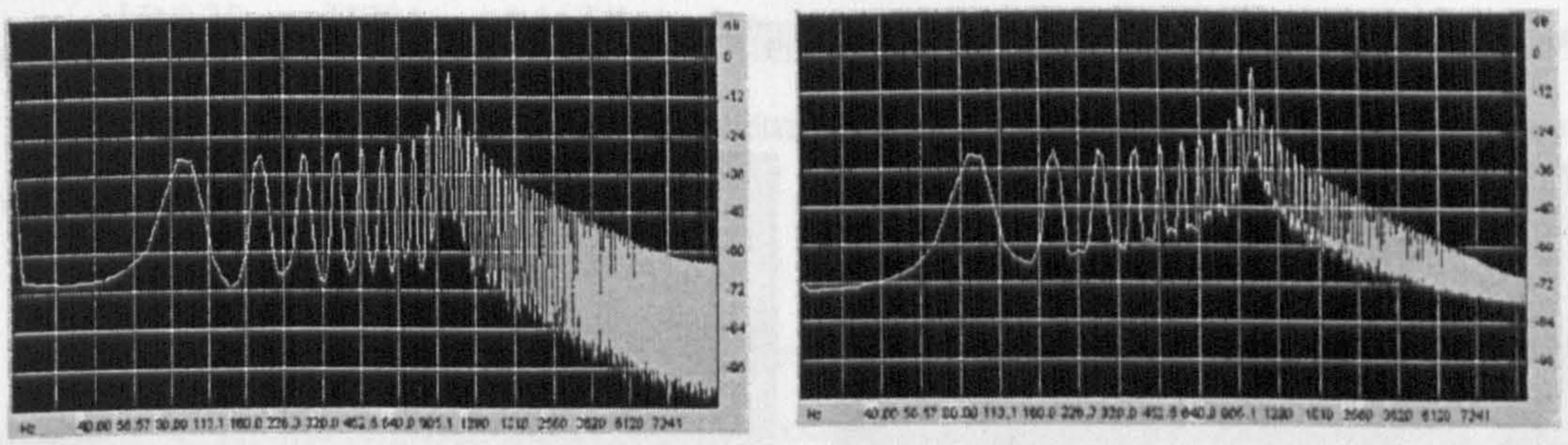


(a) signal synthesised at f0 90.9Hz

(b) TD-PSOLA modified 100Hz to 90.9Hz

FIGURE 2.6 WAVEFORMS OF SYNTHESISED AND TD-PSOLA MODIFIED VERSIONS OF 90.9HZ FUNDAMENTAL SIGNAL

Figure 2.6 shows there is little difference in the shape of the temporal envelope of the two signals. The corresponding spectra are shown in Figure 2.7. The spectral slope remains unchanged and the harmonic energy levels have suffered little attenuation.

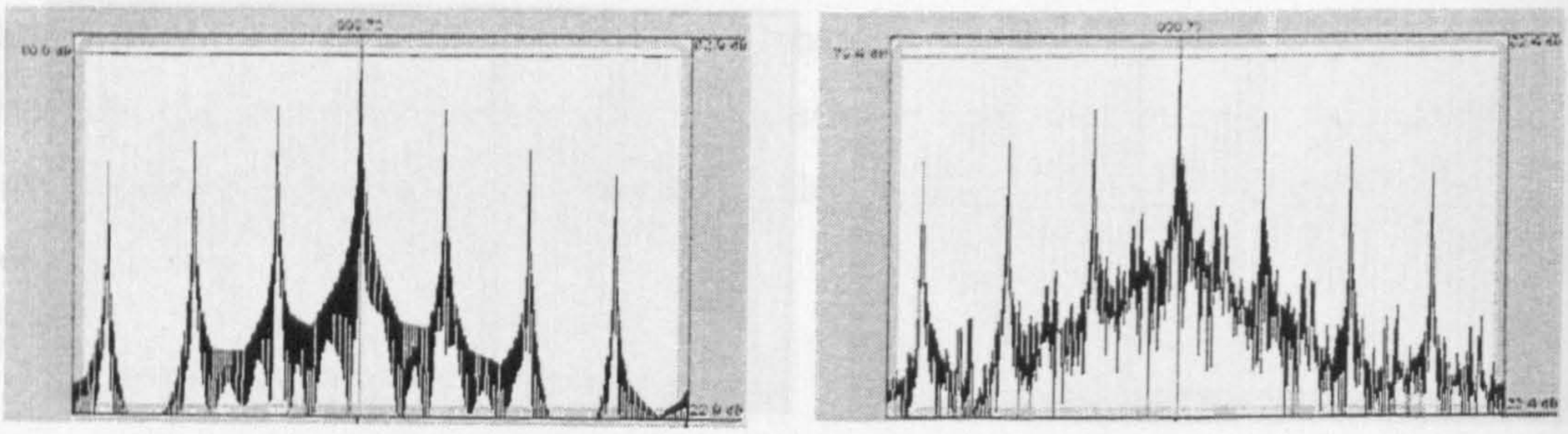


(a) signal f0 90.9Hz

(b) TD-PSOLA modified 100 to 90.9Hz

FIGURE 2.7 SPECTRA OF SYNTHESISED AND TD-PSOLA MODIFIED VERSIONS OF 90.9HZ FUNDAMENTAL SIGNAL

Figure 2.8 shows a magnified view of the first formant region. A harmonic has been produced which corresponds to the centre of the first formant at 1000Hz, resulting in minimal signal distortion. This does not appear to produce perceptible distortion; informal listening indicates it may be impossible to discriminate between the unmodified and modified signals.

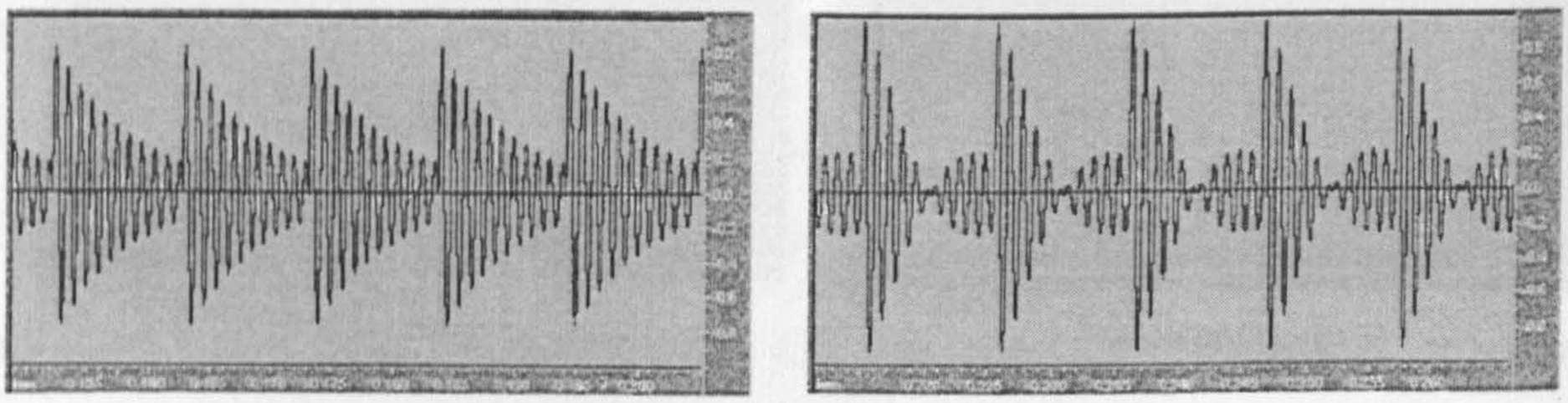


(a) unmodified signal

(b) TD-PSOLA modified signal

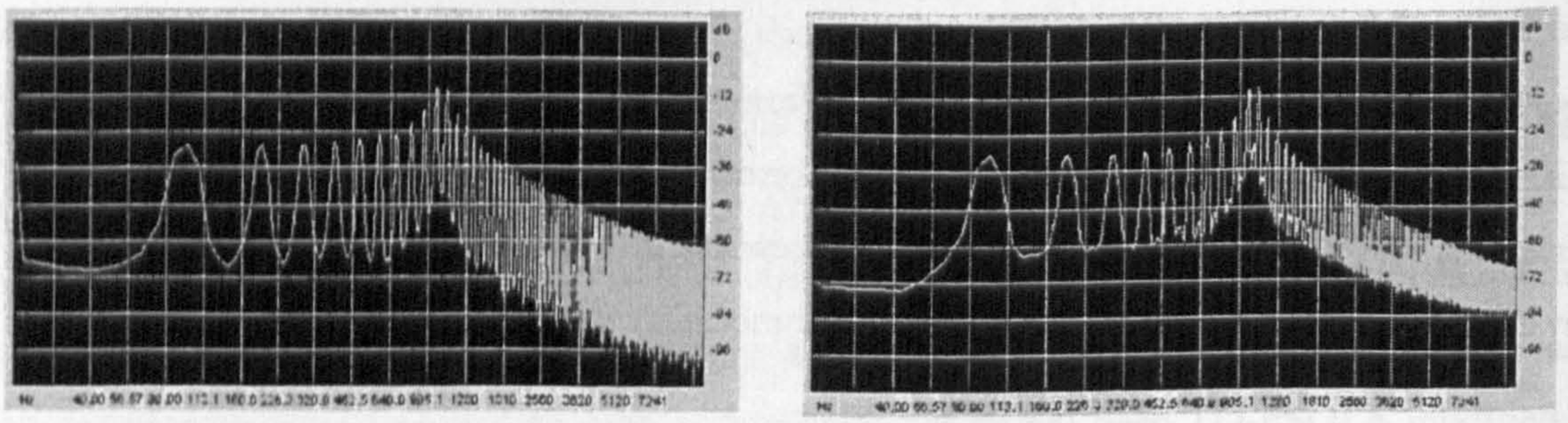
FIGURE 2.8 MAGNIFIED VIEW OF THE FIRST FORMANT REGION

The following set of diagrams in Figure 2.9 show an instance where T_s is not a multiple of the formant frequency period but has been set to 10.5ms resulting in a 95.24Hz fundamental frequency signal. The strongest harmonic is not at the centre of the first formant and a new harmonic is introduced. The diagrams on the left represent the signal synthesised at 95.24Hz, and the diagrams on the right show the TD-PSOLA signal modified from 100Hz to 95.24Hz.



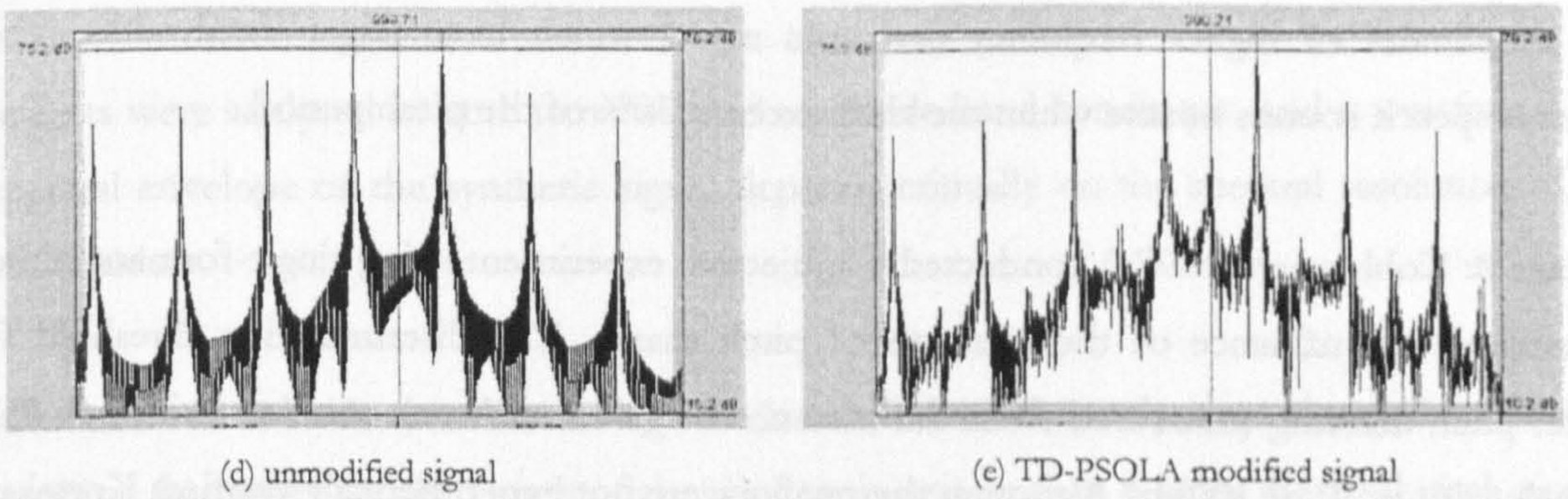
(a) synthesised f0 95.24Hz

(b) TD-PSOLA modified: 100 to 95.24Hz



(c) signal f0 95.24Hz

(d) TD-PSOLA modified to 95.24Hz



(d) unmodified signal (e) TD-PSOLA modified signal

FIGURE 2.9 WAVEFORMS, SPECTRA AND FIRST FORMANT REGION OF SYNTHESISED AND TD-PSOLA MODIFIED VERSIONS OF 95.24HZ FUNDAMENTAL SIGNAL

This distortion may be perceptible; informal listening can detect a very slight difference in timbre. The spectra show attenuation of harmonics around the first formant centre frequency of 2-3dB, although the spectral shape remains unaffected. Changes in the intensity of the spectral components that occur in the spectral region of the formant could be the cue for discrimination; formant frequency JNDs (Just Noticeable Differences) for harmonic level differences at 500, 1000 and 2000Hz formant values have been identified at 2, 2.5 and 4dB respectively (Zera *et al.*, 1993).

This section has illustrated the acoustic distortions that pitch modification with TD-PSOLA introduces into simple signals under certain conditions. Some of the distortion may be perceptible although, as Kortekaas & Kohlrausch (1997a) note, it is not known whether these conditions could be used as cues for TD-PSOLA discrimination in more complex signals, such as natural speech.

2.6 The Influence of Pitch Marker Position

The effect of incorrect pitch marker position on single formant stimuli was illustrated in Section 2.5. Moulines & Charpentier (1990) found that minimum phase distortion is achieved when pitch periods and hence windows are synchronized on the instants of glottal closure. Failure to do so results in distortions of the formant amplitudes (Moulines *et al.*, 1989). Improper synchronization

affects amplitudes of higher frequency formants more, which have larger bandwidths. They report that speech sounds hoarse when the shift exceeds 30% of the pitch period.

Kortekaas & Kohlrausch (1997a) conducted a subjective experiment using single-formant stimuli to investigate the influence of the position of pitch marks. The discrimination threshold for incorrect pitch marking (measured from the instance of glottal closure) was determined as 25% for signals with an f0 of 100Hz. Although the results were not experimentally verified, Kortekaas & Kohlrausch measured the discrimination threshold to be 10% for signals with an f0 of 250Hz. This suggests that higher fundamental frequency speech (for example female voices) may be more susceptible to small pitch marking errors.

Kortekaas & Kohlrausch (1997b) investigated the effect of pitch marker position on natural, sustained vowel stimuli. When all pitch marks were shifted by a percentage, participants reported strong 'nasality' (timbral) cues, with detection thresholds of 15%. For single pitch marker shifts, participants reported roughness as a cue for discrimination, with pitch shift thresholds as low as 2-5% of the local pitch period. For jittered pitch marking sequences (the random varying of the temporal position), roughness or unsteadyness were reported as cues, with thresholds of 0.5-1% jitter. The conclusion was that constancy of pitch marking is more important than accuracy for introducing minimal distortion.

2.7 The Influence of Analysis Window Size and Type

Moulines & Charpentier (1990) say that a PSOLA algorithm should modify the periodic part of the signal without affecting the spectrum of the stochastic component. They investigated this using an idealized mathematical model of a stationary voiced sound, consisting of a deterministic periodic signal and a stochastic component. The periodic signal represents the component in voiced speech that is the same from one pitch cycle to another, and the stochastic component models the variations from cycle to cycle that occurs during natural speech due to irregularities in the vocal cords or turbulent airflow from the lungs. The stochastic component may be dominant in certain speech, such as voiced fricatives, or during certain phonation types, such as 'breathy' speech.

The effects were analysed in both narrow-band and wide-band conditions, and it was found that the spectral envelope of the synthetic signal depends critically on the spectral resolution of the analysis window.

Narrow-band refers to the condition when the bandwidth of the analysis window is less than the fundamental frequency i.e. the length of the window is greater than 4 times the local pitch period. Considering the deterministic part of the signal, TD-PSOLA pitch modifications may cause attenuation of certain pitch harmonics due to the difference between the inherent periodicity of the Short Term signals and the synthesised pitch. A pitch harmonic remains unaffected if its frequency corresponds exactly to a pitch harmonic of the original signal. The amplitude of the synthesis pitch harmonics is affected more as it departs from the original pitch harmonics. The worst case occurs when the synthesis harmonic falls between two adjacent original pitch harmonics; the synthesis harmonic amplitude will be almost zero. This was illustrated in Figure 2.9(b). Moulines & Charpentier (1990) described this to be perceived as reverberation.

TD-PSOLA pitch-modification also affects the stochastic part of a signal under narrow-band conditions; it is no longer white noise. The stochastic signal is altered into a "pseudo-periodic structure" (resembling the output of a comb filter). Moulines & Charpentier (1990) described this structure as tonal noise, like whistling. Charpentier & Moulines (1989) and Larreur *et al.* (1989) state that narrow-band conditions are inappropriate for TD-PSOLA manipulations.

Wide-band conditions may be defined as situations where the bandwidth of the analysis window is greater than f0 i.e. the length is less than 2 pitch periods. Deterministic signals suffer as the Short Term spectra appear as a smoothed estimate of the true spectrum. The bandwidth is greater than the frequency spacing between pitch harmonics so the window cannot resolve the individual pitch harmonics. The bandwidth of a formant is usually much less than the bandwidth of the analysis window so the estimates of formant bandwidths are not good and consequently the bandwidths of the formant resonances are broadened. This problem appears more severe for higher f0 voices as the spectral resolution of the analysis window has to be reduced to meet the wide-band analysis conditions. In a worst case, fusion of closely spaced formants is observed. This may not be perceptible most of the time as the difference limen for the perception of

formant bandwidth modification is high (40% for steady vowels and larger for continuous speech (Flanagan, 1972)). Some modification of the overlap between frequency regions dominated by harmonics and those dominated by noise-like energy may also occur.

Moulines *et al.* (1990) state that achieving maximum quality is dependent upon the size and choice of analysis window. Window length should be 2P where P is the local pitch period and the window type should adhere to the desired properties of any spectral analysis window (Harris, 1978). Hanning windows are a popular choice although Kawai *et al.* (1994) advocate the use of a Tukey window that has a flat portion in its centre. This is reported to be more successful than a Hanning window when reducing pitch.

2.8 Extent of Manipulation

Much research reports that pitch and duration modifications with TD-PSOLA can introduce unacceptable degradation into speech. Laroche *et al.* (1993) say PSOLA allows high quality pitch and time modifications for moderate transformations. Breen (1998) and van Santen (1997) also report that the resulting speech quality is dependent upon the size of the modification performed and Black & Campbell (1995) found that using a corpus-based system where less signal processing was required introduced less perceptible distortion.

Some research has been performed to quantify the acceptable extent of manipulation, before distortion is perceived. Moulines & Charpentier (1990) reported that acoustical distortions are negligible for moderate duration factors. The slowing of unvoiced speech by factors greater than 2 introduces 'a short term correlation' that is perceived as tonal noise or buzzyness, due to the repetition of the ST-signals. This may be minimised in certain cases by reversing every repeated ST-signal.

Kawai *et al.* (1994) determined an "allowable modification ratio" at the word-level for pitch and duration manipulations without "significant loss of naturalness". These ratios are 55% to 149% for pitch modifications, and 45% to 122% for duration modifications, tested over a range of 0.4 – 2.0 modification factors in steps of 0.2. Significant loss of naturalness was judged to correspond

to a MOS score of less than 4/5. Hirokawa & Hakoda (1990) for sentence level stimuli gave values of f0 modification of 80% to 125% for a 75% acceptability rate of 13 listeners. Blouin & Bagshaw (2000) evaluated French vowels embedded in a CVCV (C=Consonant, V=Vowel) structure over 50% to 200% in 25% steps for both duration and pitch modifications. They found that greater modification appeared to cause greater distortion for averaged values of all vowels and report similar "acceptable" values to Kawai *et al.* (1994). Kortekaas & Kohlrausch (1997a) say that although not experimentally verified, TD-PSOLA discrimination for f0 modification of pure tone stimuli may be as low as 2%.

Breen (1998) notes that the type of modification affects perceived distortion levels; TD-PSOLA suffers when large abrupt pitch changes are necessary. Blouin & Bagshaw (2000) found the worst-case scenario occurred when static pitch contours, as opposed to dynamic contours, were imposed on speech.

Table 2.1 summarises the extent of 'acceptable' modifications.

Researcher	**Stimuli**	**Max F0 modification**	**Max duration modification**
Breen (1998)	-	<1 octave (pitch doubling)	<twice original length
Hirokawa & Hakoda (1990)	Sentence-level	80% - 125%	-
Kawai *et al.* (1994)	Word-level	55% - 149%	45% - 122%
Moulines & Charpentier (1990)	Mathematical model	-	Slowing of unvoiced speech < factor of 2
Kortekaas & Kohlrausch (1997a)	Pure tones	2%	-
Blouin & Bagshaw (2000)	CVCVC French vowels	States similar values to Kawai *et al.* (1994)	States similar values to Kawai *et al.* (1994)
Donovan & Woodland (1999)	-	Good for factor of 1.2, moderate for 0.5 – 2.0	Good for factor of 0.5 – 1.5

Table 2.1 Summary of Acceptable Extent of TD-PSOLA Modifications

The following sub-sections examine the effects of positive versus negative pitch modifications, and the original fundamental and formant frequencies on the amount of perceived distortion introduced into TD-PSOLA modified stimuli.

2.8.1 Positive versus Negative Pitch Shifts

Kortekaas & Kohlrausch (1997a) found for that for pure tones, participants were less able to discriminate negative pitch shifted stimuli. Kawai *et al.* (1994) determined the allowable range of f0 modification to be larger when pitch is decreased rather than increased, and the allowable range of duration modification to be much larger for compression than stretching. Blouin & Bagshaw (2000) found negative and positive manipulations produced similar amounts of perceptible distortion.

2.8.2 Original Fundamental Frequency and First Formant Frequencies

Kortekaas & Kohlrausch (1997a) reported discrimination ceiling effects for the 250Hz single formant stimuli, suggesting that higher f0 voices may suffer more. Blouin & Bagshaw (2000) tested both high-tone and low-tone starting frequency stimuli and found high-tone frequency French vowels suffered more perceptible distortion for duration and especially pitch modifications.

Kortekaas & Kohlrausch (1997a) also noted that formant roving may explain the success of TD-PSOLA; formant roving lessened distortion discrimination especially for lower f0 and/or low f1 values although stimuli of higher f0 (e.g. 250Hz) or higher first formant (e.g. 2000Hz) were not greatly affected by roving.

2.9 Speech Type

Breen (1998) states that TD-PSOLA cannot cope well for breathy or creaky voices that have a long or short open phase in the larynx cycle. Also, it does not perform duration modification satisfactorily for unvoiced sounds such as fricatives, or complex sounds such as affricates and plosives.

Moulines & Charpentier (1990) say slight tonal noise or buzzyness may be perceptible for voiced fricatives when raising the pitch since this involves time-scale modifications. The buzzyness is due to local periodicity caused by the repetition of Short-Term (ST) signals. This may be avoided for unvoiced speech by reversal of some ST-signals; voiced fricatives suffer as their spectrum may combine both voiced and unvoiced regions, making ST reversal impossible due to the voiced components.

Blouin & Bagshaw (2000) evaluated the effect of TD-PSOLA on French vowels and showed that the vowels may be grouped according to their first formant frequency (f1). Those that had formants less than 250Hz, such as front vowels, were perceived to be less distorted overall. Nasal vowels suffered more than oral vowels when lowering f0.

2.10 Analysis of Previous Research

This research concentrates on pitch modification as this also inherently requires duration modification; increasing pitch requires repetition of ST-signals, and decreasing pitch requires deletion of ST-signals. The existing research into the effect of the TD-PSOLA algorithm when used for pitch modification raises several issues. These are discussed below:

1. Extent of manipulation. Various 'acceptable' levels for pitch modification have been reported, ranging from ±2% to ~50%. The discrepancy between these values may be dependant on the individual implementation of TD-PSOLA, stimulus choice and definition of 'acceptable'. Kortekaas & Kohlrausch (1997a) are able to discriminate pure tone stimuli that have undergone modifications of as low as 2% of the original f0. The fact that TD-PSOLA modified stimuli can be discriminated from unmodified stimuli may be deemed unacceptable. At larger modifications, TD-PSOLA may cause changes in voice quality leading to unnaturalness, which may also be defined as 'unacceptable'.
2. Increasing versus decreasing pitch shifts. Kawai *et al.* (1994) and Kortekaas & Kohlrausch (1997a) state that pitch modification performs better when decreasing rather than increasing pitch, although Blouin & Bagshaw (2000) report similar effects in either

direction. This may be dependent on the implementation of the algorithm or on the range of manipulation under investigation.

3. Type of modification. Breen (1998) found TD-PSOLA cannot cope well with rapid pitch changes, and Blouin & Bagshaw (2000) found that imposing static pitch contours onto speech was more problematic in terms of increased perception of distortion, than imposing dynamic contours.
4. Fundamental frequency. Kortekaas & Kohlrausch (1997a) report that higher f0 single formant stimuli suffer more perceived distortion than lower f0 stimuli. Blouin & Bagshaw (2000) and Moulines & Charpentier (1990) both state that stimuli with higher starting frequencies suffer more. Conclusions are drawn that female voices may suffer more than male voices.
5. F_{wa}, F_{ws} and f1 relationship. Kortekaas & Kohlrausch (1997a) found a relationship between the original f0, F_{wa}, the target f0, F_{ws}, and the first formant, f1, in single formant stimuli. Basic distortions were illustrated such as harmonic attenuation, formant broadening and the influence of the position of the new synthesis harmonic at f1. Sommers & Kewley-Port (1996) state that level changes of the three harmonic closest to the formant frequency are most important. Although these distortions were visible in the spectra in Figure 2.8 (b), informal listening found the change in quality between the modified and unmodified signals to be perceptually minor. As Kortekaas & Kohlrausch note, it is uncertain as to whether they will be perceptible at all during more complex signals such as natural speech. They also determined that single formant stimuli with unresolved harmonics (e.g. f0 200Hz, f1 500Hz as opposed to f0 200Hz, f1 1000Hz) performed better. This may not be an issue for natural speech, as it is highly dynamic in nature and formants do not often remain steady, making such situations improbable for any appreciable duration.

Extant research also suggests several situations or parameters that may cause distortion to occur. These are discussed below.

1. Pitch marking. Moulines & Charpentier (1990) found that pitch shifts of 30% may cause speech to sound 'hoarse' and distort formant amplitudes. This was especially so for high frequency formants which have larger bandwidths. Kortekaas & Kohlrausch (1997b) report

the thresholds as 25% for stimuli with an f0 of 100Hz, and 10% for stimuli with an f0 of 250Hz stimuli. When all pitch marks are shifted by a percentage, participants report 'nasality' as a timbral cue. For single shifts, roughness was reported as a cue, with thresholds of 2-5% shift. Jittered pitch marker sequences (the random variation of temporal positions of glottal pulses) cause roughness or unsteadyness to be reported, with thresholds of 0.5-1% shift. Overall, constancy of pitch marking was found to be more important than accuracy.

2. Phonation type. Different phonation types are caused by changes in the excitation pulse or glottal waveform by varying the tension and position of the vocal cords. Breathy voice is characterized by a longer fall phase, a more symmetrical pulse and a lower f0. Breathy voice is also dominated by a stochastic component. Creaky voice (or vocal fry) is characterized by a short rise time, irregular pulses and a very low f0. Both types of phonation have been reported as problematic for TD-PSOLA.
3. Speech type. The phoneme type was reported to have a marked effect on the success of TD-PSOLA. Aspects of various speech types are discussed below.

The following points outline the major characteristics of speech types grouped according to their manner of articulation:

1. Duration. Stops have the shortest duration and diphthongs the longest. Checked vowels are shorter than diphthongs which have longer steady states, and tense vowels have a longer duration than lax vowels. Vowels have longer durations than sonorants, affricatives are longer than stops, unvoiced obstruents are longer than voiced, and unvoiced fricatives are much longer than voiced fricatives.
2. Intensity. The intensity of speech sounds varies greatly e.g. the /A:/ in "calm" is 700 times more powerful than /T/ in "think". In order, vowels are loudest, followed by liquids, nasals, affricatives, fricatives, and finally stops. Tense vowels have a higher intensity than lax vowels.
3. First formant frequency. F1 is a function of tongue height determined by the size of the back cavity behind the tongue hump; front tongue elevation produces a larger volume in the back cavity associated with lower f1 frequencies (e.g. /i:/ in "see"), whereas back tongue elevation decreases this volume and raises f1 frequencies (e.g. /Q/ in "top").

4. Speech sound category. There are three types of speech sound category: frication, plosion and voicing. All speech sounds are made up from one category or from more than one category, either simultaneously or in sequence.

Fricatives are produced by turbulent air streams causing noise-like signals. The strident fricatives /s/ and /S/ have higher amplitudes than non-strident fricatives such as /f/ and /T/.

Plosives consist of a silence and then a release burst. The release burst shows a split second of turbulence resembling the fricative with the same place of articulation e.g. a /t/ looks like a brief /s/. As the vocal tract moves from the consonant to the position for the following vowel, there are brief influences on the formants of the vowel. There would also be influences on the formants at the end of preceding vowels. Voiced plosives are periodic whilst the vocal tract is closed and the release burst is less prominent.

The voiceless affricative /tS/ begins as plosives with a silence, but when the closure in the vocal tract is released a fricative sound is produced. The voiced affricative /dZ/ has an initial part that is periodic and a second part that is a mixture of periodic and aperiodic signals. Moulines & Charpentier (1990) say there can be some modification of overlap between frequency regions dominated by harmonics and those dominated by noise-like energy, which may cause problems for the TD-PSOLA algorithm and produce distortion.

Vowels and voiced consonants contain voiced speech. Phonemes such as voiced fricatives may combine both voiced and unvoiced parts simultaneously. Vowels may often consist of both voiced and unvoiced speech due to the occurrence of large amounts of glottal induced noise that may be superimposed on their basic periodic waveform.

2.11 Summary

TD-PSOLA is a highly successful and popular algorithm used to modify the pitch of speech segments when required to create new utterances during the speech synthesis process. Chapter 2

initially described the operation of the TD-PSOLA algorithm, which involves the three stages of analysis, modification and synthesis. The Praat software implementation of the algorithm, which is used during this research, was then presented.

It is well documented that the main drawback is the occurrence of perceptible distortion when TD-PSOLA is applied to speech which, although less than for parametric speech models, is nonetheless seen as detrimental. The research concentrates on the effect of pitch modification using TD-PSOLA, as pitch modification inherently requires duration modification. Some of the basic signal distortions introduced when the algorithm is used for pitch modification were investigated by examining its effect on pure sine waves and then on more complex single formant stimuli.

As these distortions are not always perceptible especially with more complex natural speech stimuli, extant research was reviewed and suggested some possible situations of when this perceptible distortion may occur. These possible contributions consisted of incorrect pitch marking, the influence of the analysis window size and type, the extent of manipulation, and the speech type. The review of existing literature highlighted some issues requiring further investigation, which may inform the design of a framework for a speech corpus approach used in conjunction with TD-PSOLA that results in an output with less perceptible buzzyness. Part of the framework consists of the design of a corpus containing optimised segments which when TD-PSOLA is applied, result in minimal introduction of perceptible distortion. Issues which may affect the corpus design include the various 'acceptable' levels of pitch modification cited, ranging from 2 to ~50% of the original pitch, the issue of whether increasing or decreasing pitch modifications is preferable, the type of modification (static versus dynamic and abruptness of f0 changes), and a possible relationship between the starting f0, the target f0 and first formant value, and the speech type. The issue of choosing a speaker to record the speech segments for the corpus may be dependant on the effect of the starting f0 of speech (female voices may suffer more), the phonation type of the speaker, and possible noise on the periodic parts of the speech. Some parameters of the various speech types were identified, when the phonemes were grouped according to their manner of articulation, of duration, intensity, first formant values, and speech sound category (frication, plosion, voiced). These parameters may lead to increased distortion.

Finally, some situations where anomalous, or increased amounts of distortion may occur were identified. These consisted of incorrect pitch marking, phonation type, analysis window size, and extraneous noise on periodic waveforms. If these were found to be possible contributions to the occurrence of distortion, segments exhibiting such aspects could be removed from the corpus.

To address these issues, a series of experiments will be carried out to investigate the effect of the TD-PSOLA algorithm on natural speech. The results of the experiments will then be used to suggest a design for the speech corpus. The following chapter reviews existing speech assessment techniques and experimental practice to aid the design of these investigative experiments.

Chapter 3. Evaluation of Synthetic Speech Output

3.1 Introduction

Rapid progress has been made in the field of speech synthesis, but it is still possible to identify the output of any commercial speech synthesiser as non-human sounding. Although synthetic speech may be perfectly intelligible, listeners often report synthetic voices as irritating (Cowley & Jones, 1993) and find more natural speech easier to process (Hawkins *et al.*, 2000). While this remains the case, it is important to evaluate synthetic speech effectively to highlight areas where further improvements are necessary.

There are two main forms of assessment: diagnostic tests and overall performance assessments. Diagnostic tests evaluate the synthesiser performance at individual levels. Such levels may be the intelligibility or naturalness of the speech, and subsets of these such as speaker style, emotion, accentuation etc. Overall or global performance assessments provide judgements on the quality or fitness of a synthesiser for a certain application. They may be used to evaluate more than one quality aspect at a time. Overall performance tests often involve field-testing, using real situations and listeners. For example the synthesiser may need to produce intelligible output in a noisy environment or over a low communication bandwidth medium, such as for telephony applications.

Major applications of synthesis systems are to provide user-friendly interfaces between human and machine. It is therefore important that the synthetic output is easy to listen to, and as intelligible and natural as possible to prevent listener fatigue (Morton, 1991). It is necessary to identify the application of the synthesiser to decide upon the most relevant aspects to be tested. For example, intelligibility would be of paramount importance for applications such as speaking timetables or clocks. For applications such as electronic mail or web page readers, it may be additionally desirable to generate expressive or emotional speech, with natural prosody.

There are two approaches to evaluating the quality of either parts of, or the whole of, a speech synthesis system: objective and subjective evaluation. Objective approaches attempt to measure physical features of a speech signal that are related to the quality of the speech. Subjective measures rely on human perception to judge speech quality. The branch of science concerned with the perception of sound is known as psychoacoustics.

Speech is one of the most complex signals in existence (Deketelaere *et al.*, 2001); it does not lend itself easily to acoustic measures of quality. Due to the difficulty in establishing a link between physical factors of a speech signal and the resulting perceptual quality required for objective assessments, this research relies on evaluations based upon the perception of speech quality by human listeners.

The drawback to perceptual-based, qualitative measures is that they require careful design. There are many factors that may influence the results of an evaluation and it is imperative to eliminate factors that are not part of the desired evaluation. Factors of influence may be the test methods and test material chosen, the criteria for evaluation, the selection of participants, and the intention of the assessment (Jekosch & Pols, 1994). Each of these issues will be discussed in the following sections.

3.2 Existing Test Procedures

Due to the nature of speech, and the difficulties in quantifying terms such as quality, intelligibility, and naturalness, it is important to determine appropriate criteria for assessment and develop appropriate test procedures. There are currently no existing standards for these, although there are many popular individual tests in existence, with various accuracy and validity (Jekosch 1993, van Bezooijen & van Heuven, 1997).

Some of the existing tests are listed. These can be used for intelligibility, or overall quality evaluations. Such features can be assessed at various levels such as segmental or phoneme, word and sentence-level.

For segmental intelligibility:

- the Diagnostic Rhyme Test (DRT)
- the Modified Rhyme Test (MRT)
- the CLuster IDentification tests (CLID) (consonant-vowel-consonant (CVC) or vowel-consonant-vowel (VCV)).

For sentence-level intelligibility:

- the Semantically Unpredictable Sentences test (SUS)

For Overall Quality:

- the Paired Comparison (PC)
- the Mean Opinion Score (MOS) tests
- the ITU-T Overall Quality Test

Aspects of speech such as intelligibility and naturalness are not completely independent of one another, and it is almost impossible to predict which test procedure will give the more valid results. Several reports have been compiled containing recommendations on assessing speech using these tests, such as the ESPRIT 'SAM' project (Pols & SAM-Partners, 1992), the EAGLES project (Expert Advisory Group on Language Engineering Standards) (Gibbon *et al.*, 1997), and the JEIDA report (Japan Electronic Industry Development Association) (Jeida, 1995). Additionally a PC based package 'SOAP' (Speech Output Assessment Package) (Howard-Jones *et al.*, 1992a) has been implemented to simplify and standardise test procedures.

3.2.1 Segmental Intelligibility Tests

The Segmental Intelligibility Tests evaluate the intelligibility of single phonemes only. The more popular segmental intelligibility tests are the rhyme tests (DRT and MRT) and the CLID test.

3.2.1.1 The Diagnostic Rhyme Test (DRT) (Voiers, 1983)

The DRT is a widely used phonemic intelligibility test. Stimuli are constructed as meaningful, monosyllabic CVC (consonant-vowel-consonant) words. Participants are acoustically presented with a pair of words and asked to identify which is which from a list of two words (closed

response). Each stimulus differs only in variation of a single acoustic feature of the initial consonant, for example, "dune" and "tune". The total number of stimuli presented to the participant is 96 word pairs.

Six contrasting speech sounds are represented; voiced/unvoiced, nasal/oral, sustained/ interrupted, sibilant/non-sibilant, graveness/acute, and compactness/diffuse. For example, the stimuli "veal" and "feel", for voiced/unvoiced speech. Each of these contrasts is represented 32 times, with a combination of 8 vowels.

The total time for such a test averages 15 minutes, with 3 seconds between each stimulus presentation.

Intelligibility is expressed simply as the percentage of correctly identified initial consonants overall and for each contrast.

3.2.1.2 The Modified Rhyme Test (MRT) (House et al, 1965)

The MRT is also a phonemic intelligibility test, although less widely used than the DRT, perhaps due to the larger stimuli set required by the MRT. Stimuli are meaningful CVC words that can differ in both their initial or final consonants, but not both simultaneously. The MRT is a closed response test, with six possible alternatives. 50 sets, each consisting of 6 stimuli, are presented during the test. The first 25 sets have different initial consonants and the second have different final consonants. An example of a response list may be "peas, peak, peal, peace, peach, and peat". Participants are played one word at a time, and asked to identify it in such a list of six words.

The average time for such a test is 25 minutes, with an interval of 4 seconds between each stimulus.

Intelligibility is expressed as the percentage of correctly identified initial and final consonants, or as an overall consonant correctness.

3.2.1.3 Advantages and Disadvantages of the DRT and MRT

Both the DRT and MRT tests allow the phonemic identity of those speech sounds that are generated less successfully by the synthesiser to be easily identified. Both these test procedures have proven reliability (Jekosch 1993). A minimal set of stimuli is necessary, giving short test runs. Naïve listeners need no training, as the test employs meaningful stimuli, allowing listeners to respond with familiar spelling. Reliable results can be gathered with a small number of participants (10 - 20) making these tests easy and economic to implement (Jekosch, 1993). The measure of intelligibility is simply the number of correctly identified stimuli. Confusion matrices can provide data on misidentified phonemes, and how they are confused with one another, to highlight aspects of the synthesiser requiring attention or further development. For the DRT, large amounts of previous test data are readily available.

The drawbacks are that only the initial consonant for the DRT and initial and final consonants for the MRT are tested. Additionally, there are only a limited number of meaningful, or semantically bearing, CVC combinations that fulfil the rhyme conditions. This means that not all possible confusions between phonemes can be evaluated. Additionally, they do not address any vowel intelligibility or prosodic features so they are not suitable for overall quality evaluation.

3.2.1.4 The CLuster IDentification Test (CLID) (Jekosch, 1992)

To overcome the problems of using a finite list of meaningful words, nonsense strings are also presented in this test. The stimuli consist of sequences of one or more consonants (consonant clusters) and sequences of one or more vowels (vowel clusters). This overcomes the limitations of testing only initial and final single consonants, when 40% of all monosyllabic English words begin, and 60% end, with consonant clusters (Speigel *et al.*, 1990). Example stimuli are 'storp' (CCVCC) and 'fast' (CVCC). The stimuli may also be representative of the frequency of occurrence of phonemes in the language using Phonetically Balanced (PB) stimuli lists.

An open response modality is employed. Participants are given the overall list of sounds they may hear during the test. They respond using normal spelling or simple notation. Jekosch (1992)

presented 900 stimuli, giving an overall test time of 2 hours, however, a smaller subset can be used. The SOAP software (Howard-Jones *et al.*, 1992a) implements this test, allowing stimuli to be automatically generated and scored, in terms of percentage correctness of initial, medial and final clusters separately or as whole constructs. Additionally, confusion matrices can be generated to investigate confusion between individual phonemes.

The test is relatively fast. A small number of participants is required, 4 being typical. CVC tests have been recommended by the CCITT (CCITT, 1993), whereas a VCV version of the test is proposed as a standard for European languages by the SAM project (Pols & SAM-Partners, 1992). One disadvantage of using both meaningful and nonsense stimuli is that participants expect them to make sense and naïve listeners often need to adjust to this test.

3.2.2 Sentence-level Intelligibility

The drawback to testing intelligibility at sentence-level is that, for the sake of naturalness, prosody is included. Tests at sentence-level then become more like perception of normal communication and it becomes difficult to restrict the results to intelligibility only. If incorrect prosody occurs, it may affect the perceived content of the sentence.

3.2.2.1 Semantically Unpredictable Sentences (Pols & SAM-Partners, 1992)

The aim of this test is to reduce contextual information present in a sentence available to the participant by using semantically unpredictable sentences. Simple grammatical sentence structures are used into which nonsense strings are inserted.

Five grammatical structures are used (examples taken from Jekosch, 1993):

Subject – verb – adverbial: "The table walked through the blue truth."

Subject – verb – direct object: "The strong way drank the day."

Adverbial – verb – direct object: "Never draw the house and the fact."

Q-word – transit. verb – subject – direct object: "How does the day love the bright word?"

Subject – verb – complex direct object: "The plane closed the fish that lived."

During the test, listeners are presented with fifty sentences, 10 of each sentence structure in a random order. The response modality is open; participants record the sentence they perceive using normal spelling.

Intelligibility is scored simply as the percentage of correctly identified sentences. Using 50 sentences, with 15-second intervals between each, the average test time is 15 minutes.

The main drawback to this test is that it may only be used as a comparative tool, rather than a diagnostic one. The intelligibility scores may not be based solely on phonemic identification but may be affected by prosody, and semantic content.

3.2.3 Overall Quality Tests

Improving quality is an on-going problem in synthetic speech development. Research questions exist as to why people would rather listen to natural speech rather than synthetic speech. It is important to extract the qualities of speech that separate natural speech from synthetic speech. Conversely, it may be desirable to identify the aspects of synthetic speech which participants find most irritating.

Overall quality tests are desirable to test a larger, subtler range of speech synthesis properties. The following tests are often used to evaluate specific aspects of speech quality or can be used to assess overall quality.

3.2.3.1 Paired Comparison (PC) (Kraft & Portele, 1995)

Participants are presented with a set consisting of two stimuli and must choose the one that fulfils the test criteria better, often referred to as a two-alternative forced choice (2AFC) test. Identical sets can be included to verify the participant reliability. Transitivity must be respected to avoid conclusions such as A is better than B which is better than C which is better than A.

3.2.3.2 MOS test (ITU-T, 1996)

Mean Opinion Score (MOS) tests rate perception of quality on a scale, such as 1 to 5, where 1 represents poor and 5 represents excellent. Participants simply rate individual aspects or the overall quality of the speech on this scale. This method provides information about which aspects of the speech need attention. A DMOS (Degradation MOS) may also be used to measure the magnitude of disturbances.

The main drawback of the MOS test is that a participant's score has meaning only relative to the scores from other respondents and the data should therefore be treated as ordinal rather than interval.

Both the paired comparison and the MOS test can be used to evaluate intelligibility.

3.2.3.3 ITU-T Overall Quality Test (ITU-T, 1994)

This test evaluates aspects such as acceptability, overall impression, listening effort, comprehension problems, articulation, pronunciation, speaking rate and voice pleasantness. Four synthesis systems can be rated in such a test.

3.3 Test Conditions

Test conditions must be reported to allow experiments to be replicated. The chosen speech output device e.g. speakers, headphones or telephone, depends upon the application of the synthesiser. The acoustic environment, such as an office, train station etc. must be recorded and the type of background noise, if any, noted. The synthesiser specifications such as voice details, sampling frequency, and synthesis method are also required.

3.4 Participants

Human participants are inconsistent in their judgements or task performance (Eagles, 1996), hence it is important to use at least 5 participants or more (Jeida, 1995). The number of

participants is also dependent on the statistical tests that will be employed to analyse the data to give the test adequate power (Clark-Carter, 1999).

The participants' personality, motivation, education, attitudes and expectations (Jekosch, 1993) can all affect the results. A listener's conclusions may be affected by variables such as speaker dialect, gender etc. It is advantageous to choose participants from both genders and from various regions to ensure that individual preferences for either male or female voices, or regional accents, do not greatly affect the evaluation.

Howard-Jones *et al.* (1992a, 1992b) investigated the reliability of participants with age. Little evidence was found to restrict age participation in subjective experiments. It is imperative though to establish that all participants of any age, participating in listening tests, are of normal hearing ability. Decibel-level perception tests can be administered to achieve this.

Howard-Jones *et al.* (1992a, 1992b) also investigated variations in results when using experts, or participants with previous experience of synthetic speech, and naïve listeners. Experts were found to score higher on intelligibility tests than inexperienced participants. It is therefore important to identify the end user of the synthetic speech (expert or naïve listener) and reflect this in the choice of participant.

Eagles (1996) recommend that the same participant should not be used more than once, due to learning effects, whereby the speech becomes more acceptable, and to use only participants speaking the same language as the test stimuli when performing diagnostic tests.

3.5 Experimental Procedure

The order of presentation of the stimuli to the participants should be randomised for each test run. The results may be influenced by a learning effect whereby synthetic speech becomes easier to understand and more acceptable in terms of prosodic patterns and naturalness with familiarity (Neovius & Raghavendra, 1993). Additionally, concentration may lessen over time, depending on the participants' motivation and length of experiment.

Some participants may have experience in listening to synthetic speech and carrying out qualitative testing. Unfamiliar participants should be familiarised with the testing method and the definitions of the criteria as applied to synthetic speech before the test. Due to this, preliminary training should be carried out to familiarise inexperienced listeners with synthetic speech, the terminology or criteria, and the test procedures. To prevent participants becoming familiar with synthetic speech, human speech can be used in training where possible (Jeida, 1995).

Where a MOS scale rating is used, examples of the range of speech quality should be given before the actual test, to give participants a baseline for evaluation. Examples should be few to minimise the learning effect.

The stimuli are normally presented with an interval of approximately 5 seconds between test speeches. Each stimulus should be presented only once, or with one possible repetition (van Santen, 1993), and due to listener fatigue, the length of session should be 20 minutes maximum (ITU-T, 1996).

3.6 Summary

There are many procedures in use for the evaluation of both intelligibility and quality of synthetic speech. There is currently no standard, although this problem has been recognized in the speech synthesis community and various ideas have been proposed. Subjective assessment is a vast area of research in itself, with many variables that must be taken into consideration.

Perception of quality of speech is a complex relationship of many factors e.g. naturalness and intelligibility; speech may sound natural but not be understandable or it may be smooth with no audible jumps and have correct prosody, but be very buzzy. In applications, it is important to identify which aspects are most critical.

The principal aim of this research is to investigate the occurrence of perceived distortion when speech is pitch manipulated using the TD-PSOLA algorithm. This research is not concerned with intelligibility; research has already ascertained that concatenative synthesis systems using the TD-PSOLA algorithm produce extremely intelligible speech. The research is concerned with the

evaluation of quality, by detecting perceived distortion in the form of buzzyness. The majority of the following five investigative experiments make use of MOS tests, which are reported to be sensitive to subtle differences (Blouin & Bagshaw, 2000). MOS tests were chosen as they allow differences in amounts of perceived buzzyness to be investigated for a relatively small data set. The use of MOS tests rather than Paired Comparison tests are discussed in more detail in Section 4.2.2.2. MOS scale data will be treated as ordinal when testing for statistical significance as participants' scores are only meaningful relative to each other.

The fifth experiment is concerned with determining whether any distortion is perceived at all, not with the magnitude of this distortion, therefore a yes-no categorical response modality will be employed.

Guidelines for the documenting of test conditions, choice of participants, and experimental procedure outlined in this chapter were adhered to in the following experiments wherever possible. To allow the experiments to be replicated, the test conditions will be reported for each. These consist of the output device, the acoustic environment in which the experiment takes place, background noise levels, the PC running the experiment, and the speech specification (voice identity, gender of speaker, sampling frequency, type of stimuli and the algorithm under investigation).

The participants are all university staff or students, and are relatively small in number, due to the constraints of expense and availability. This restricted sample population is acknowledged in each experiment as it may bias the results, although no aspects of the occupation (such as working in a noisy environment for certain occupations) are assumed to have any effect on hearing abilities. No restriction on age participation was imposed, but all will be asked if, to the best of their knowledge, they have normal hearing ability. All participants were known to the author, and so it was assumed that they may admit more readily to any known hearing defects when asked. It is expected that the majority will be naïve listeners with little or no experience of taking part in synthetic speech experiments, and so full training and familiarisation with the criteria for evaluation will be provided before each test run commences. A set of instructions for each experiment will be prepared, giving a clear description of the procedure and criteria for evaluation.

The same participants may be required to take part in more than one experiment. As these experiments were conducted with at least a month between each, the learning effect encountered when listening to synthetic speech, was not expected to affect the results.

The order of presentation of the stimuli will be randomised for each test run to reduce the influence of the learning effect and potential loss of concentration as the experiment progresses. Due to the problems of listener fatigue, no test run will be longer than twenty minutes. Stimuli will be presented at a rate controlled by the participant and each stimulus may be presented once only, to minimise the learning effect.

A MOS scale rating will be used, and examples of the possible range of speech quality to be encountered will be provided before the test. A minimal number of examples will be used in an attempt to avoid any learning effects.

Chapter 4. Investigative Experiments

4.1 Introduction

Concatenative synthesis makes use of pre-recorded units stored in an inventory or speech corpus. These units may not have the desired pitch and duration when used to create new, arbitrary sentences. TD-PSOLA is an efficient, successful and widely used algorithm capable of modifying the pitch and duration of such speech units. However, these modifications can degrade the signal by the introduction of unwanted perceptible distortion in the form of buzzyness. This chapter describes the experiments that were undertaken to investigate the behaviour of the TD-PSOLA algorithm in terms of the occurrence of this buzzyness. The results of these experiments will be used to guide the development of a framework for concatenative synthesis, which is capable of producing an output with reduced distortion. The framework consists of a speech corpus design for use with TD-PSOLA, a signal processing distortion measure to enable segments to be selected from the corpus that will result in minimal distortion, and a special selection process for especially problematic phonemes.

In a speech synthesis system using TD-PSOLA, both pitch and duration modifications of speech are necessary to produce the desired prosody. To keep the extent of this study to manageable proportions, only the pitch modification deficiencies of the algorithm were explored during the following experiments. Pitch modification was chosen as it inherently requires duration modifications; increasing pitch necessitates repetition of ST-signals (simulating increasing duration), and decreasing pitch involves removal of ST-signals (simulating decreasing duration).

In Section 2.8, previous research suggested that in general, the greater the distance a speech signal was pitch-modified, the greater were the perceptible levels of distortion. The first experiment undertaken here investigates this relationship between the degree of pitch manipulation and the resulting distortion levels in speech using vowel stimuli. The results from this experiment will be used to inform the design of the speech corpus by suggesting the maximum extent of pitch modification possible without the introduction of perceptible distortion. The identity of the individual phoneme stimuli used in the experiment appears to affect the levels of perceptible

distortion, and this is investigated post hoc. The results will be used to determine whether the content of the corpus should be balanced to take into consideration the effect of the algorithm on segment types, rather than just phonetically balanced. These results will also be used to develop a signal processing distortion measure for vowel phonemes reflecting the levels of distortion perceived for each phone class.

Conflicting views on the effect of positive pitch shifts (signal modification to a higher fundamental frequency) and negative pitch shifts (signal modification to a lower fundamental frequency) were also presented in Section 2.8; a second experiment compares the effect of the algorithm on distortion levels when speech is pitch-modified positively versus negatively. The results from this experiment will be used to determine whether segments in the speech corpus should be represented at lower or higher pitches. They may also be selected having a f0 below or above the target pitch in order to be pitch-modified in a positive or negative direction, to result in minimal perceptible distortion.

These two experiments evaluate the effect of the algorithm using stimuli consisting of isolated syllables, therefore a third experiment was designed to evaluate the effect of pitch modification on distortion at the sentence level. The results of this experiment will be analysed to determine whether the results of the previous experiments using word level stimuli may be generalised to the sentence level. This will indicate whether the data from the previous experiments, used to design the speech corpus, may provide a valid design when the corpus is used for synthesising new sentence level utterances.

A fourth experiment evaluates the response of various voices (two male and two female) to the application of TD-PSOLA to determine whether conclusions for one voice may be generalised to others. In addition, consonant stimuli are used, and the effect of the individual phoneme identity of each stimulus on perceived distortion is investigated post hoc. These results will be used to inform the design of the speech corpus and signal processing distortion measure with regard to consonant phonemes.

The results of these four experiments indicated that aspects of the original recordings of the stimuli have a large effect on the resulting perceived distortion levels. A fifth experiment explores

these issues. These aspects could then be eliminated from the speech corpus to reduce levels of perceived distortion when using TD-PSOLA to modify speech.

The following sections document the investigative experiments undertaken during this research. For each experiment, the aims and hypotheses are identified, followed by the experimental design, which justifies chosen statistical tests, documents any pilot studies undertaken, and describes stimulus preparation, procedure, and participants. The resulting data are then analysed to determine statistical significance, and finally discussions and conclusions are presented.

4.2 Experiment 1: The Effect of Pitch Manipulation using the TD-PSOLA Algorithm on Distortion Levels in Speech Sounds

Abstract

A listening test was undertaken to determine the amount of distortion, in the form of buzzyness, present in TD-PSOLA pitch-modified speech. The stimuli had been modified by various standard distances from their original pitch. Participants were presented aurally with the stimuli in a random order and asked to judge on a scale of 1 to 5 the levels of distortion present in each. Significantly greater distortion was perceived for increasing degrees of pitch manipulation, with pitch modifications as small as 1% introducing perceptible distortion into the speech signal. The effect of the individual phoneme identity of the stimuli on distortion levels was investigated post hoc.

4.2.1 Introduction

The pitch and duration of speech units are the major contributors to the prosodic aspects of speech. The TD-PSOLA algorithm is used to successfully manipulate both the pitch and duration, although it can introduce distortion into the signal (Breen, 1998). It has been reported (van Santen 1997, Breen 1998) that in general the larger the pitch modification, the more distortion that is introduced. Kortekaas & Kohlrausch (1997a) argue that although this may be true, their work with more abstract, single-formant stimuli indicates that the results may depend on a relationship between the values of the starting fundamental frequency, the target fundamental frequency and the first formant value. Various "acceptable" pitch manipulation levels have been cited in the literature ranging from 2 to ~50%; this large variation is thought to be due to the definition of "acceptable". Studies citing the larger values may use changes in voice quality as a percept of unnaturalness and unacceptability, although it could be argued that even the smallest perceivable distortion is unacceptable. The variation in opinion may also be due to a lesser extent to the individual implementation of the TD-PSOLA algorithm evaluated. This experiment aims to address these issues by determining the relationship between the degree of pitch manipulation and the resulting distortion levels in speech, and determine the minimum

perceptible level of such distortion for the Praat (Boersma & Weenink, 1999) implementation of TD-PSOLA.

4.2.2 Design

This section states the experimental hypotheses, describes considerations in the design of the experiment to control or account for variables, and documents any biases that could influence the results, to allow reasonable conclusions to be drawn.

4.2.2.1 Hypothesis

Hypothesis H_1: Increasing degrees of pitch manipulation using the TD-PSOLA algorithm introduce significantly increasing levels of perceptible distortion, in the form of buzzyness, into speech.

4.2.2.2 Structure of experiment

A listening test was designed to evaluate the amount of distortion introduced by TD-PSOLA when speech is pitch-modified. By systematically varying the degree of pitch manipulation, the behaviour of the TD-PSOLA algorithm can be described, in terms of resulting distortion levels.

Initially, the range of pitch manipulation to be evaluated was chosen. Synthesis systems have progressed from simple diphone synthesis to more sophisticated systems, such as corpus-based strategies. These approaches make use of inventories or corpora containing speech units whose fundamental frequencies are closer to the target values. Hence, the chosen range of the independent variable *degree of pitch manipulation* does not reflect the possible variation of pitch during natural speech, but has been chosen to reflect typical modifications required for corpus-based synthesis systems. The independent variable *degree of pitch manipulation* has a maximum level of 15% modification from the initial pitch. It should be noted that it is not suggested that 15% would be the maximum pitch modification required in such a speech synthesis system, but is one that would be required often; the segments are extracted form many prosodic contexts with various values of pitch, therefore potentially requiring less signal-processing to attain target values.

The method of measurement of the dependent variable *distortion* was then chosen. A possible option was to use a paired comparison test where participants would be presented with two stimuli and asked to judge the more distorted. Presentation of an unmodified stimulus paired with its TD-PSOLA modified version at one of the IV levels may lead participants to realise that the stimulus at the standard starting pitch was usually the less distorted of the pair. This could be avoided by having various starting frequencies but would involve larger amounts of stimuli. Alternatively, two identical frequency instances of the speech sound (i.e. a stimulus recorded at the target frequency and a stimulus TD-PSOLA modified to the target frequency from the starting frequency) could be provided. However, the stimuli may differ due to variations in voice and recording quality, which could affect participants' judgement and make such tests invalid. Paired comparison tests would also not facilitate the determination of differences in perceived amounts of distortion between all levels of the IV without impractically large data sets.

A Mean Opinion Score (MOS) test was adopted which made the measurement of different amounts of distortion between the IV levels possible with a manageable sized data set. The ITU-T recommendations (ITU-T, 1996) advocate the use of MOS tests when evaluating small impairments due to their sensitivity. The dependent variable *distortion* was measured on a MOS scale of 1 to 5. The stimuli were assessed using the amount of perceived distortion as the criterion for evaluation, where distortion was defined as *buzzyness*, or *electronic* sounding.

Participants rated each stimulus using a number between 1 and 5, corresponding to the following definitions:

1 no perceived distortion/ very natural

2 quite undistorted/ quite natural

3 distorted/ slightly unnatural

4 quite distorted/ quite unnatural

5 very distorted/ very unnatural

A within-subjects design was used to reduce the effect of differences in the level of response ratings between participants which would obscure the effect of the treatment; each participant rated the distortion at all levels of pitch manipulation.

A statistical test was required to compare the differences between the medians of the IV levels to determine whether increasing pitch modification may have contributed to significant increases in distortion. The design was within-subjects but as the data were ordinal, the assumptions of the parametric ANOVA were not met. Hence a within-subjects Friedman test (corrected for ties for greater accuracy) was performed on the non-parametric data. The hypothesis stated that increasing pitch manipulation would lead to increasing distortion, so a one-tailed test was carried out.

When the number of IV levels is greater than three the Friedman test is conservative, increasing the likelihood of committing a Type II error (rejection of the Research Hypothesis when it is true). The power of the test determines the likelihood of avoiding a Type II error and depends on the number of IV levels and the sample size. The greater the number of IV levels the greater the power of the test. The probability that a Type I error will be made (rejecting the Null Hypothesis when it is true) is given by α. By convention α was set to 0.05, to avoid making the test too conservative.

4.2.3 Stimuli

The choice of stimulus type was important. Choices ranged from abstract synthetic signals to natural signals, from short to artificially long duration, and from isolated speech sounds, CVC mono-syllables, disyllabic CVCVC syllables, to sentences.

Abstract synthetic signals, such as double formant stimuli (Kortekaas & Kohlrausch, 1999) provide greater control but it may not be possible to generalise these results to natural speech. Alternatively, natural speech stimuli mimic the input to the signal processing stage in a real synthesis system and hence contain variables that would be encountered in such a system.

It is difficult to produce an individual speech sound, for example a phoneme, in isolation at a set pitch (Lenzo & Black, 2000), although CVC syllables are easy to produce in isolation. CVC syllables are also more akin to human communication than isolated speech sounds, which do not represent natural communication structures. Participants are familiar with the linguistic structures of syllables and words, therefore no perceptual adjustment is necessary for such stimuli (Jekosch, 1993); presenting CVC stimuli would provide a less contrived and abstract framework for the perception of distortion. Furthermore, CVC stimuli would provide longer bursts of speech sound, which make the perception of distortion less difficult than for shorter bursts. Tones lasting a second or more may be viewed as infinite by the auditory system (Gelfand, 1998), whereas auditory perception is altered for sound bursts of less than one second. Exceptionally short durations of 10ms or less result in transients that spread energy across the frequency range, which may affect experimental results (Wright, 1960).

Larger linguistic structures, such as bi-syllabic structures or sentences introduce many additional variables; the use of longer constructs such as sentences, may alter participants' perception due to the influence of ideals of prosody. Blouin & Bagshaw (2000) advocate the use of nonsense disyllables, as the judging of distortion in mono-syllables was reported to be difficult for participants. On the other hand, investigating phenomena in isolation, or with a minimum of factors present (the use of mono-syllabic structures as opposed to bi-syllabic structures containing many speech sounds, or sentences having prosody), provides greater control.

This experiment was concerned solely with the evaluation of distortion present in the stimuli. The effect of TD-PSOLA on vowel sounds was investigated. Vowel sounds only were evaluated during this experiment, due to the large number of possible stimuli and corresponding lengths of test runs. A CVC structure was chosen in which to embed each of the vowel phonemes, which would provide a natural communication framework with no perceptual adjustment necessary for the participants. Whole CVC syllables were chosen rather than recording individual CV and VC diphones to be concatenated. The concatenation process itself may introduce unwanted variables such as potentially poor segment joins.

A string list (Appendix B) consisting of 20 CVC (Consonant Vowel Consonant) syllables was generated. The vowel was varied for each syllable with the surrounding C*C structure kept

constant. The list is representative of most of the vowel sounds in the English language; it is commonly agreed that there are 14 pure vowels (IPA, 1949) although it is uncertain as to the number of combined vowels. The MBROLA (Dutoit *et al.*, 1996) speech synthesiser for British English isolates 6 combined vowels and these, with the 14 pure vowels, were included in the string list. The majority of possible vowel sounds were used to model the general effect of the TD-PSOLA algorithm, in terms of introduced distortion levels.

The string list contained both meaningful and non-meaningful CVC syllables. This mix has been encountered as a problem in existing speech assessment tests, for example the CLID test (Jekosch, 1992) where it was reported that participants needed to adjust to this concept. In an attempt to minimise this effect, it was stressed to participants that the syllables were both meaningful and non-meaningful, but that the phonetic identities of the speech sounds were of no importance; the evaluation criterion was solely the amount of distortion. The string list was also designed to contain syllables that would not cause an emotive reaction in the participants. Efforts have been made to reduce bias due to the content of the string list, and it will be assumed for this experiment that the effects are negligible, although further investigation would be desirable.

All vowel sounds to be evaluated were embedded between the same unvoiced consonants /k/ and /t/.

e.g. kEt, k{t, kA:t

The two short, unvoiced, plosive segments /k/ and /t/ were chosen as a contrast to the voiced vowel sounds to be evaluated. They also ensured the vowels' phonemic targets were reached by minimising coarticulation (Blouin & Bagshaw, 2000). They were standard for all the stimuli in the syllable set to minimise their influence on the results.

The natural CVC reference syllables were all uttered and recorded at a pitch of 220 Hz by a female speaker. The reference syllables were spoken at a steady rate and pitch. This frequency, the neutral pitch of speaker's voice, was determined by frequency analysis of the production of the 'schwa' sound and confirmed as the average fundamental frequency over three neutrally uttered sentences. The recordings were performed using the RPP (Reference Pitch Prompt) recording technique (Vine *et al.*, 1999) where a tone of 220Hz was played prior to recording to guide the speaker. The CVC recordings were checked to ensure a fundamental frequency of

220Hz using the Praat software (Boersma & Weenink, 1999). The fundamental frequencies of the waveforms were also measured by hand and confirmed using alternative speech software (Cool Edit 96, Syntrillium Software Corporation). The stimuli were recorded at a standard volume level, which was achieved by positioning the microphone at a set distance from the speaker and using standard recording settings for the Volume Control options on the PC.

To provide the modified stimuli for the levels of the IV, each of the CVC reference recordings was pitch-manipulated by various standard percentages using the Praat software implementation of the TD-PSOLA algorithm. This was achieved as described previously in Section 2.3. The fundamental frequency values were determined from the chosen percentages using the *mel* scale (Stevens & Volkmann, 1940), which provides a linear relationship between fundamental frequency and pitch:

$$m = 1125\log(0.0016f + 1) \qquad \text{Eqn 4.1}$$

where m = frequency in mels, f = frequency in Hz

The stimuli were presented to the participants via headphones to minimise any reverberations or other extraneous noises, caused by conducting the experiment in an office environment, which may affect participants' judgement.

All stimuli were recorded in one session, using the same computer and software so any variables due to the recording process, such as background noise, would be similar for all stimuli. Clicks due to the starting and stopping of recording were minimised by fading the stimuli in and out.

The female speaker used to record the CVC reference syllables, which have been modified to obtain the stimuli for the experiment, was the author of this work. This voice was chosen for practical reasons of cost and availability throughout the course of the research. This is not assumed to introduce bias for this experiment, as the author's voice is not compared to any other voice. The author recorded the several versions of each of the CVC reference syllables and selected the one that sounded most clear and consistent when replayed.

4.2.4 Pilot Study

A small pilot study was carried out to test the experimental design and procedure. Stimuli consisted of 20 vowel sounds embedded within a constant CVC structure. Initially, 7 IV levels of -15, -10, -5, 0, 5, 10 and 15% were used giving 140 stimuli.

Three participants participated and were debriefed at the conclusion of the experiment. The instructions for the test were given verbally. An example of the distortion to be judged was provided by the presentation of two unmodified speech waveforms and their corresponding TD-PSOLA modified versions. Participants were told that the change in quality, in terms of the buzzyness introduced, was called *distortion*.

During the experiment, the stimuli were presented via headphones in a random order using the C++ software in Appendix A. This software was written by the author to automate the experiment. Appendix A provides a brief description of the functionality of the code, shows screenshots of the interface and gives the code listing. Participants were asked to make their judgements using the 5-point MOS scale interface provided by the software.

The data analysis was found to be unnecessarily complex with two independent variables: pitch modification and positive versus negative manipulations. It also raised issues that were unanswerable with the current size of dataset. It was decided to split the experiment in two to enable more stimuli to be evaluated so such issues could be investigated. The first, which forms the main experiment presented in this section, investigates purely positive pitch modifications, and a second experiment, presented in Section 4.3, investigates positive versus negative modifications.

The data were analysed and the results indicated that increasing pitch manipulation led to greater distortion, supporting the research hypothesis. A large increase in distortion levels was found to occur between the 0 and -/+5% levels. The effect of the algorithm on distortion levels needed to be modelled for smaller pitch modifications, therefore an additional 1% modification level was introduced into the main experiment.

The size of the dataset (140 stimuli) seemed acceptable; participants reported slight boredom and fatigue towards the end of the experiment. It was concluded that stimulus numbers should be below this value for following experiments.

The test time for this pilot study was approximately 10-15 minutes. This time was used as an estimate of test duration to be given to participants for the main experiment.

No actual range of distortion level that the participants would encounter was provided before the test. This was identified as a problem when judging the level of distortion using a MOS scale. Consequently, unmodified stimuli and those judged having most distortion for this study would be used in the main experiment as examples of the range of distortion the participants may encounter.

It was discovered that without typewritten instructions, the participants were given varying descriptions of what was required from them. For the main experiment, a typewritten set of instructions was prepared, based on guidelines provided by the ITU-T recommendations for subjective assessment of quality (ITU-T, 1996).

The software interface was altered following the pilot study; it originally incorporated a stimulus number counter, but one participant noted it was distracting and this was removed from the interface.

Participants reported that it was a little difficult to judge such small differences in distortion levels on the MOS scale but thought they had succeeded. Initial indications from analysis of the pilot study data suggest that the use of a MOS scale facilitated the judgement of these variations in perceived distortion.

4.2.5 Choice of IV Levels for Main Experiment

Five IV levels were adopted for the main experiment of 0, +1, +5, +10 and +15%, giving 5 fundamental frequency levels of 220, 223, 233, 246, 259Hz. The 0% level (the unmodified version of each stimulus) was included as a "control" level to monitor the perceived quality of the

original recording of each CVC syllable. The 1% level allowed the effect of very small pitch manipulations on distortion levels to be evaluated. 1% represented a fundamental frequency change of ~3Hz, where 1 JND (Just Noticeable Difference), the smallest perceivable frequency difference, is cited as 2-3Hz for frequencies below 1000Hz (Ladefoged, 1996). The 5, 10, and 15% levels illustrated the effect of the algorithm on increasingly larger pitch manipulations.

Initially, the fundamental frequency of 220Hz (the speaker's neutral pitch) was converted to *mels*, then +1, +5, +10, and +15% *mel* values were calculated. These were then converted back into Hz. These values are shown in table 4.1.

Frequency (Hz)	frequency (mels)	% pitch manipulation
220 Hz	147 mels	0%
223 Hz	149 mels	1%
233 Hz	155 mels	5%
246 Hz	162 mels	10%
259 HZ	170 mels	15%

Table 4.1 % Pitch Manipulation and Corresponding F0 Values in Mels and Hz

Twenty CVC stimuli representing the majority of the vowel sounds in the English language were presented at five levels of pitch manipulation giving 100 stimuli.

4.2.6 Procedure

Participants were familiarised with the procedure and criterion via a fixed set of typed instructions (Appendix C). They were informed that they would be aurally presented via headphones with 100 stimuli each consisting of a CVC syllable. The CVC syllables were either non-meaningful or meaningful. It was made clear to the participants that the experiment was concerned solely on determining the amount of perceived distortion present rather that placing any importance on the phonetic identity, intelligibility, or meaning of the speech sounds. They were told that they would hear each stimulus once only and then must make a judgement.

To provide participants with a range for the MOS scale ratings and to familiarise them with the term *distortion* in the context of the effect of TD-PSOLA on speech, a short training session was conducted. This involved the aural presentation of two stimuli judged to be very distorted from the pilot study, and their non-modified versions to which TD-PSOLA had not been applied. Listeners were instructed that the change in the quality of the stimuli, in the form of buzzyness, was called *distortion*. Any further clarification of terms or procedure was given if necessary.

Participants were asked to judge using the MOS scale the amount of distortion present in each stimulus. To minimise the 'learning effect', the order of presentation of the stimuli was randomised for each test run. Additionally, judgement of a stimulus may be affected by the memory of the previous stimulus; this effect has also been minimised by the random presentation. The stimuli presentation, the randomisation of the stimuli, and the MOS scale interface was provided by the C++ software in Appendix A. Test-runs of 100 stimuli, with a delay of approximately 5 seconds between each stimulus presentation, lasted approximately 10 minutes.

All stimuli were presented at a standard volume level, which was achieved by using standard output settings for the Volume Control options on the PC. It should be noted that perceived loudness of a stimulus is frequency dependent; two auditory signals at different frequencies with equal power will have different perceived loudness. Under 500Hz lower frequency signals require more power to be perceived as having the same loudness as higher frequency signals (Robinson & Dadson, 1956). This has not been taken into consideration due to the small range of f0 manipulation (39Hz), although small variations in the recording or output level may be a factor.

4.2.7 Participants

Fifteen participants took part. All participants were university students or university staff, which makes use of a constrained sample population due to the issues of availability and cost. Participants ranged from 23-52 years of age, and from both genders (10 male, 5 female). All had self-reported normal hearing. The participants were unfamiliar with speech test procedures and were introduced to the testing method and the definition of the criterion as applied to synthetic speech prior to the test. They were not paid to participate in the test.

4.2.8 Test Conditions

Output Device: headphones

Acoustic Environment: quiet office

Noise Levels: Minimum background noise

PC: Pentium, 133MHz

Speech Spec:

- Voice: J. Longster
- M/ F: F
- Sampling Frequency: 44100Hz
- Speech units: CVC syllables
- Algorithm: TD-PSOLA, Praat Software (Boersma & Weenink, 1999).

4.2.9 Results

% pitch manipulation	Median of distortion levels (MOS scale rating)
0 %	1.75
1%	1.95
5 %	2.80
10 %	2.85
15 %	3.15

Table 4.2 Summary Statistics: Distortion Levels for % Pitch Modifications

Table 4.2 shows the percentage pitch manipulation of the stimuli from their original pitch and the corresponding median distortion levels, as judged on the MOS scale. The most obvious effect is that the perceived distortion ratings increase for increasing degrees of pitch manipulation from the original pitch.

The results for the five pitch manipulation levels are presented in Figure 4.1 where the medians of the distortion ratings are plotted as a function of pitch manipulation using a boxplot, to show how the distributions differ for each level.

Figure 4.1 also shows the spread of the data, illustrating that there is some overlap between the sets of results for each of the IV levels due to the variation in participants' level of judgement.

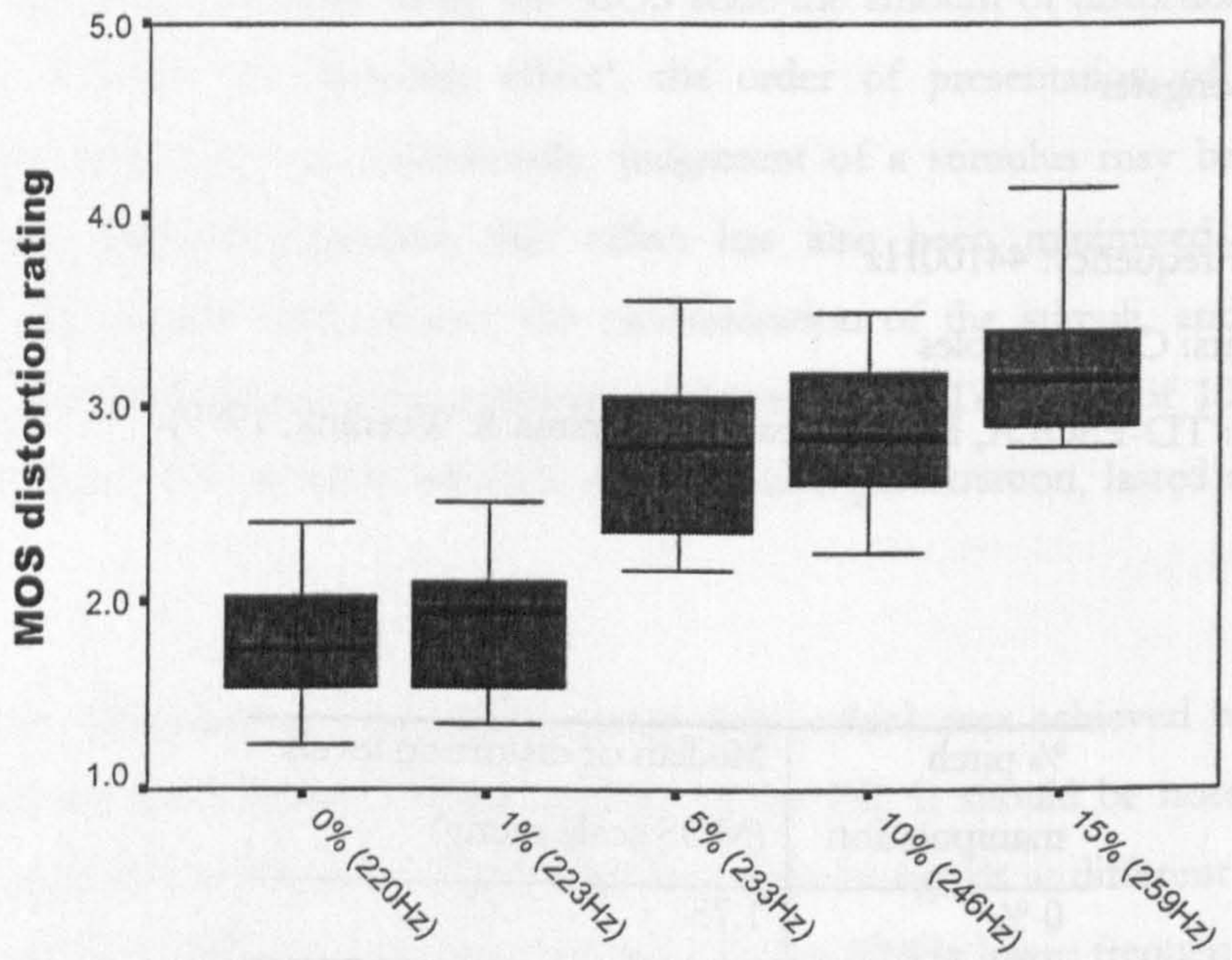


FIGURE 4.1 BOXPLOT OF PITCH MANIPULATION AND DISTORTION LEVELS

A within subjects Friedman test (corrected for ties) was performed on the data to compare the differences between the resulting levels of distortion. Unless stated, all statistical analysis was performed using SPSS for Windows, Release 10.0.5. A large significant effect was found, indicating a strong increase in perceived distortion with increasing pitch manipulation (χ_F^2 =56.564, df=4, N=15, $p<0.01$, *one-tailed test*). The test indicated that at least one of the medians of the five pitch manipulation levels differed and further analysis in the form of contrast tests was conducted to pinpoint the source of significance.

A set of Willcoxon Signed Rank tests was performed between the 0 and 1%, the 1 and 5%, the 5 and 10%, and the 10 and 15% manipulation levels to measure the size of differences between the levels of the IVs. As, the contrasts were unplanned, the α-level was adjusted using a Bonferroni adjustment by dividing the α-level by the number of contrasts to be carried out. For four contrasts, the error rate per contrast (EC) becomes:

EC=0.05/4=0.0125

Results of the set of contrasts are shown below:

0% and 1% manipulation level: Z=-2.7, $p<0.05$

1% and 5% manipulation level: Z=-3.4, $p<0.01$

5% and 10% manipulation level: Z=-2.2, $p<0.05$

10% and 15% manipulation level: Z=-3.4, $p<0.01$

The differences in the distributions were significant for all contrasts, except that of the 5% and 10% contrast when the Bonferroni adjustment is taken into account. The differences between the distributions for the 0% and 1% levels indicated that distortion may be introduced for even the smallest perceivable pitch modification. Observing Figure 4.1, it appears that distortion per unit percentage increase was greatest between 1 and 5% manipulation levels, which may be a critical pitch modification region in a corpus-based system.

A barchart (Figure 4.2) shows the average distortion ratings provided by the 15 participants for each of the five levels of pitch manipulation. All agreed in general, that there is an association between the extent of pitch manipulation and resulting levels of perceived distortion in the stimuli, although there was a variation in the level of response.

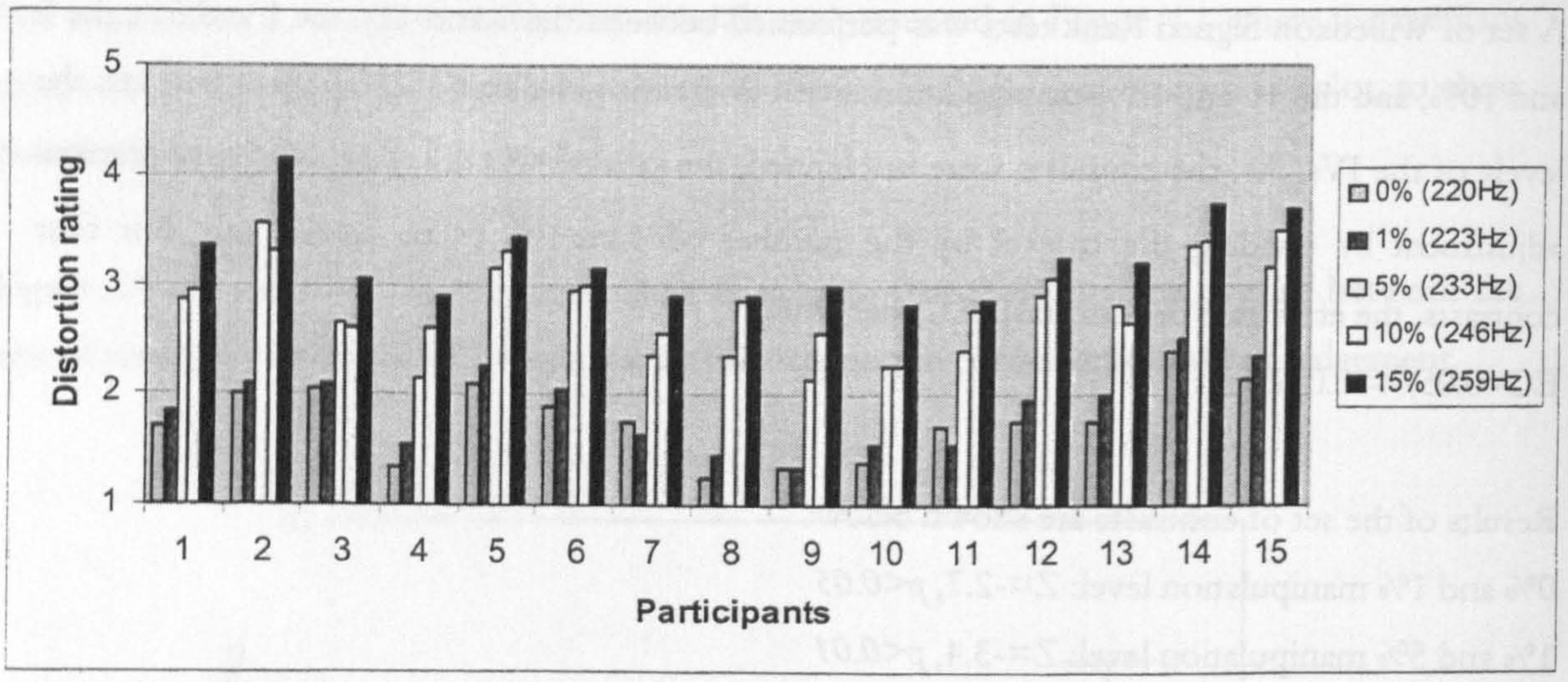


FIGURE 4.2 COMPARISON OF PARTICIPANT RESPONSE

4.2.10 Discussion

The experiment evaluated the effect of pitch manipulation using the TD-PSOLA algorithm on perceived distortion levels, or buzzyness, in speech. Significantly greater distortion was found to occur with increasing levels of pitch manipulation supporting the claims of Breen (1998), van Santen (1997), Black & Campbell (1995) and Blouin & Bagshaw (2000) that there may be an association between the extent of pitch manipulation and the resulting levels of perceived distortion.

This experiment evaluated distortion levels occurring due to small pitch modifications, unlike the works of Kawai *et al.* (1994) and Hirokawa & Hakoda (1990), who determined "thresholds of acceptability" over a larger range of pitch modifications. They cited large "acceptable" modifications of up to ~50% although it is argued for this work, that any perceived distortion is unacceptable. Such large modifications degrade voice quality leading to unnaturalness, which may be a larger perceptual cue for discrimination than distortion.

The work of Kortekaas & Kohlrausch (1997a) found a relationship between original fundamental frequency, target fundamental frequency, f1 value and resulting distortion for single formant stimuli. It was not possible to verify this as the sub-discrimination target fundamental frequency

values were outside the pitch manipulation range of this experiment. Although not experimentally verified, it appears that results for single formant stimuli could not be generalised to natural speech as major unnatural changes in voice quality occur at these larger pitch modifications, providing larger discrimination cues than distortion.

Kortekaas & Kohlrausch (1997a) also stated that PSOLA discrimination for fundamental frequency modification of a pure tone stimulus, although unverified, may be as low as 2%. The Wilcoxon Signed Rank test performed between the control level of 0% (no pitch modification) and the 1% level indicated that even the smallest perceivable pitch modification might cause significant perceptible distortion in these CVC stimuli. This result has a large impact on the design of the speech corpus and indicates that any amount of signal-processing applied may introduce some perceptible distortion. In an ideal situation, all occurrences of all phonemes in every prosodic context would eliminate the need for signal-processing totally, although this may never be the case due to the extreme variability of speech. This research will therefore focus on minimising the introduction of some of the perceived distortion, as a certain amount may be unavoidable when even the smallest modification is performed.

The 1% to 5% region of pitch manipulation showed a marked increase in distortion per percentage increase. The occurrence of distortions in this region would be critical for any corpus-based speech synthesis systems, where pitch modifications in this region may occur frequently. The impact of these results on the speech corpus design would suggest that the larger the speech corpus, and hence the variability of the pitch and duration of the segments, the less signal processing would be required, leading to a higher quality output with less distortion.

Figure 4.3 shows a bar chart of the CVC stimuli and their corresponding distortion ratings at each of the 5 levels of pitch manipulation. In general, all of the CVC stimuli suffer greater perceptible distortion levels when greater degrees of pitch modification are applied, although individual CVC stimuli appear to suffer varying amounts of distortion.

This raises the issue of whether the individual stimulus identity and hence spectral content may have an effect on the levels of perceived distortion that is introduced. A Friedman test for a within-subjects design, having an IV with more than two levels and consisting of non-parametric

data, was used to test whether participants rated the perceived distortion levels present in the stimuli differently depending upon their phonetic identity.

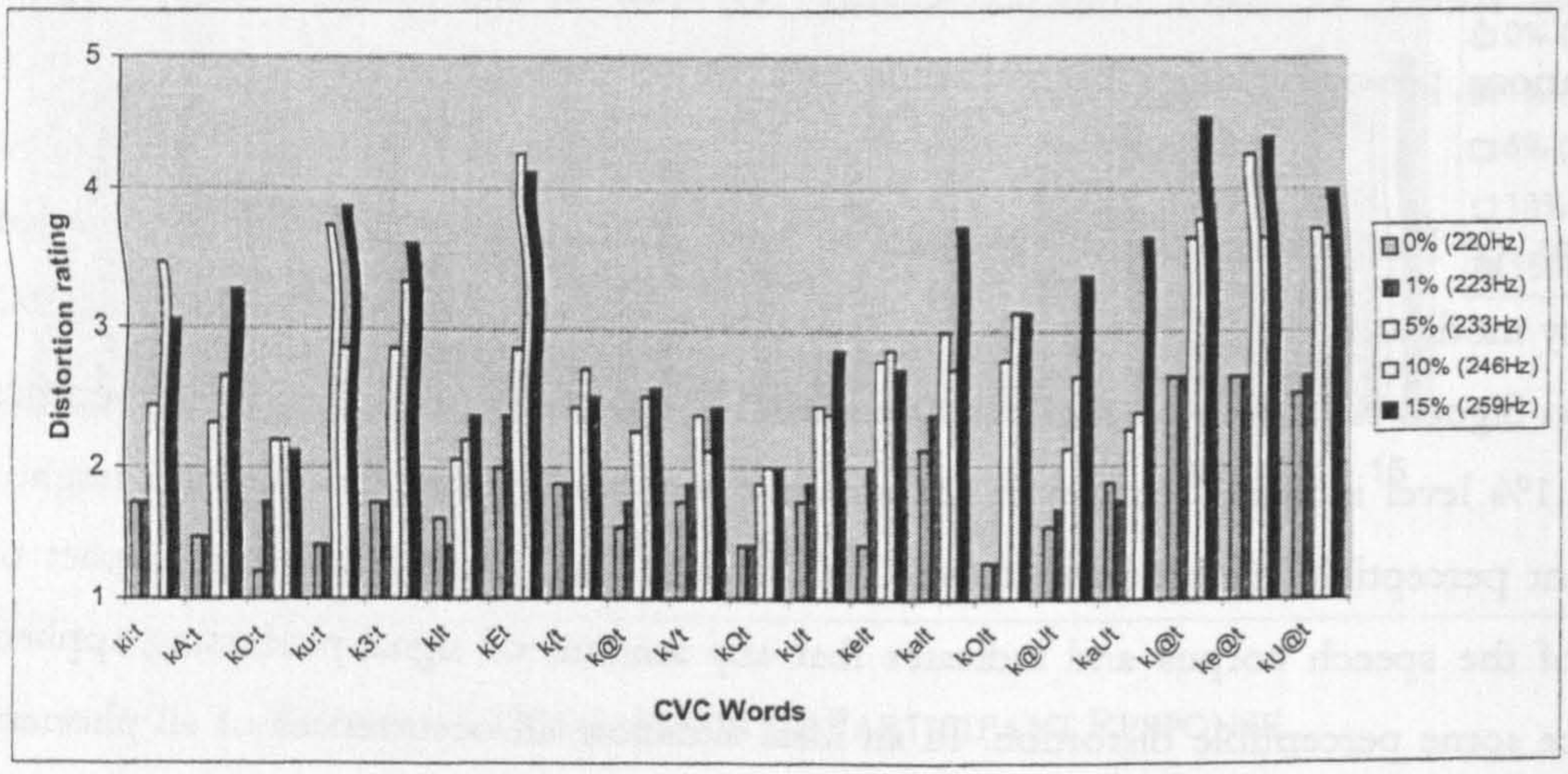


FIGURE 4.3 BARCHART OF CVC SYLLABLES AND DISTORTION RATINGS

Friedman's test determines whether the medians of the levels of the IV differ. There was a significant difference between the ratings for the twenty stimuli (χ_F^2 =169.5, df=19, N=15, $p<0.01$). This can be seen in the boxplot in Figure 4.4. The effect of stimulus identity is addressed further in the fourth experiment in Section 4.5, which is concerned with consonant stimuli.

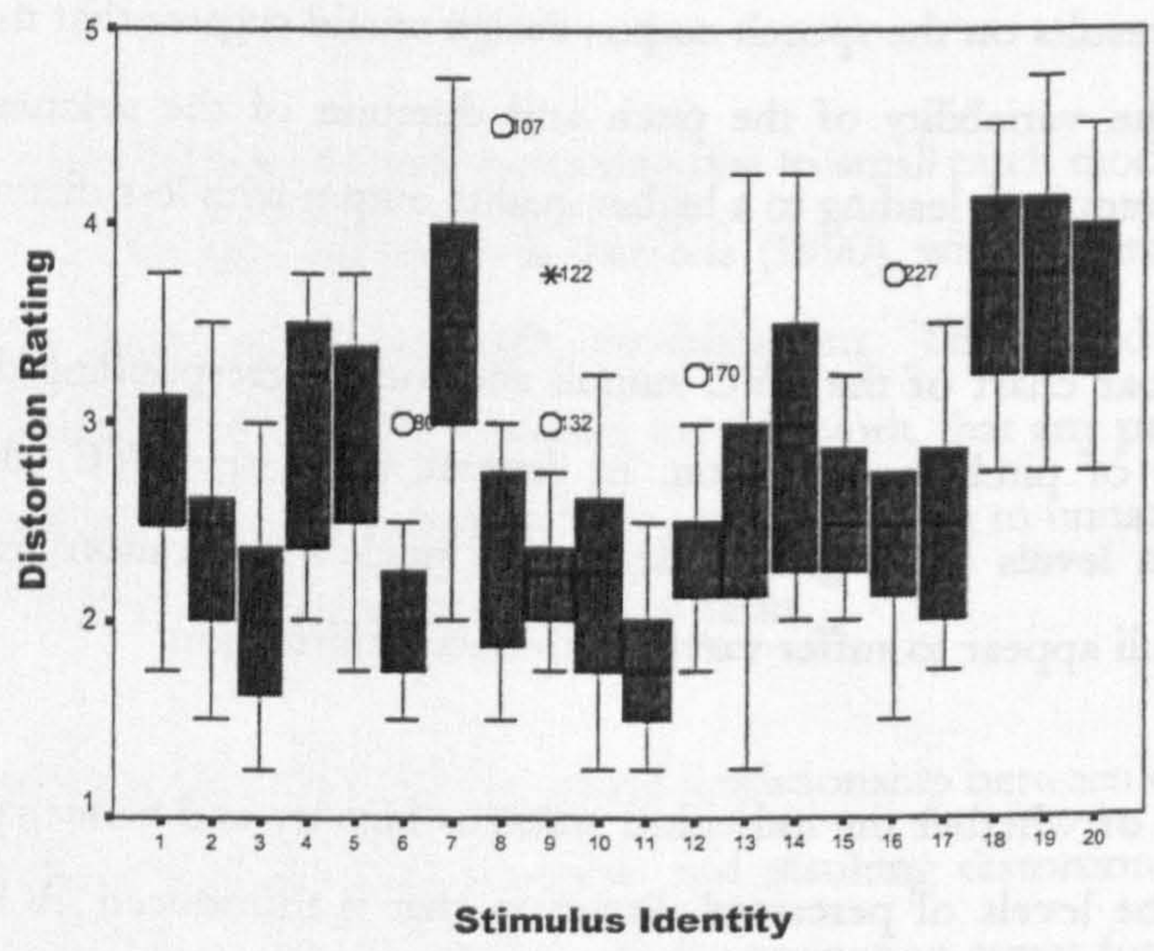


FIGURE 4.4 STIMULUS IDENTITY AND DISTORTION RATINGS

4.2.11 Conclusions

To conclude, the research hypothesis which stated that increasing degrees of TD-PSOLA pitch manipulation may introduce significantly larger levels of perceived distortion, in the form of buzzyness, has been supported.

The maximum pitch modification of 15% level was chosen to reflect some of the most frequently occurring modifications required in a corpus-based system. Previous research has cited "acceptable modification levels" of up to 50%, the point where significant loss of voice quality may occur and the speech begins to sound unnatural. It is argued here that in a corpus-based system such modification ratios would not be required often, or even not at all. With increasingly larger speech corpora available, modifications that cause extreme voice quality changes are no longer an issue. Distortion that occurs for smaller levels of modification is still an issue though, as any perceived distortion is undesirable and reduces the resulting speech quality.

The smallest perceivable pitch modification (1% or 3Hz) with the Praat implementation of TD-PSOLA may introduce perceptible distortion. The largest percentage increase in distortion could be seen for pitch modification levels of 1-5%. Both regions would be critical for a corpus-based system, where many of the modifications would be required. This indicates that for an ideal synthesis system, every occurrence of every phoneme combination in every prosodic context must be present in the speech corpus so that no signal processing is required. Due to the variability in speech, this may never be achieved, so steps need to be taken to minimise the distortion introduced when a signal processing algorithm such as TD-PSOLA is applied for necessary modifications in this region.

Finally, individual stimuli suffered significantly different distortion levels indicating that speech segment identity and hence composition may be a factor. This would be an important aspect in the design of the speech corpus; segments that respond adversely (in terms of large levels of perceived distortion) would require greater representation in the corpus in many prosodic contexts, and segments that respond better would require less representations. In this way, the amount of signal processing may be reduced for the adversely affected segments, and hence

minimise resulting distortion levels. The data from this experiment are analysed in Section 5.2.1 to design the speech corpus by determining the balance of the vowel representations in the corpus to be created. The results will also be used in Section 5.2.2 to develop a signal processing distortion measure for vowel phonemes, based on the levels of distortion perceived for each phoneme identity.

4.3 Experiment 2: The Effect of the TD-PSOLA Algorithm on Distortion Levels in Positive versus Negative Pitch Manipulated Speech

Abstract

A listening test evaluated the effect of positive pitch shifts (signal modification to a higher fundamental frequency) versus negative pitch shifts (signal modification to a lower fundamental frequency) with TD-PSOLA on distortion levels, in the form of buzzyness. Participants were presented aurally with 100 CVC stimuli in a random order, which had been pitch-modified positively and negatively by various standard amounts from their original pitch. The participants were asked to judge, on a scale of 1 to 5, the level of distortion present in each of the stimuli. Significantly greater distortion was found with increasing levels of pitch manipulation in both positive and negative directions, with similar distortion levels for both directions within the range tested.

4.3.1 Introduction

In a speech synthesis system, speech segments selected from an inventory or speech corpus may need to be manipulated either to a higher pitch or to a lower pitch, or both over various parts of the speech waveform. Research opinions are divided over whether positive or negative manipulations may suffer greater or less, or similar levels of distortion. Kortekaas & Kohlrausch (1997a) using single formant stimuli found that negative pitch shifted signals suffered less for a pitch modification range of ±20% for a 100Hz f0 signal. Kawai *et al.* (1994) also found that positive pitch shifted speech, at the word level, was more problematic for the range of 40-200%. Blouin & Bagshaw (2000) evaluating a range of 50-200% manipulation of CVCVC syllables found that positive and negative pitch modification appeared to suffer similar amounts of distortion. This experiment therefore investigates the effect of the TD-PSOLA algorithm on natural speech in terms of distortion for both positive and negative manipulations using the Praat implementation of TD-PSOLA.

Corpus-based systems contain multiple versions of speech segments, which may allow candidate segments to be chosen with pitches below or above the target pitch. If one direction of modification were found to be less problematic, it would be advantageous in terms of minimising resulting distortion, to select the speech segment that required this direction of modification. If no significant difference were found, this factor would not need to be taken into account in either the design of the corpus or the signal processing distortion measure. The pilot study carried out in Section 4.2.4 suggested that positive and negative pitch shifts using the Praat implementation of TD-PSOLA introduce similar amounts of perceptible distortion in stimuli over the range tested.

4.3.2 Design

This section states the experimental hypothesis and documents the design of the experiment, describing the statistical tests, the stimuli, and any other design considerations.

4.3.2.1 Hypothesis

H_1: There will be a significant correlation between the amounts of perceptible distortion introduced when speech is pitch manipulated either positively or negatively.

4.3.2.2 Structure of the Experiment

A listening test was undertaken to evaluate the amount of distortion, in the form of buzzyness, introduced by the TD-PSOLA algorithm into speech that has been pitch modified in a positive or negative direction from the original pitch. The independent variable *direction of pitch manipulation* has two levels of positive and negative modification.

The method of measurement of the dependent variable *distortion* was chosen. The use of a paired comparison test would involve the presentation of a positively and negatively modified stimulus and participants would be required to judge the more distorted. Paired comparison tests would not allow differences in amounts of distortion at different levels of pitch modification to be measured or allow individual pairs to be judged as having similar distortion. Using a direct

comparison may also lead to participants using other perceptual cues such as bass being preferable to treble for lower or higher fundamental frequency stimuli respectively. The dependent variable *distortion* is to be measured on a MOS (Mean Opinion Score) scale of 1 to 5 using the definitions from Experiment 1, Section 4.2.2.2.

A within-subjects design was used to reduce the effect of differences in response rating between participants; each participant rated the distortion for both levels of the IV. The data were ordinal, so a Spearman rank-order correlation coefficient was applied between the two levels to measure the linear relationship. A one-tailed test was conducted as the hypothesis predicted a positively correlated relationship.

4.3.3 Stimuli

The CVC string list (Appendix B) from Experiment 1, Section 4.2 was used for this experiment (see Section 4.2.3 for a full discussion of the choice of stimuli and string list). The use of the same material may introduce some bias, as certain participants are common to both experiments although these two experiments were carried out with a month delay between. A new set of stimuli was uttered and recorded at a pitch of 220Hz by the same female speaker as in the previous experiment, using the RPP recording technique (Vine *et al*, 1999). These were checked using the Praat and CoolEdit software to ensure a fundamental frequency of 220Hz. To provide the stimuli for the two levels of the IV, these were then pitch-manipulated by standard amounts with the Praat software (Boersma & Weenink, 1999) using the process described in Section 2.3. The manipulations were performed in steps of –8, -4, 0 +4, and +8% from the original pitch, giving 5 fundamental frequency levels of 200, 210, 220, 230 and 240Hz. The *mel* scale was used to determine these frequencies from the chosen percentages, providing a linear relationship between fundamental frequency and pitch:

$$m = 1125\log(0.0016f + 1) \qquad \text{Eqn 4.3}$$

where m = frequency in mels, f = frequency in Hz.

The stimuli were presented via headphones to minimise the effect of any extraneous noises due to conducting the experiment in an office environment. All stimuli were recorded in one session

using the same computer and software to ensure that any noises caused by the recording process, such as background noise would be similar for each stimulus. Stimuli were faded in and out to avoid clicks caused by the starting and stopping of the recording process. The female speaker used is the author of this work; no bias is assumed although each stimulus was checked to be clear and consistent when replayed.

4.3.4 Procedure

Participants were familiarised with the procedure and criterion via a fixed set of typed instructions (Appendix C). They were informed that they would be aurally presented with 100 stimuli each consisting of one CVC syllable. The stimuli that they would be presented with were either nonsense or meaningful syllables. It was made clear to the participants that the experiment was concerned solely on determining the amount of distortion present rather that placing any importance on the phonetic identity, intelligibility, or meaning of the speech sounds. They were told that they would hear each stimulus once only and then must make a judgement. The stimuli were assessed using the amount of perceived distortion as the criterion for evaluation. Distortion was defined as *buzzyness*, or *electronic* sounding.

To provide participants with a range for the MOS scale ratings and to familiarise them with the term *distortion* in the context of the effect of the TD-PSOLA algorithm on speech, a short training session was conducted. Two of the more distorted stimuli from Experiment 1 were presented as the examples of distorted speech. Non-manipulated speech, to which TD-PSOLA had not been applied, was provided as the contrasting example of speech with no distortion. Participants were told that the change in quality, in the form of buzzyness, between the pairs of stimuli was termed *distortion*.

Participants were asked to judge on the MOS scale how distorted or unnatural each stimulus sounded. To minimise the 'learning effect', the order of presentation of the stimuli was randomised for each test run. Test-runs of 100 stimuli, with a delay of approximately 5 seconds between each stimulus presentation, lasted 10 minutes. The test-runs were fully automated, and the MOS scale interface was provided by the C++ software in Appendix A.

4.3.5 Participants

Ten participants took part. The sample population was composed of university students or university staff due to the constraints of cost and availability. Participants ranged from 25-48 years of age, from both male and female genders (5 male, 5 female). All were asked whether to the best of their knowledge, they had normal hearing. Some of the participants were unfamiliar with speech test procedures and all were introduced to the testing method and the definition of the criterion as applied to synthetic speech prior to the test.

4.3.6 Test Conditions

Output Device: headphones

Acoustic Environment: quiet office

Noise Levels: minimum background noise

PC: Pentium, 133 MHz

Speech Spec:

- Voice: J. Longster
- M/ F: F
- Sampling Frequency: CD quality (44100Hz)
- Speech Units: CVC syllables
- Algorithm: TD-PSOLA, Praat Software (Boersma & Weenink, 1999).

4.3.7 Results

Modifications		Median distortion rating
Positive modification	+8%	2.55
	+4%	1.83
No modification	0%	1.43
Negative modification	-4%	1.88
	-8%	2.28

Table 4.3 Summary Statistics: Median Distortion for +ve and –ve Modifications

Table 4.3 shows the summary statistics for the two levels of positive and negative pitch manipulation, each having two percentage pitch manipulations from the original pitch. The 0% level was included as a control level to judge the participants' perception of unmodified speech. For each modification their median distortion levels judged on the MOS scale is given.

The most obvious effect is that positive and negative modifications appear to introduce similar amounts of distortion. Additionally, distortion increases with increasing degree of pitch manipulation in either direction, concurring with the results of Experiment 1. The results are presented in Figure 4.5 which shows the median distortion rating for each level of pitch manipulation: zero modification, and both positive and negative modification.

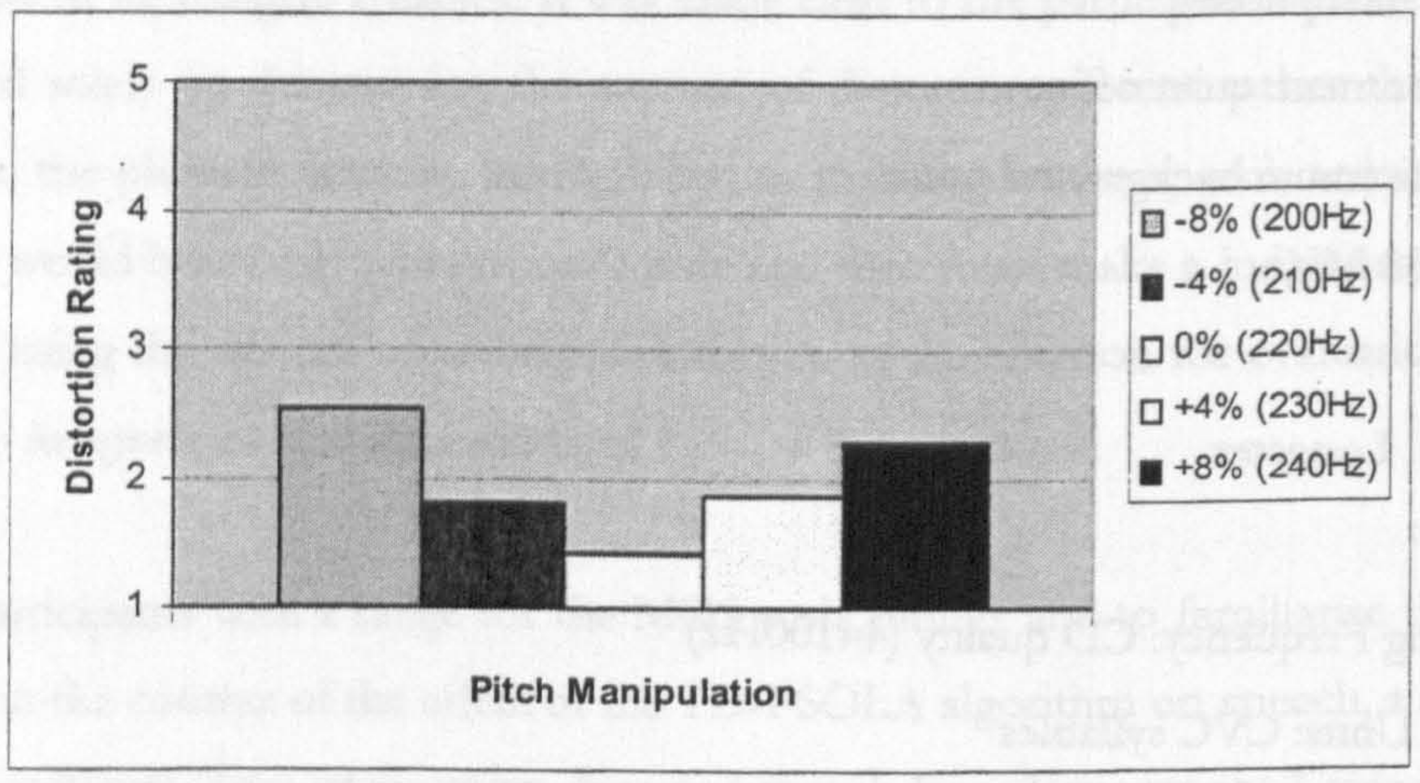


FIGURE 4.5 POSITIVE AND NEGATIVE PITCH MANIPULATION AND DISTORTION RATING

The linear relationship between the positive and negative modifications was measured by calculating Spearman's rho for the data from each of the ten participants, averaged for all CVC stimuli. There was a significant positive correlation between positive and negative pitch manipulation directions (rho=0.788, N=20, $p<0.01$, *one-tailed test*).

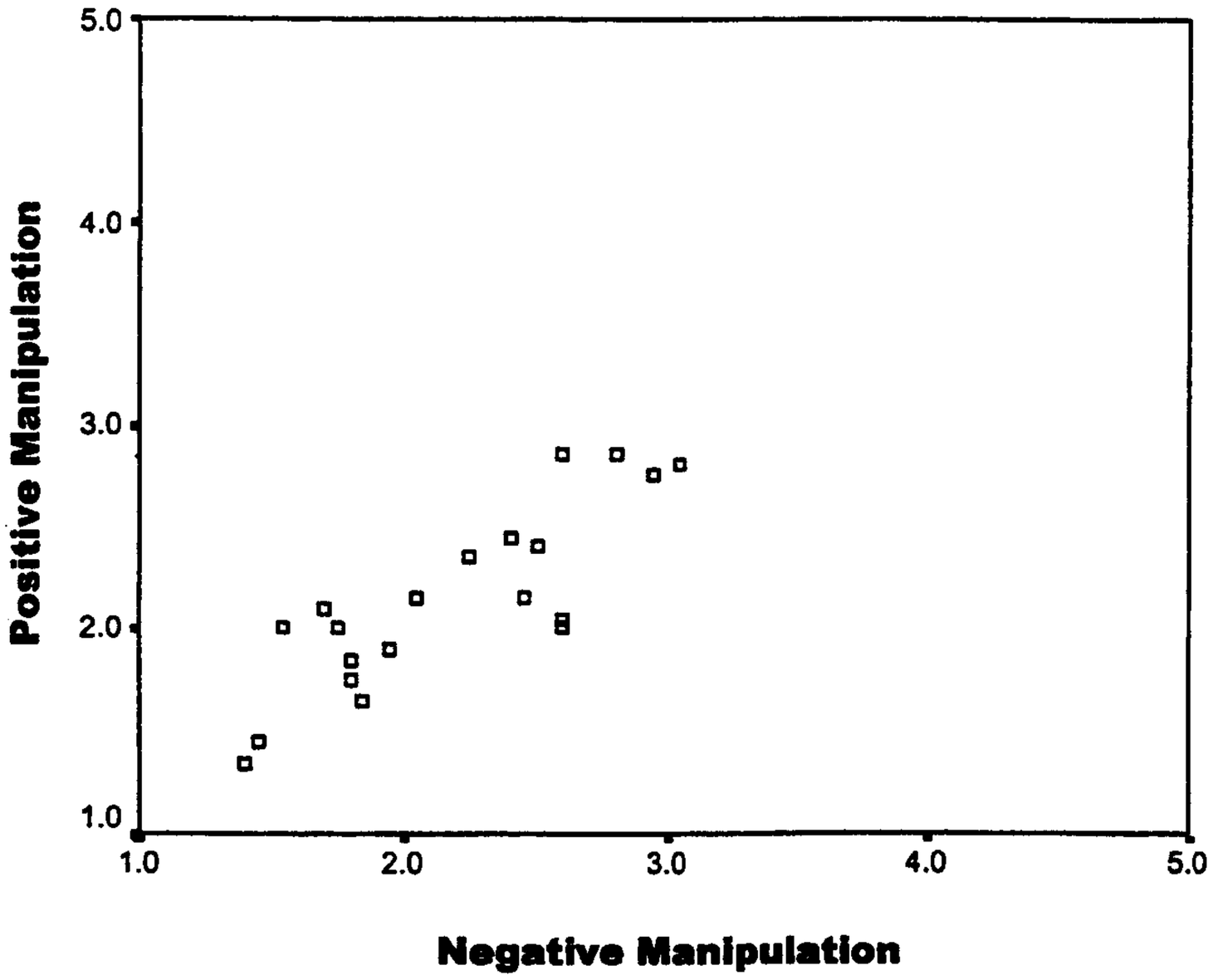


FIGURE 4.6 SCATTERGRAM OF RELATIONSHIP BETWEEN +VE AND –VE MODIFICATIONS

A scattergram (Figure 4.6) shows the relationship of distortion levels for the positive and negative modifications, with MOS values averaged for all CVC stimuli. The scattergram indicates that distortion levels for positive and negative manipulations are correlated, although it suggests that negative modifications may be slightly more problematic. The scattergram also shows there are no outliers which would affect the value of rho.

4.3.8 Discussion

The results indicate that in general, there is a correlation between resulting amounts of perceived distortion for both positive and negative modifications, although negative modifications may be slightly more problematic. This supports the work of Blouin & Bagshaw (2000) but is contrary to the work of Kortekaas & Kohlrausch (1997a) and Kawai *et al.* (1994). This difference of opinion may be due to the stimuli under investigation (from abstract to sentence level), the implementation of the algorithm, and the range over which this was tested.

4.3.9 Conclusions

To conclude, there is a significant relationship in the form of a positive correlation between the distortion ratings for positive and negative frequency manipulation between the limits of -8% and 8%. This suggests that in general speech is affected similarly in terms of perceived distortion when pitch is increased or decreased using TD-PSOLA. This supports the work of Blouin & Bagshaw (2000) who investigated this over a greater pitch modification range.

It also appeared that individual stimuli were affected differently when pitch was increased or decreased although it would require further experimentation to verify this. This would not have any bearing on the design of the speech corpus (segments would not require more representation at either higher or lower frequencies) but may influence the segment selection process. For example, certain phonemes may respond better to the application of the algorithm in terms of less perceived distortion when performing pitch modifications in a positive direction. In this case, it would be advantageous to select a segment with a lower f0 than the target value for the sentence to be synthesised. For the purpose of this research, it will be assumed that all phonemes respond similarly to positive and negative modifications over a small range. Future experiments will evaluate only positive modifications, although the investigation of the effect on individual stimuli will be discussed as further work in Section 7.2.

Figure 4.7 illustrates the effect of modifications on the individual stimuli for this experiment.

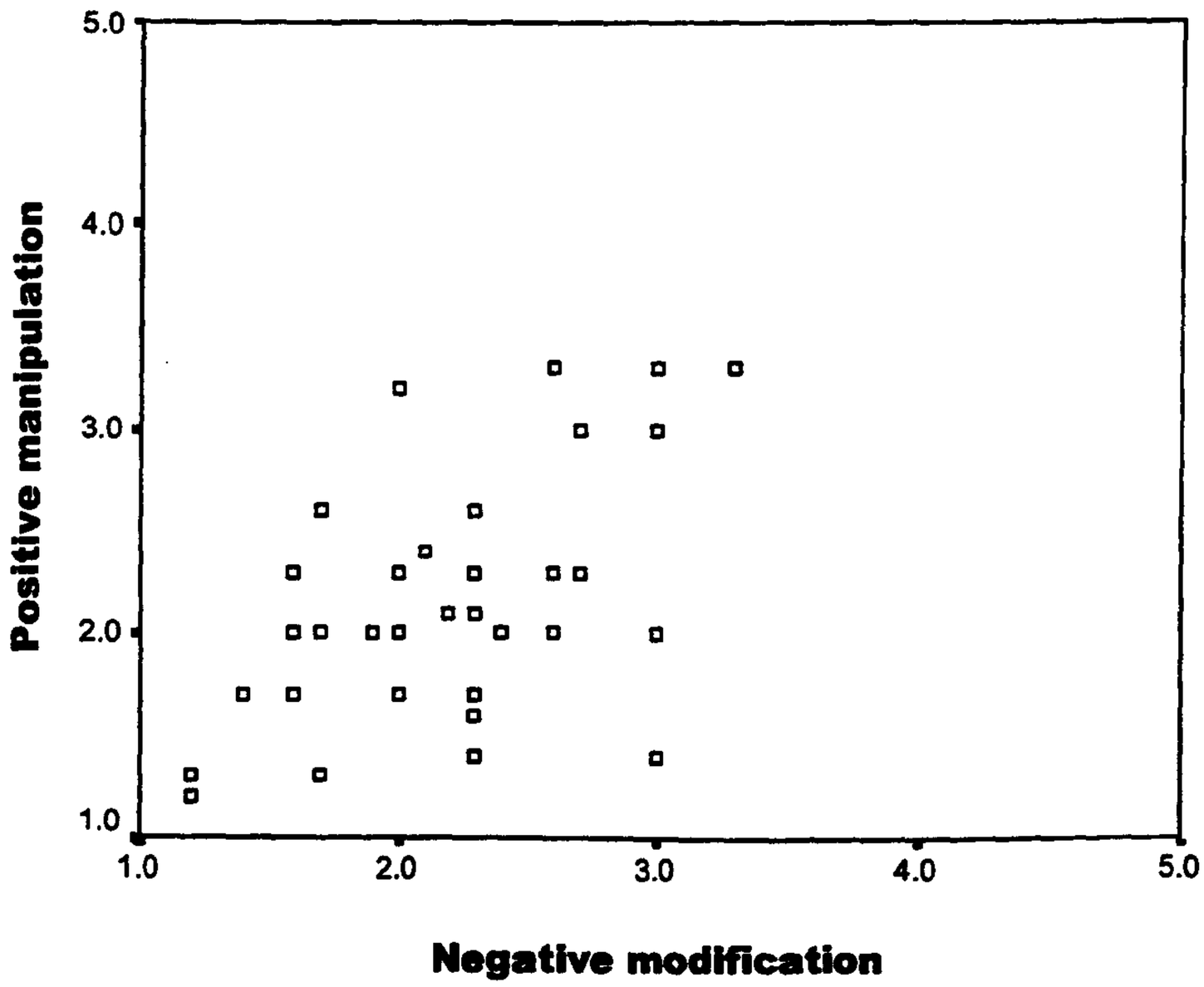


FIGURE 4.7 SCATTERGRAM OF DISTORTION LEVELS FOR INDIVIDUAL STIMULI

It appears that individual identity stimuli are not affected similarly, which concurs with the findings from Experiment 1. Calculation of Spearman's rho is significant although less so (rho=0.519, N=40, $p<0.01$, *one-tailed test*), when comparing the relationship between positive and negative modifications for individual stimuli. This suggests that individual stimuli may not be affected similarly for positive and negative modifications, which may have an impact on the selection of segments from the speech corpus. This would require more investigation and is discussed in Section 7.2 as further work. For the purposes of this research, over the pitch manipulation range tested, segments will be assumed to respond similarly in terms of the distortion introduced when modified either positively or negatively.

4.4 Experiment 3: The Effect of Pitch Manipulation using the TD-PSOLA Algorithm on Distortion Levels in Synthetic Speech at the Sentence Level

Abstract

A listening test was undertaken to determine the effect of TD-PSOLA pitch manipulation on perceptible distortion levels in synthetic speech at the sentence level. Participants were presented aurally with 8 stimuli, consisting of simple parts of sentences. The sentences were synthesised from different syllable inventories, both using the TD-PSOLA algorithm for final pitch manipulation. The first inventory consisted of CV, VC and CC syllables recorded at the neutral pitch of the speaker, and the second consisted of CV, VC and CC syllables recorded at pitches closer to the target pitch of the sentences to be synthesised. Participants were asked to judge, on a scale of 1 to 5, the levels of distortion present and the humanness of each of the synthesised sentences. Humanness was defined as the naturalness and smoothness of speech and was used to evaluate the effect of the concatenation process. On average, greater distortion and humanness were found with sentences synthesised using the first inventory, which required greater degrees of pitch manipulation to achieve the target prosody. The results were not significant and possible reasons for this are discussed.

4.4.1 Introduction

Experiments 1 and 2 indicated that perceived distortion levels increase significantly with greater pitch manipulation at the CVC syllable level. These experiments evaluated stimuli at the word level only, and without concatenation. This experiment was designed to determine whether such observations could be generalised to the sentence level, and also what effect the concatenation of speech segments has on the perception of the output. Two segment inventories will be used, with one containing segments closer to the target pitch values, to determine whether perceived distortion levels increase significantly with greater pitch modification at the sentence level.

In previous experiments, only static pitch contours have been imposed on stimuli having original static pitch contours. According to Blouin & Bagshaw (2000), this would be the worst-case scenario in terms of greater resulting levels of distortion. Imposing dynamic pitch contours may

mask some of the distortion, leading to less distortion being perceived. The results of this experiment would indicate whether the results for CVC stimuli from Experiment 1 could be generalised to the sentence level and therefore be used to inform the design of the speech corpus and segment selection measure to be used for sentence level constructs.

Additionally, previous stimuli have not undergone the concatenation process. The concatenation process may degrade the smoothness of speech, depending on the properties of the segments to be concatenated, which may affect participants' judgements of distortion. This was evaluated during this experiment using the criterion 'humanness', defined as naturalness and smoothness of the speech.

4.4.2 Design

This section states the experimental hypotheses and documents the design considerations.

4.4.2.1 Hypotheses

H1: Sentence level stimuli requiring greater TD-PSOLA pitch manipulation will suffer significantly greater distortion levels.
H2: Sentence level stimuli constructed from syllables having similar fundamental frequencies will be more human sounding than those constructed from syllables having different fundamental frequencies when pitch mauipulated with TD-PSOLA.

4.4.2.2 Structure of the Experiment

A listening test was designed to evaluate the effect of TD-PSOLA pitch-modification on synthetic speech at the sentence level. The independent variable was *inventory type*. Two inventories were evaluated; the first inventory consisted of CV, VC and CC syllables recorded at the neutral pitch of the speaker, and the second consisted of CV, VC and CC syllables recorded at pitches closer to the target pitch of the sentences to be synthesised. In this way, sentences synthesised using Inventory 1 required greater pitch modification than those synthesised using Inventory 2. Additionally, the method of concatenating the segments differed between the

inventories. Two dependent variables *distortion* and *humanness* were measured on a MOS (Mean Opinion Score) scale of 1 to 5 shown below:

1 no perceived distortion/ very human
2 quite undistorted/ quite human
3 distorted/ slightly human
4 quite distorted/ quite inhuman
5 very distorted/ very inhuman

The first dependent variable *distortion* was defined as in previous experiments, as *buzzyness*, or *electronic sounding*. The second dependent variable *humanness* was included to determine the effect of the concatenation process on the perception of stimuli. The aim was to evaluate the presence of any audible discontinuities in the speech, so humanness was defined as *naturalness* or *smoothness*.

A within-subjects design was used; each participant rated the distortion and humanness of both sets of sentences to reduce the effects of variations in participants' level of response and hence provide a more powerful test. The requirements of a within-subjects t-test were not fulfilled, as the data were ordinal, therefore a Wilcoxon Signed Rank test was applied to the non-parametric data to detect differences in the distribution of the distortion for sentences from each inventory.

4.4.3 Stimuli

Identical short parts of sentences were synthesised using both inventories. The following four sentence parts were used:

"No way!"
"My cat?"
"Look here….."
"Prove it."

Short parts of sentences were chosen to minimise ideals of prosody that may affect participants' perception, but still allow the effects of pitch modification at the sentence level to be investigated. Initially, these sentences were recorded by the speaker, and then analysed for pitch and duration information using the Praat software. As all pitch movements are not perceptible, the pitch contour was stylised using the algorithm provided by Praat. The syllables and their fundamental frequencies are given in Table 4.4. For unvoiced speech such as /k/ in the CV syllable /k{/, there is no pitch information, so the pitch contour given in the table applies to the voiced part only.

Sentence	*CV, VC, and CC syllables and frequencies*
No way!	n@U (400 - 340Hz) @Uw (340 - 270Hz) weI (330 - 360Hz - 280Hz)
My cat?	maI (480 - 515Hz) aIk (515 - 380Hz) k{ (240 – 260Hz) {t (260 – 280Hz)
Look here....	lU (260 - 280Hz) Uk (280Hz) hI@ (320 - 300Hz - 280Hz)
Prove it.	pr(-) ru: (330 - 300Hz) u:v (300 - 260Hz) vI (180 - 170Hz) It (170 – 160Hz)

Table 4.4 Syllables and Fundamental Frequency Contours of Test Sentences

This prosody information was imposed on both sets of sentences after synthesis to provide standard values.

To control for the effect of the necessary duration manipulation, the syllables were recorded using the RPP recording technique; all segments were recorded with similar durations, which can be controlled to a certain extent by recording them in isolation. Similar duration values were then imposed on the sets of syllables from both inventories.

Inventory 1: The CV, VC and CC syllables required to synthesise the sentences were recorded using the RPP recording technique at the neutral pitch of the speaker (220Hz). All of the syllables possessed similar fundamental frequencies, so they were simply concatenated to form the test sentences by cutting the syllables at the stable mid-point of each phoneme in the syllable structure and abutting them together. Careful attention was paid to concatenate the syllables at zero crossings on the time-domain waveform and at similar positions in the voiced cycles. The TD-PSOLA algorithm was applied to manipulate the pitch and duration of the synthetic

sentences to match the prosody of the previously recorded natural sentences (the f0 values are shown in Table 4.4).

Inventory 2: The CV, VC, and CC segments required to synthesise the sentences were recorded at the frequencies shown in Table 4.5.

Sentence	CV, VC, and CC syllables and frequencies
No way!	n@U (370Hz) @Uw (305Hz) weI (320Hz)
My cat?	maI (498Hz) aIk (448Hz) k{ (250Hz) {t (270Hz)
Look here....	lU (270Hz) Uk (280Hz) hI@ (300Hz)
Prove it.	pr(-) ru: (315Hz) u:v (280Hz) vI (175Hz) It (165Hz)

Table 4.5 Synthesis Fundamental Frequencies of Syllables

The frequencies were chosen to be midway between the maximum and minimum pitch changes from the start of the voiced part to the temporally central voiced part for each CV syllable, and from the temporally central part of the voiced part to the end of the voiced part of each VC syllable. The syllables were individually pitch modified using TD-PSOLA to midway between the fundamental frequencies of the segments to be joined. The aim is to minimise audible discontinuities that may occur when concatenating segments of different fundamental frequencies. For example, /n@U/ and /@Uw/ were modified to an intermediate, static f0 value of 337Hz. These were joined by cutting the syllables and abutting them together, as performed for Inventory 1 stimuli. TD-PSOLA was then used to impose the final target prosody provided by pitch and timing data extracted from the natural sentences (the f0 values are shown in Table 4.4). The difference in concatenation methods between the two inventories should be noted. For Inventory 1, segments were concatenated and then TD-PSOLA was applied to achieve the target prosody. For Inventory 2, the segments were individually modified to a common intermediate fundamental frequency, then concatenated and modified to their final target prosody values. The concatenation method used for Inventory 2 may give rise to larger audible discontinuities than the concatenation method used for Inventory 1 due to the additional modification stage, and if so, this may have an effect on the perception of the output.

4.4.4 Procedure

Participants were familiarised with the procedure via a set of typed instructions (Appendix C). They were informed that they would be aurally presented via headphones with 8 short sentences. The sentences were assessed using the criteria of 'perceived distortion' and 'humanness'. Distortion was defined as *buzzyness* or *electronic* sounding, and humanness was defined as *naturalness*, or *smoothness of speech*. It was made clear to the participants that the experiment was concerned with determining the amount of distortion present and determining the humanness of the sentences; no importance was to be placed on the intelligibility, prosody or meaning of the sentences. Each sentence would be presented once only and then the participants must make a judgement. To provide the participants with a range of potential distortion they may hear and to familiarise them with the terms distortion and humanness, a short training session was conducted. The training for the distortion criterion involved the presentation of the training stimuli from Experiment 2; listeners were told the change in voice quality, in terms of buzzyness, was called distortion. For the definition of humanness, participants were told to rate any audible discontinuities present in the speech.

To minimise learning effects, the order of presentation of the stimuli was randomised for each test run. Test-runs, with a delay of approximately 5 seconds between the stimulus presentations, lasted approximately 5 minutes each, including time taken for training and explanation of the test procedure.

4.4.5 Participants

Ten participants took part. The restricted sample population, consisting of postgraduate students or university staff, was due to availability of participants. The participants ranged from 22-46 years of age (7 male, 3 female). All were asked whether to the best of their knowledge, they had normal hearing. The participants were unfamiliar with speech test procedures and were familiarised with the definition of the criterion as applied to synthetic speech and the procedure prior to the test.

The stimuli presentation, randomisation and MOS scale interface were provided by the C++ software in Appendix A.

4.4.6 Test Conditions

Output Device: headphones

Acoustic Environment: quiet office

Noise Levels: minimum background noise

PC: Pentium, 133 MHz

Speech Spec:

- Voice: J. Longster
- M/ F: F
- Sampling Frequency: CD quality (44100Hz)
- Speech Units: CV, VC and CC syllables concatenated to form short sentences
- Algorithm: TD-PSOLA, Praat Software (Boersma & Weenink, 1999).

4.4.7 Results

Synthesis Inventory		Median distortion rating
Inventory 1	Distortion	2.7
	Humanness	3.55
Inventory 2	Distortion	2.45
	Humanness	3.75

Table 4.6 Summary Statistics: Medians of Distortion and Humanness

Table 4.6 shows the distortion and humanness ratings for speech synthesised from the two inventories. The most obvious effect is that Inventory 1 stimuli, which have undergone greater pitch manipulation, appear to suffer greater distortion. Inventory 2 stimuli, which were synthesised using the more complex concatenation method have on average a greater humanness rating than Inventory 1 stimuli, indicating that speech synthesised from this inventory may be

smoother with less audible discontinuities. Figure 4.8 shows a bar chart illustrating the two synthesis methods and the corresponding median distortion and humanness ratings.

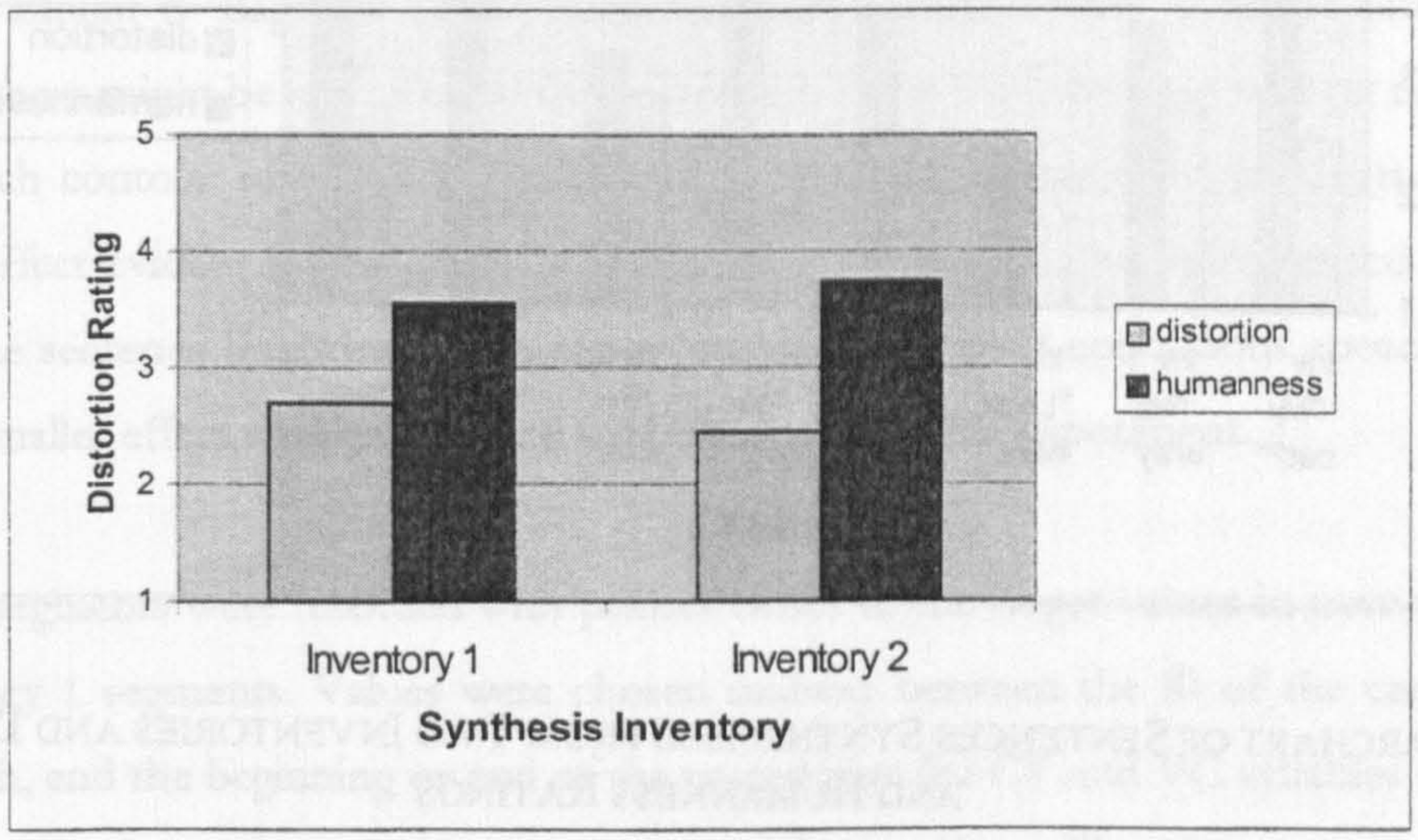


FIGURE 4.8 BARCHART OF SYNTHESIS INVENTORIES WITH DISTORTION AND HUMANNESS RATINGS

Figure 4.9 shows a bar chart of the two sets of sentences synthesised from the two inventories and their corresponding average distortion and humanness ratings. For three of the four sentences, those synthesised from the Inventory 1 appear to suffer greater distortion. Three of the four sentences synthesised from Inventory 2 show greater humanness ratings.

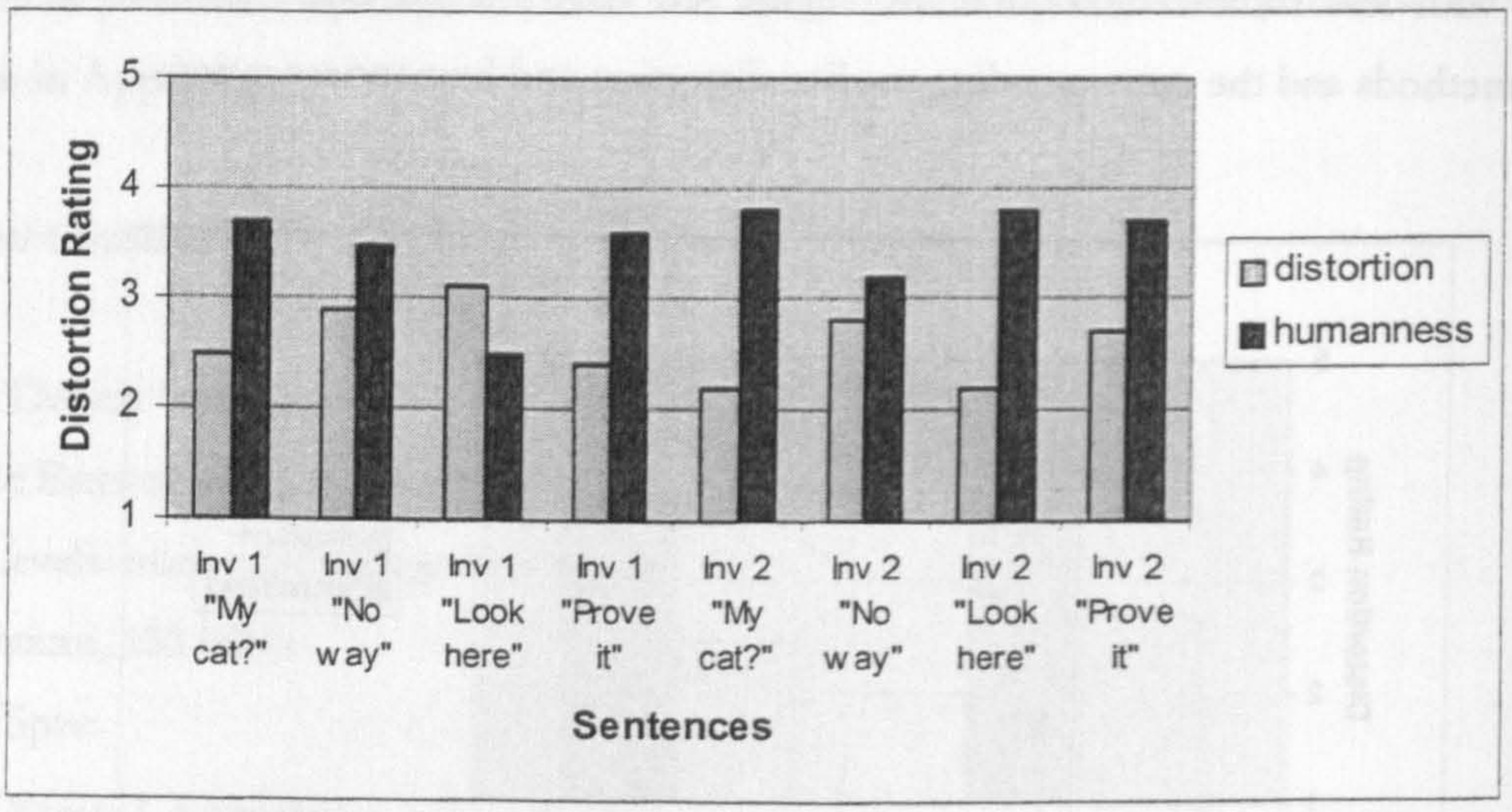


FIGURE 4.9 BARCHART OF SENTENCES SYNTHESISED FROM TWO INVENTORIES AND DISTORTION AND HUMANNESS RATINGS

A Wilcoxon Signed Rank test was performed on the data to detect differences in the distributions of the two IV levels. Neither the distortion distributions ($Z=-1.28$, $N=10$, *$p>0.05$, one-tailed test*) nor the humanness distributions ($Z=-1.07$, $N=10$, *$p>0.05$, one-tailed test*) were found to be significantly different for speech synthesised using the different inventories.

4.4.8 Discussion

Retrospective power analysis was conducted on the data, as the results were not found to be significant. The effect size was calculated from the data using Cohen's d, and was found to be 0.41 for the distortion measure and 0.34 for the humanness measure. Clark-Carter (1999) provides tables to determine the power of this test; for a one-tailed test, $\alpha=0.05$, with an effect size of between 0.41 and 0.34, having 10 participants, the power is given as 0.2-0.4. To achieve a power of 0.8, 40-70 participants would have been necessary. Cohen (1988) states an effect size of $d=0.2$ is small, $d=0.5$ is medium, and $d=0.8$ is large. For comparative purposes, the effect size for the data from Experiment 1 was calculated. Clark-Carter (1999) uses the effect size of ANOVA, η^2, as an estimate of the effect size for a Friedman test. The effect size of Experiment 1 was found to be $\eta^2=0.47$, which Cohen describes as large. The smaller effect size for the current

experiment and hence power of the test could therefore be responsible for the rejection of the experimental hypotheses.

Research by Blouin & Bagshaw (2000) suggested that stimuli having dynamic pitch contours imposed on them might be less affected than ones with static contours imposed on them. Results for static pitch contour stimuli may not be as pertinent for dynamic pitch targets. This could reduce any effect evident in Experiments 1 and 2 at the static-pitch contour word level, when applied to the sentence level due to the highly dynamic nature of continuous speech. This may explain the smaller effect sizes of 0.41 and 0.34 observed for this experiment.

Inventory 2 segments were recorded with pitches closer to the target values in terms of mean f0 than Inventory 1 segments. Values were chosen midway between the f0 of the central part of voiced speech, and the beginning or end of the voiced part for CV and VC syllables respectively. Inventory 2 segments, although closer to the target frequencies, still underwent extremely large pitch modifications of up to 68 Hz (30%). Such large modifications were outside levels investigated so far. It is possible that such large modifications caused changes in voice qualities that masked the occurrence of buzzyness under investigation, making the levels indistinguishable from those occurring in Inventory 1 stimuli.

The lack of significance may also be due to the unpredictable nature of the algorithm already observed; subjecting all segments to smaller degrees of pitch manipulation may not be enough to minimise distortion. In Experiments 1 and 2, segment identity appeared to have a large effect on the perceived distortion. It may be that certain segments can withstand greater or lesser pitch modifications and that this should be taken into consideration.

4.4.9 Conclusions

Sentence level stimuli requiring greater TD-PSOLA modification did not suffer significantly different distortion levels than those requiring less modification. The Wilcoxon test indicated that there was no significant difference between the distributions of the stimuli synthesised from the two inventories.

This may be partly due to the low power of the test caused by the smaller effect size. The smaller effect size could be due to imposing dynamic pitch contours onto the stimuli which may not be as problematic for TD-PSOLA as static pitch contours. This means that the design of the speech corpus and signal processing distortion measure using data from experiments making use of static pitch contour word-level stimuli may not be as pertinent when used for sentence level stimuli.

Sentence level stimuli created by the concatenation of different pitch CV and VC segments did not sound significantly less smooth than those created from CV and VC segments having the same f0. It was expected that segments synthesised from Inventory 2 would be perceived as significantly less smooth than sentences synthesised from Inventory 1, due to the need to concatenate segments with different original fundamental frequencies. Evidence suggests that this may not be a problem, but stimulus preparation using the two-stage modification approach was a more difficult process and some discontinuities were audible. After the final dynamic pitch contour was applied though, these effects were almost imperceptible. Although this claim is not experimentally verified, it does support the view that the dynamic nature of speech may explain much of the success of TD-PSOLA.

To conclude, the smaller effect size was thought to be due to the imposition of dynamic pitch contours, and that any sentence level speech requiring dynamic pitch contour applications would suffer less distortion. In addition, the large modifications required for both sets of segments may have masked the effect for smaller modifications investigated in the previous experiments at the word level. The implications for the speech corpus and signal processing distortion measure design are that effects seen at the word level, which are used to inform their design, provide the worst-case scenarios. Such effects may be less evident when the speech corpus is used for sentence level synthesis and the signal processing distortion measure is used as part of the segment selection process.

4.5 Experiment 4: The Effect of Pitch Manipulation using the TD-PSOLA Algorithm on Distortion Levels in Speech for Various Voices

Abstract

An experiment was undertaken to determine whether the effect of pitch manipulation using the TD-PSOLA algorithm on distortion levels in speech is significantly different for various voices. Four voices, two male and two female, were evaluated. Twenty participants were presented aurally with 92 CVC stimuli, 23 of each for the four voices, in a random order. The stimuli had been pitch manipulated using the TD-PSOLA algorithm, by various standard amounts from their original pitch. Participants were asked to judge on a MOS scale of 1 to 5 the levels of distortion present in each of the stimuli. Voices were found to respond differently to the algorithm, some appearing to suffer less distortion, although in general, greater distortion was found with increasing levels of pitch manipulation for all voices. The CVC stimuli had varying initial consonants, and the effect of this phoneme identity on distortion was investigated post hoc.

4.5.1 Introduction

Certain voices may respond better to the application of the TD-PSOLA algorithm (Lowry, 1999) in terms of perceived distortion levels. In addition, some evidence suggests that higher f0 voices, such as female voices may suffer more (Moulines & Charpentier 1990, Blouin & Bagshaw 2000). This experiment investigates whether greater pitch manipulation may contribute to greater distortion for all voices and whether certain voices suffer more or less. The results from this could provide some guidance for speaker selection when recording a corpus, to choose a voice that will respond well to the TD-PSOLA algorithm, in terms of minimal introduction of distortion in the form of buzzyness.

Experiment 1 showed that certain stimuli were more affected than others depending on the phonemic identity; this experiment investigates whether there is a relationship between distortion levels for individual stimuli for each voice, to determine whether results for one voice can be

generalised to others. This would suggest whether the design of the speech corpus and signal processing distortion measure developed during this work might be used for other voices.

Experiment 1 evaluated the amount of perceived distortion present in CVC stimuli in which only the central vowel was altered. This experiment uses CVC stimuli in which only the initial consonant is altered, and the effect of the phoneme identity is investigated post hoc, to determine the effect of the algorithm on consonant phonemes. This data will then be used to inform the design of the speech corpus and signal processing distortion measure in Chapter 5.

4.5.2 Design

This section states the experimental hypothesis and details considerations in the experimental design.

4.5.2.1 Hypothesis

H1: Different voices will suffer significantly different amounts of perceived distortion when speech is pitch-modified using TD-PSOLA.

4.5.2.2. Structure of Experiment

A listening test was designed to evaluate the amount of distortion introduced by TD-PSOLA into natural speech recorded by various voices to determine whether certain voices suffer different amounts of distortion. The independent variable *voice* had 4 levels; two female and two male, to allow comparisons between different gender voices. The dependent variable *distortion* was measured on a MOS scale (Mean Opinion Score) scale of 1 to 5. A MOS scale was chosen to allow the measurement of different amounts of distortion between the IV levels. The stimuli were assessed using the amount of perceived distortion as the criterion for evaluation where distortion was defined as *buzzy*, or *electronic sounding*. The participants rated each stimulus from 1 to 5 using the following definitions:

1 no perceived distortion
2 quite undistorted
3 distorted
4 quite distorted
5 very distorted

A within-subjects design was used; each participant rated the distortion at all 4 levels of the IV. A within-subjects design was used to reduce the effect of differences in levels of response between participants, reducing the number of participants required for the test to achieve the same power as a between-subjects design.

A statistical test was needed to compare the differences between the IV levels to determine whether different voices suffered significantly different amounts of distortion. The design was within-subjects, but the data were ordinal, so the assumptions of the parametric ANOVA were not met. A Friedman test (corrected for ties) was therefore performed on the nonparametric data. A two-tailed test was performed as the hypothesis was non-directional.

4.5.3 Stimuli

A string list given in Appendix B consisting of CVC (Consonant Vowel Consonant) syllables is representative of the initial consonants occurring in the English language (IPA, 1949). There are 24 consonant phonemes, but only 23 appear in the initial position. The list contains both meaningful and non-meaningful syllables. The initial consonant was varied in the CVC structure with the following VC structure kept constant. These stimuli would allow the effect of the TD-PSOLA algorithm on initial consonants to be evaluated, which was investigated post hoc. All stimuli were prepared with the following VC segment /{n/

e.g. b{n, k{n, dZ{n

The checked vowel /{/ was chosen as it had a low distortion rating from Experiment 1 and 2. The sonorant /n/ was chosen rather than the plosive used in previous experiments which caused problems with the recordings due to 'plosive pop' overloading the microphone.

All possible initial consonants were included to model the general effect of the algorithm in terms of introduced distortion (see Section 4.2.3. for an in depth discussion of the choice of stimulus type). The string list contained meaningful and non-meaningful syllables. To minimise any effect this may have on the results of the test, it was stressed to participants that the phonetic identities of the individual stimuli were not important and that only distortion, in the form of buzzyness, was to be evaluated. The string list contained only syllables that would not cause an emotive reaction in participants.

The CVC stimuli were recorded at the speakers' neutral pitch. The neutral pitch was determined by production of the "schwa" sound. The reference syllables were spoken at a steady rate and pitch, using the RPP recording technique to guide the pitch of the speaker. These pitches are shown below:

SPEAKER	NEUTRAL PITCH
Speaker 1 (Female)	220Hz
Speaker 2 (Female)	200Hz
Speaker 3 (Male)	130Hz
Speaker 4 (Male)	120Hz

To provide the stimuli for the experiment, the recordings were manipulated from their original pitch using the TD-PSOLA algorithm to 0, +1, +5, +10 and +15%, using the *mel* scale (see Section 4.2.3) to provide a linear relationship between fundamental frequency and pitch. This gave for each voice, the fundamental frequency levels shown in Table 4.7.

Voice	**% pitch manipulation**				
	0%	**1%**	**5%**	**10%**	**15%**
Voice 1 (F)	220Hz	223Hz	233Hz	246Hz	259Hz
Voice 2 (F)	200Hz	203Hz	212Hz	223Hz	235Hz
Voice 3 (M)	130Hz	132Hz	137Hz	144Hz	152Hz
Voice 4 (M)	120Hz	122Hz	127Hz	133Hz	140Hz

Table 4.7 Fundamental Frequency Values for Voices

The 0% control level was included to monitor the quality of the original recording and also to examine participants' possible bias towards a particular voice.

4.5.4 Procedure

The procedure and test conditions were identical to those in Experiment 1 (Section 4.2.6), and the same software was used to automate the test.

4.5.5 Participants

Twenty participants took part. All participants were university students or university staff due to the constraints of cost and availability. Participants ranged from 18-52 years of age, and of both male and female gender (13 male, 7 female). All reported to have, to the best of their knowledge, normal hearing. The participants were familiarised with test procedures and the definition of the test criterion prior to the test.

4.5.6 Results

Voice	% pitch manipulation	Median distortion (MOS scale rating)
Voice 1 (F)	0 %	1.83
	1%	2.00
	5 %	2.54
	10 %	2.76
	15 %	2.98
Voice 2 (F)	0%	1.43
	1%	2.46
	5%	3.22
	10%	3.61
	15%	3.52
Voice 3 (M)	0%	2.17
	1%	2.20

	5%	2.50
	10%	2.65
	15%	2.80
Voice 4 (M)	0%	1.78
	1%	2.11
	5%	2.65
	10%	3.13
	15%	3.41

Table 4.8 Summary Statistics: Distortion Rating for four Voices

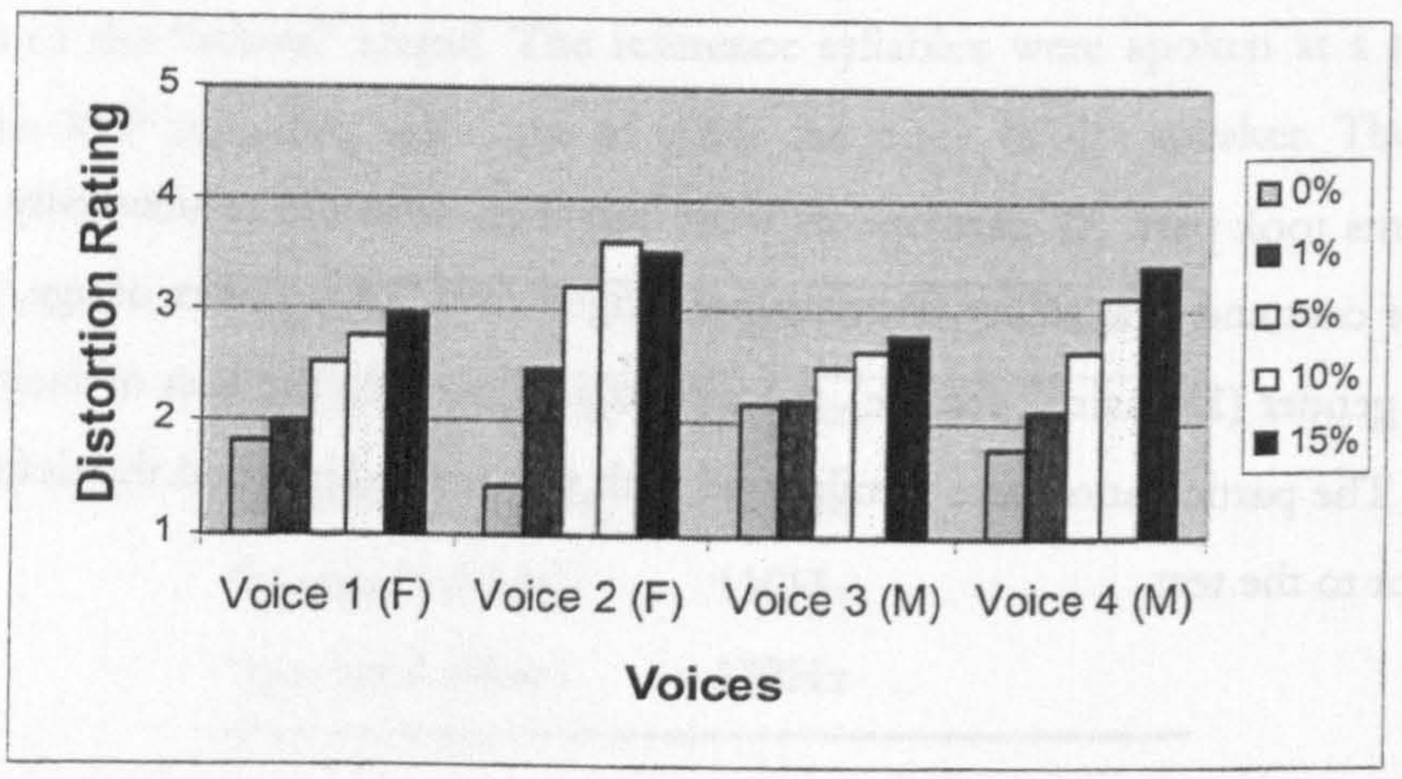


FIGURE 4.10 COMPARISON OF DISTORTION FOR FOUR VOICES AT 5 LEVELS OF PITCH MANIPULATION

Table 4.8 shows the summary statistics for each voice. Figure 4.10 illustrates the effect of pitch manipulation with TD-PSOLA on perceived distortion levels for the four voices. The most obvious effect is that all voices did not suffer similar amounts of distortion over the range of pitch modification evaluated. The 0% level represented participants' perception of the voices that had not been manipulated by the algorithm. Voice 2 was the participants' preferred voice when unmanipulated by the TD-PSOLA algorithm. Conversely, this voice suffered most, in terms of perceived distortion, when the algorithm was applied for even small 1% manipulations. Voice 3 was the least preferred when unmanipulated, but performed best when the algorithm was applied, appearing to be unaffected by manipulations of only 1%. Voice 1 appeared to be able to sustain small manipulations of 1% but suffered when 5% and larger manipulations were required. Voice 4 exhibited a greater distortion level at 1% than Voice 1, and a linear response between 5 and

15% manipulation. It may also be observed that in general more distortion was introduced the greater the manipulation for all four voices.

A within-subjects Friedman test (corrected for ties) was performed on the data to compare differences between the medians of the distortion ratings, using SPSS for Windows, Release 10.0.5. A significant effect was found indicating that voices may respond differently to the algorithm in terms of introduced distortion: for all averaged CVC stimuli and averaged pitch manipulation levels (χ_F^2 =42.2, df=3, N=20, $p<0.01$, *two-tailed test*).

In addition, a set of contrasts was performed at each pitch modification level, and all were found to be significant.

At 1% pitch manipulation (χ_F^2 =26.4, df=3, N=20, $p<0.01$).

At 5% pitch manipulation (χ_F^2 =32.1, df=3, N=20, $p<0.01$).

At 10% pitch manipulation (χ_F^2 =44.9, df=3, N=20, $p<0.01$).

At 15% pitch manipulation (χ_F^2 =39.1, df=3, N=20, $p<0.01$).

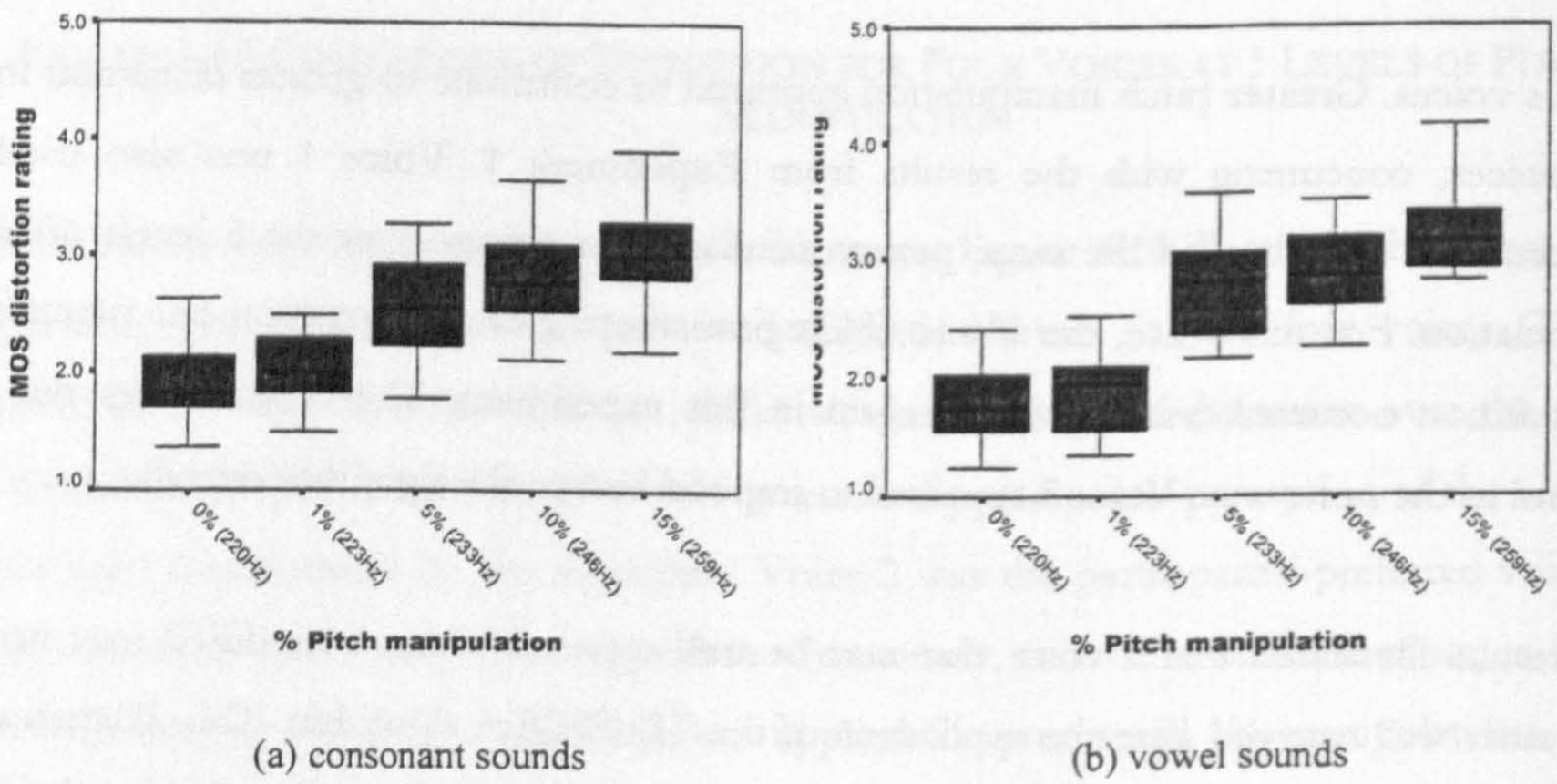
4.5.7 Discussion

1. Various voices. Greater pitch manipulation appeared to contribute to greater distortion in all four voices, concurring with the results from Experiment 1. Voice 1 was also used in Experiment 1 and showed the same pattern of distortion ratings over the 5 levels of pitch manipulation. For this voice, the 1% to 5% region where greatest distortion per percentage manipulation occurred was illustrated again in this experiment. The other voices did not respond in the same way. Voice 3 appeared to respond best to the algorithm overall.

   The results illustrated that a voice that may be well received when unmodified may not be necessarily well received after the application of the TD-PSOLA algorithm. This illustrates the importance of carefully selecting a speaker to record a database or speech corpus and testing the response of the voice to TD-PSOLA before the recording process is begun.

2. Male versus female voices. Literature reports (Moulines & Charpentier 1990, Blouin & Bagshaw 2000, Kortekaas & Kohlrausch 1997a) that female or higher fundamental frequency voices may suffer more from the application of the algorithm. This experiment did not show a clear indication of this, as the female voices responded second and fourth well of the four. Voice 1 (F) and Voice 3 (M) had both had experience in recording speech synthesis stimuli so were potentially more aware of criteria for high quality recording i.e. they were more akin to professional speakers. Voice 3 (M) performed better than Voice 1 (F) for the expert speakers, and Voice 4 (M) performed better than Voice 2 (F) for the novice speakers, so there is some evidence, although not conclusive, to suggest that female voices may suffer more.

3. Consonant versus vowels. A comparison may be made between the pattern of results of Experiment 1 and 4 (for Voice 1). Experiment 1 investigated the effect of the TD-PSOLA algorithm on stimuli where only the central vowel was altered, whereas Experiment 4 investigated the effect of the algorithm on initial consonants for the same Voice. The similar pattern of response to the algorithm over the 5 levels of pitch manipulation for both vowel and consonant seen in Figure 4.11 suggests that these stimuli are affected by TD-PSOLA in a similar manner. The largest % increase in distortion is still visible between the 1 and 5% pitch manipulation levels.



(a) consonant sounds (b) vowel sounds

FIGURE 4.11 BOXPLOT OF DISTORTION LEVELS IN CONSONANT AND VOWEL SPEECH SOUNDS

4. Stimulus identity and distortion. Experiment 1 showed that certain CVC stimuli were affected in terms of perceptible distortion more than others were. This was also evident in this experiment. The bar chart in Figure 4.12 shows the stimuli identities and their corresponding distortion levels for Voice 1.

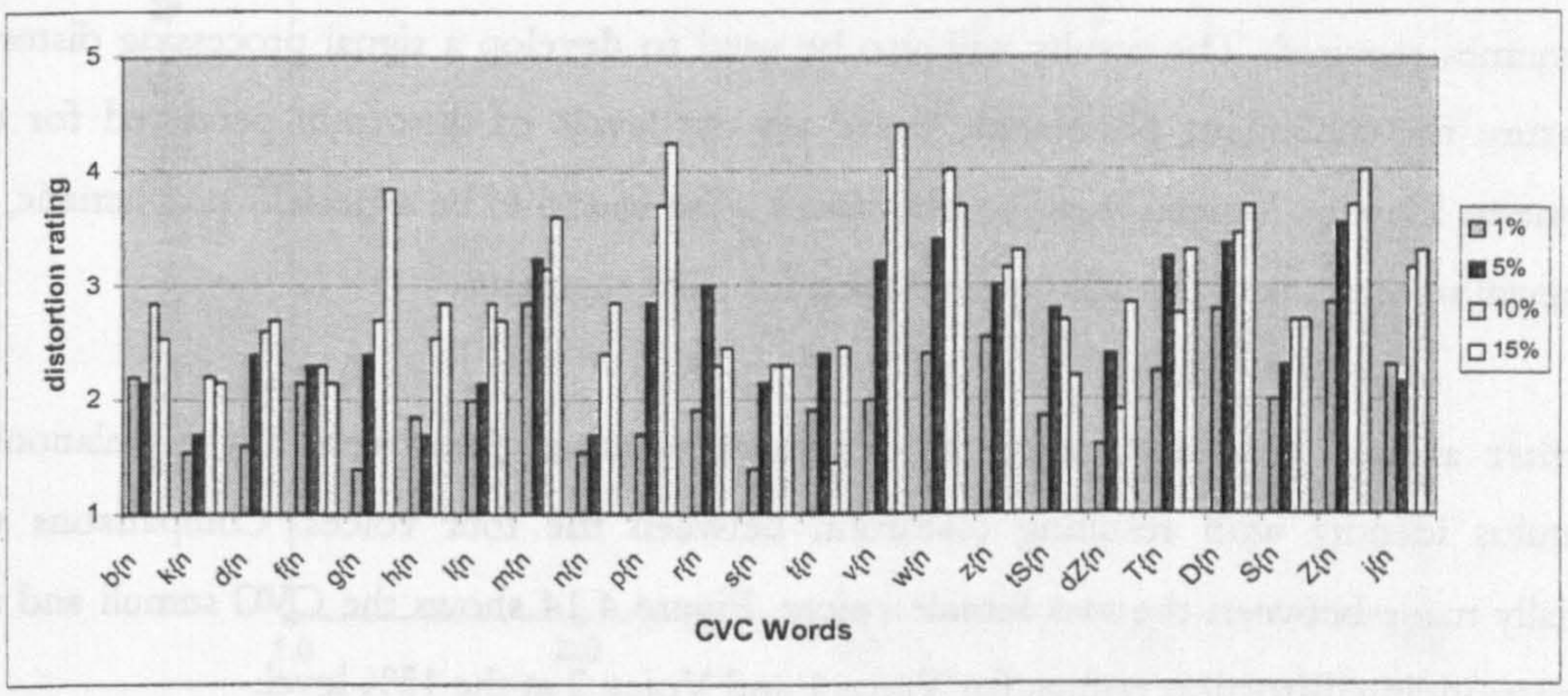


FIGURE 4.12 BARCHART OF VOICE 1 CVC STIMULI and DISTORTION

This is also shown in Figure 4.13, which illustrates large differences in the medians of distortion levels for individual stimulus identities and the spread of the data.

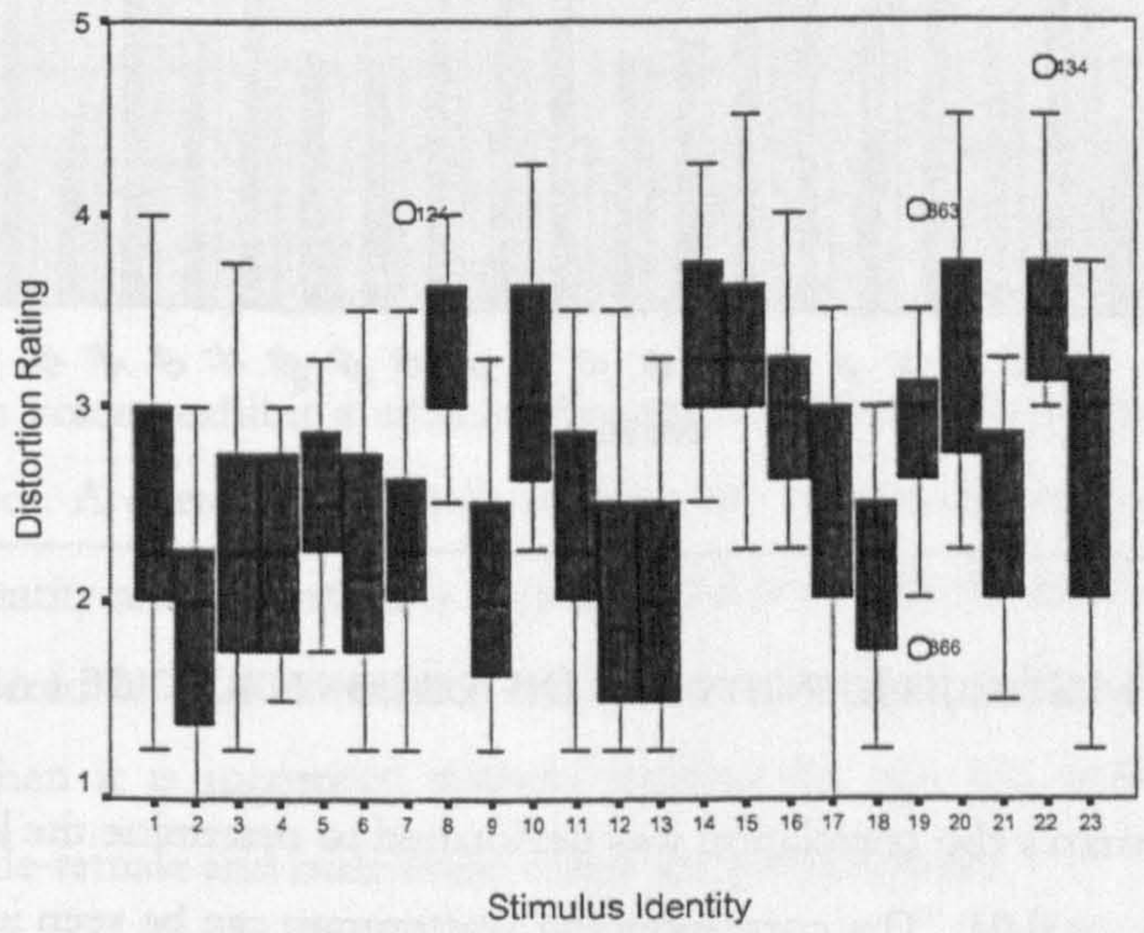


FIGURE 4.13 BOXPLOT OF VOICE 1 CVC STIMULI VERSUS DISTORTION

To examine this, a Friedman test was performed on the non-parametric data to see if there was significant variance in the distortion levels of the individual stimuli. A significant effect of phoneme identity was found (χ_F^2 =227.8, df=22, N=20, $p<0.01$). The data are also analysed in Chapter 5 to design the corpus by determining the balance of consonant phonemes required. The results will also be used to develop a signal processing distortion measure for consonant phonemes, based on the levels of distortion perceived for each phoneme identity. Voiced fricative phonemes were found to be especially problematic, and a special selection process will be developed for such segments.

Further analysis was undertaken to determine whether there were any correlations of stimulus identity with resulting distortion between the four voices. Comparisons were initially made between the two female voices. Figure 4.14 shows the CVC stimuli and their corresponding distortion ratings for Voice 1 and Voice 2 at the 15% level.

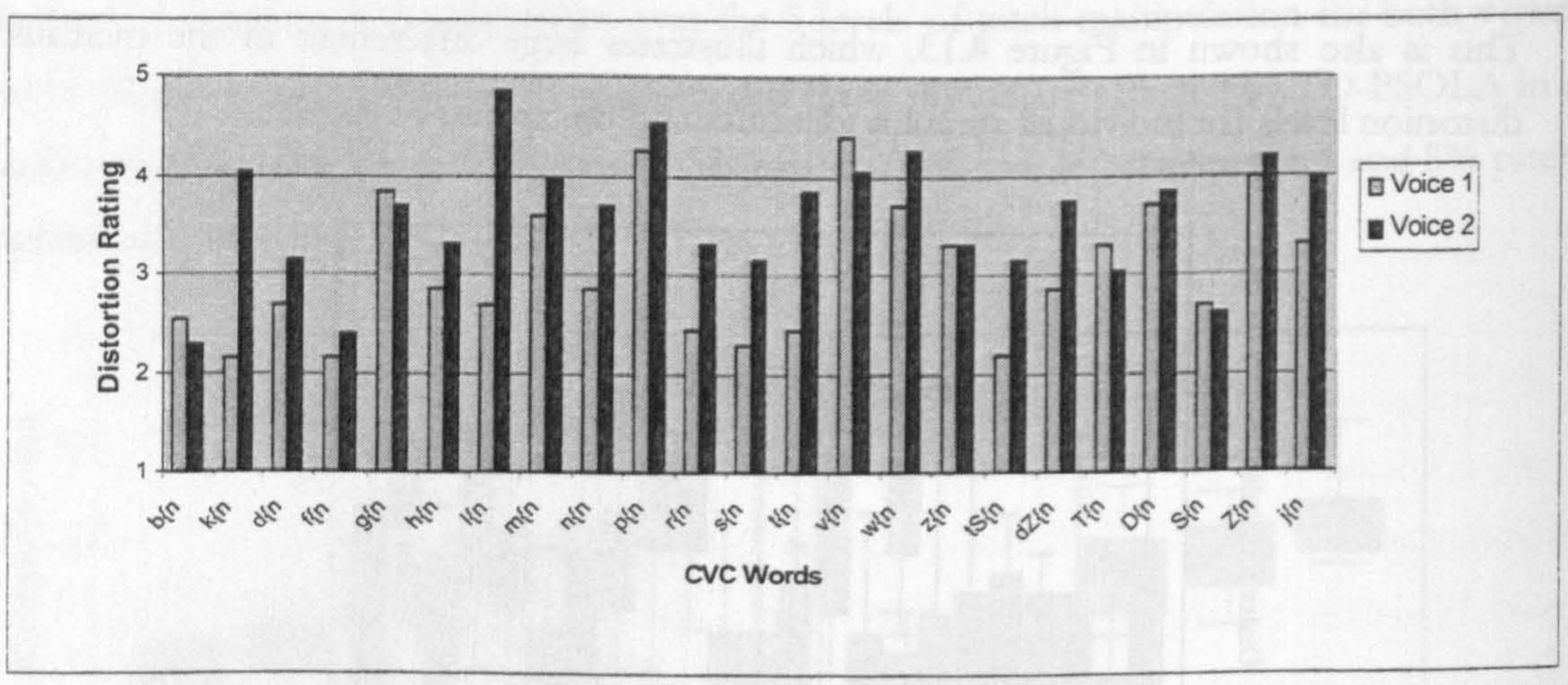


FIGURE 4.14 STIMULI IDENTITY AND DISTORTION FOR VOICE 1 AND 2

A one-tailed Spearman's rho correlation was performed to determine the level of correlation (rho=0.524, N=23, $p<0.05$). The corresponding scattergram can be seen in Figure 4.15.

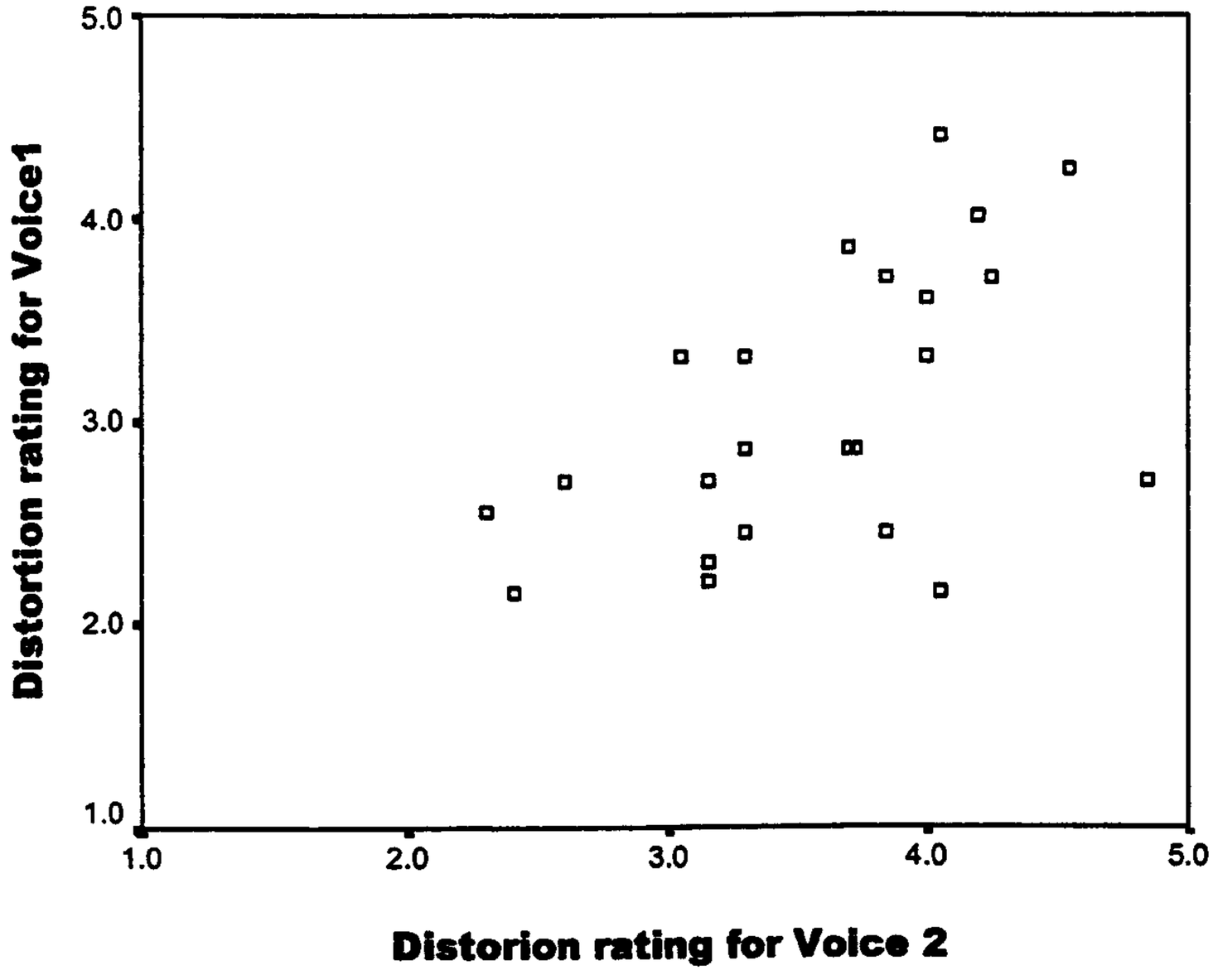


FIGURE 4.15 SCATTERGRAM OF STIMULI IDENTITY AND DISTORTION FOR VOICE 1 AND 2 AT 15% PITCH MANIPULATION

A correlation can be observed between the stimuli identity and distortion levels for Voice 1 and 2 at the 15% manipulation level, although Voice 1 appears less affected than Voice 2 by the TD-PSOLA algorithm overall.

The two male voices exhibit a similar phenomenon, although the male and female results together do not. A correlation matrix showing the relationship between voices in terms of consonant identity and distortion is shown in Table 4.9. As 64 tests were performed, the α level was lowered to 0.01 to reduce the chance of making a Type I error (accepting the hypothesis when it is incorrect) without making the test too conservative. Correlations between female-female and male-male voices are given in bold.

1% manip.	Voice1 (F)	Voice2 (F)	Voice3 (M)	Voice4 (M)
Voice1 (F)	X	**Not sig.**	Not sig.	Not sig.
Voice2 (F)	**Not sig.**	X	Not sig.	Not sig.
Voice3 (M)	Not sig.	Not sig.	X	**Not sig.**
Voice4 (M)	Not sig.	Not sig.	**Not sig.**	X

5% manip.	Voice1 (F)	Voice2 (F)	Voice3 (M)	Voice4 (M)
Voice1 (F)	X	**Not sig.**	Not sig.	Not sig.
Voice2 (F)	**Not sig.**	X	Not sig.	Not sig.
Voice3 (M)	Not sig.	Not sig.	X	**Not sig.**
Voice4 (M)	Not sig.	Not sig.	**Not sig.**	X

10% manip	Voice1 (F)	Voice2 (F)	Voice3 (M)	Voice4 (M)
Voice1 (F)	X	**Rho=0.46 *p<0.05***	Not sig.	Not sig.
Voice2 (F)	**Rho=0.46 *p<0.05***	X	Not sig.	Not sig.
Voice3 (M)	Not sig.	Not sig.	X	**Not sig.**
Voice4 (M)	Not sig.	Not sig.	**Not sig.**	X

15% manip	Voice1 (F)	Voice2 (F)	Voice3 (M)	Voice4 (M)
Voice1 (F)	X	**Rho=0.52 *p<0.05***	Not sig.	Not sig.
Voice2 (F)	**Rho=0.52 *p<0.05***	X	Not sig.	Not sig.
Voice3 (M)	Not sig.	Not sig.	X	**Rho=0.51 *p<0.05***
Voice4 (M)	Not sig.	Not sig.	**Rho=0.52 *p<0.05***	X

Table 4.9 Correlations of Voices at each Pitch Manipulation Level

At 1% and 5% pitch manipulation, none of the distortion levels for the individual stimuli identities for any of the voices are correlated. At 10% manipulation, the female voices were loosely correlated. Female voices become more correlated at 15%, with male voices also exhibiting a high correlation. Correlations appear to occur between voices that have similar neutral pitches, or of the same gender. These results indicate that data gathered for one voice

may not be generalised to other voices, and that the design of a speech corpus or signal processing distortion measure would have to be tailored to the individual voice.

5. Speaker Selection. Black & Lenzo (2000a) and Lowry (1999) suggest that the selection of the speaker has a great impact on the synthesis quality although it is difficult to determine acoustic aspects that differentiate good speakers from bad. Indeed it has been shown in this experiment that a preferred natural voice might not provide the best synthetic voice.

The Harmonics-to-Noise Ratio (HNR) can be used to measure the signal to noise ratio on a periodic signal where

$$HNR(dB) = 10\log_{10}(PeriodicEnergy / NoiseEnergy) \qquad \text{Eqn. 4.4}$$

Evidence suggests that TD-PSOLA cannot cope well for mixed–voice speech, which would be predominant in a 'hoarse' speaker, or one who has noise superimposed on the glottal waveform. A normal speaker producing an /{/ will typically have a HNR of 20dB, whilst a hoarse speaker will have a lower HNR.

In tests on production of the phone [{] Voice 2, which responded least favourably to TD-PSOLA, had a mean HNR of 14dB whilst Voice1 had a mean HNR of 20dB. Conversely, Voice 3 which responded best to the algorithm had a mean HNR of 14.5dB, whilst Voice 4, which was received less well, had a mean HNR of 17.2dB.

This parameter does not appear to be a conclusive indicator of the potential success of a voice. Whilst it is still impossible to predict the suitability of a voice for synthesis, much time and effort may be wasted in recording a corpus before knowing how the speaker's voice will respond to synthesis with algorithms such as TD-PSOLA.

4.5.8 Conclusions

Various voices appeared to suffer different patterns and levels of distortion when pitch manipulated using TD-PSOLA, supporting Lowry (1999). This has a great impact on the selection of voices for the recording of a speech corpus for synthesis with TD-PSOLA. The

preferred unmodified voice responded worst to the algorithm, which is in agreement with Syrdal *et al.* (1998a) highlighting the fact that voices for synthesis cannot be chosen purely on their unmodified characteristics. A parameter of signal-to-noise ratio was tested to determine whether it may be used to identify voices that may suffer greater distortion, but no consistent pattern was found.

Literature reports that higher f0 or female voices suffer more than lower f0 or male voices. The results from this experiment are inconclusive as only four voices were evaluated, although there was some supportive evidence that female voices may suffer more.

Patterns of distortion for consonant and vowel stimuli at each pitch modification level were compared and found to increase similarly with increasing pitch modification. Vowels may be affected more on average, which was to be expected as not all consonants are voiced; unvoiced sounds experience no modification during the pitch manipulation process.

Individual stimuli appeared to suffer different levels of distortion. A Friedman test was performed on the data and the effect of phoneme identity was found to be significant. These data will be analysed in Chapter 5 to inform the design of the speech corpus and the signal processing distortion measure for use with TD-PSOLA. Voiced fricative phoneme stimuli were found to be especially problematic, and a special selection process will be developed in Chapter 5 to reduce potential distortion for such segments.

4.6 Experiment 5: The Effect of Aspects of the Original Recordings on Distortion Levels in TD-PSOLA Pitch-Manipulated Speech

Abstract

An experiment was undertaken to determine the effect of aspects of the original recording of stimuli on resulting distortion levels. Ten participants were presented aurally with 2 sets of randomly ordered CVC stimuli. The first set consisted of 4 recording versions of 6 CVC syllables with a varying central vowel. The second set consisted of 4 recording versions of 7 CVC syllables with varying initial consonants. Each set was manipulated to a standard pitch of 1, 5, 10 and 15% from the neutral pitch, giving a total of 96 stimuli for the first set and 112 stimuli for the second set. Participants indicated whether or not any distortion was perceived for each of the stimuli. Perceived distortion for each version varied in each set indicating that aspects of the original recording had a high impact on the results. The recording parameter of "waveform asymmetry" that had been identified at being potentially problematic during previous experiments was found to have a significant effect on resulting distortion levels for 9 of the 13 sets of CVC syllables.

4.6.1 Introduction

During the preparation of stimuli for previous experiments, it became apparent that certain recordings of speech sounds suffered anomalously large amounts of distortion. Where identified, these were removed from the stimuli set and re-recorded.

Such anomalously large distortions seemed to occur when the time-domain waveform exhibited "asymmetry" in the pressure fluctuations. Such pressure changes are mainly caused by the production of plosives as illustrated in the waveform of the word "cart" in Figure 4.16. This is often known as plosive distortion where the production of a plosive may generate large sound pressure levels (SPLs) that can overload the microphone. The effect may be minimised by including a pop filter inside the microphone cap or by using a windscreen, at the expense of the high frequency response. An alternative approach places the microphone at 45% degrees to one side of the speaker's mouth.

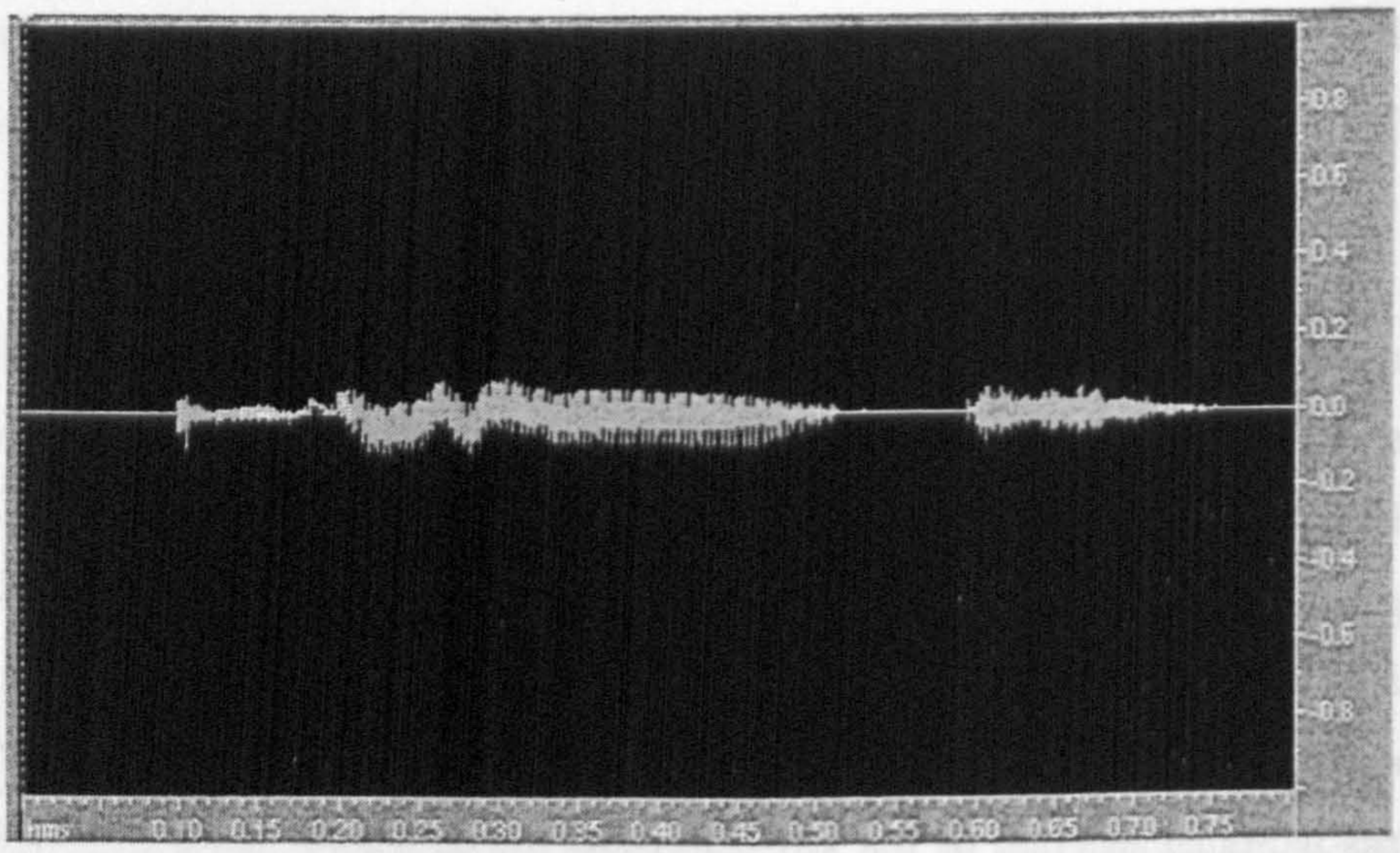


FIGURE 4.16 WAVEFORM OF THE WORD "CART" WITH ASYMMETRY

The experiment analysed whether this may have contributed to this anomalous distortion and whether there may have been any additional factors.

4.6.2 Design

This section states the experimental hypothesis and documents design considerations.

4.6.2.1 Hypothesis

H1: Aspects of the original recordings of the stimuli will have a significant perceived effect on the resulting distortion levels.

4.6.2.2 Structure of the experiment

A listening test was undertaken to investigate aspects of the original recording that may lead to anomalously large distortion levels. The independent variable *recording aspect* had two levels of "asymmetry" and "symmetry".

The dependant variable was measured by indicating whether distortion was present or not. A yes/no response was adopted to determine if any distortion was present in the modified stimuli. What was of interest in this experiment was whether certain recordings suffered significantly more distortion.

A test was required to determine whether stimuli exhibiting asymmetry suffered significantly more distortion than ones that were more symmetrical. The design was within-subjects with nominal data, so a non-parametric McNemar's test of change was performed.

4.6.3 Stimuli

The stimuli were divided into two sets to keep the data sets and corresponding lengths of test runs to a manageable size. The first set consisted of 6 different CVC syllables with varying central vowels. The three checked vowels /I/, /{/ and /Q/ were evaluated with their free counterparts /aI/, /A:/ and /u:/. The second set consisted of 7 different CVC syllables with varying initial consonants; a voiced plosive /d/, an unvoiced fricative /s/, a voiced fricative /D/, an affricative /tS/, a nasal /n/, a liquid /r/ and a glide /j/, to evaluate a phoneme from each category when grouped according to manner of articulation. Four versions of each CVC syllable were recorded. Each recording was made at a different sitting to maximise differences in voice quality for each session. Two stimuli were chosen which exhibited asymmetry in the time-domain waveform, and two were chosen with more symmetrical waveforms, to provide the four recording versions of each CVC syllable.

The CVC stimuli were recorded at the neutral pitch of the speaker (220Hz) using the RPP recording technique (Vine *et al.*, 1999). These were then pitch manipulated to the target fundamental frequencies of 223, 233, 246 and 259Hz, corresponding to 1, 5, 10 and 15% modification to provide the TD-PSOLA manipulated stimuli.

4.6.4 Procedure

Participants were familiarised with the procedure and criterion using a set of typed instructions (Appendix C). They were presented with the CVC stimuli via headphones and then made a

judgement as to whether any distortion was present in the stimuli using the software interface (Appendix A). The two test runs of 96 and 112 stimuli with a delay of approximately 5 seconds between each stimulus presentation lasted 8 and 10 minutes respectively.

4.6.5 Participants

Ten participants took part. The somewhat restricted sample population consisted of university staff or students, ranging from 20-56 years of age and from both genders (6 male, 4 female). All had self-reported normal hearing.

4.6.6 Test Conditions

Output Device: headphones
Acoustic Environment: quiet office
Noise Levels: minimum background noise
PC: Pentium, 133 MHz
Speech Spec:

- Voice: J. Longster
- M/ F: F
- Sampling Frequency: CD quality (44100Hz)
- Speech Units: CVC syllables
- Algorithm: TD-PSOLA, Praat Software (Boersma & Weenink, 1999).

4.6.7 Results

4.6.7.1 Vowel results

Version	k{t	kIt	kQt	kA:t	kaIt	ku:t
1	17.5%	15%	2.5%	42.5%	35%	57.5%
2	35%	17.5%	5%	20%	35.5%	62.5%
3	50%	87.5%	47.5%	47.5%	30%	37.5%
4	20%	52.5%	25%	77.5%	77.5%	25%

Table 4.10 Summary Statistics: % Distortion Detection for 4 Versions of 6 CVC Syllables

Table 4.10 shows the summary statistics, for 4 recording versions of the 6 CVC syllables, of the percentage distortion detection at all pitch manipulation levels. Due to the variability for each version, the aspects of the original recording can be seen to have a great impact on the success of the stimuli.

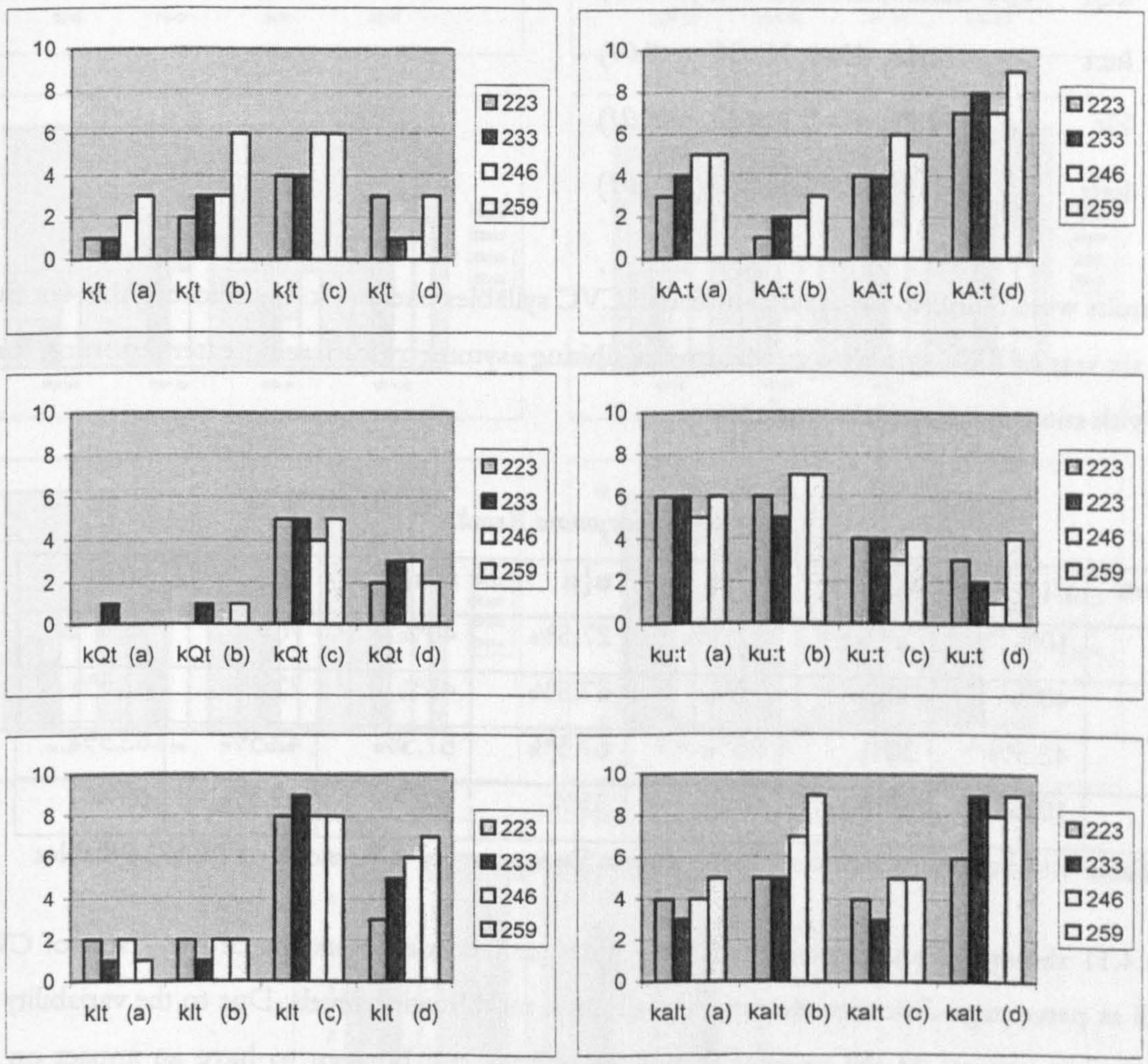


FIGURE 4.17 FOUR VERSIONS OF 6 VOWEL STIMULI AND DISTORTION DETECTION LEVELS

Figure 4.17 illustrates the importance of aspects of the original recording; each of the four versions of each CVC syllable is shown with the number of distortion detections on the y-axis for each % pitch modification.

The stimuli were grouped into those that exhibited asymmetry in their time-domain waveform and those that did not, and a McNemar's test of change was performed on the data.

- k{t (χ_2 =0.97, df=1, N=80, *p>0.05*)
- kA:t (χ_2 =12.90, df=1, N=80, *p<0.01*)
- kQt (χ_2 =20.83, df=1, N=80, *p<0.01*)
- ku:t (χ_2 =11.26, df=1, N=80, *p<0.01*)
- kIt (χ_2 =39.83, df=1, N=80, *p<0.01*)
- kaIt (χ_2 =18.58, df=1, N=80, *p<0.01*)

The results were found to be significant for all CVC syllables except /k{t/ indicating that for five of the six sets of CVC syllables, waveforms exhibiting asymmetry suffered greater distortion than those with more symmetrical waveforms.

4.6.4.2 Consonant Results

Version	d{n	s{n	tS{n	n{n	r{n	j{n	D{n
1	10%	25%	17.5%	27.5%	40%	70%	90%
2	40%	50%	55%	67.5%	45%	72.5%	87.5%
3	42.5%	30%	25%	67.5%	67.5%	42.5%	62.5%
4	10%	15%	60%	35%	72.5%	42.5%	65%

Table 4.11 Summary Statistics: % Distortion Detection for 4 Versions of 7 CVC Syllables

Table 4.11 shows the summary statistics for the four recording versions of the 7 sets of CVC stimuli as percentage distortion detection at all pitch modification levels. Due to the variability of each version, aspects of the original stimuli recordings can be seen to have an impact on its success, when subjected to TD-PSOLA pitch modification.

Figure 4.18 illustrates the impact of aspects of the original stimuli recordings; each of the 4 versions of each CVC syllable is shown plotted against the number of distortion detections for each modification level.

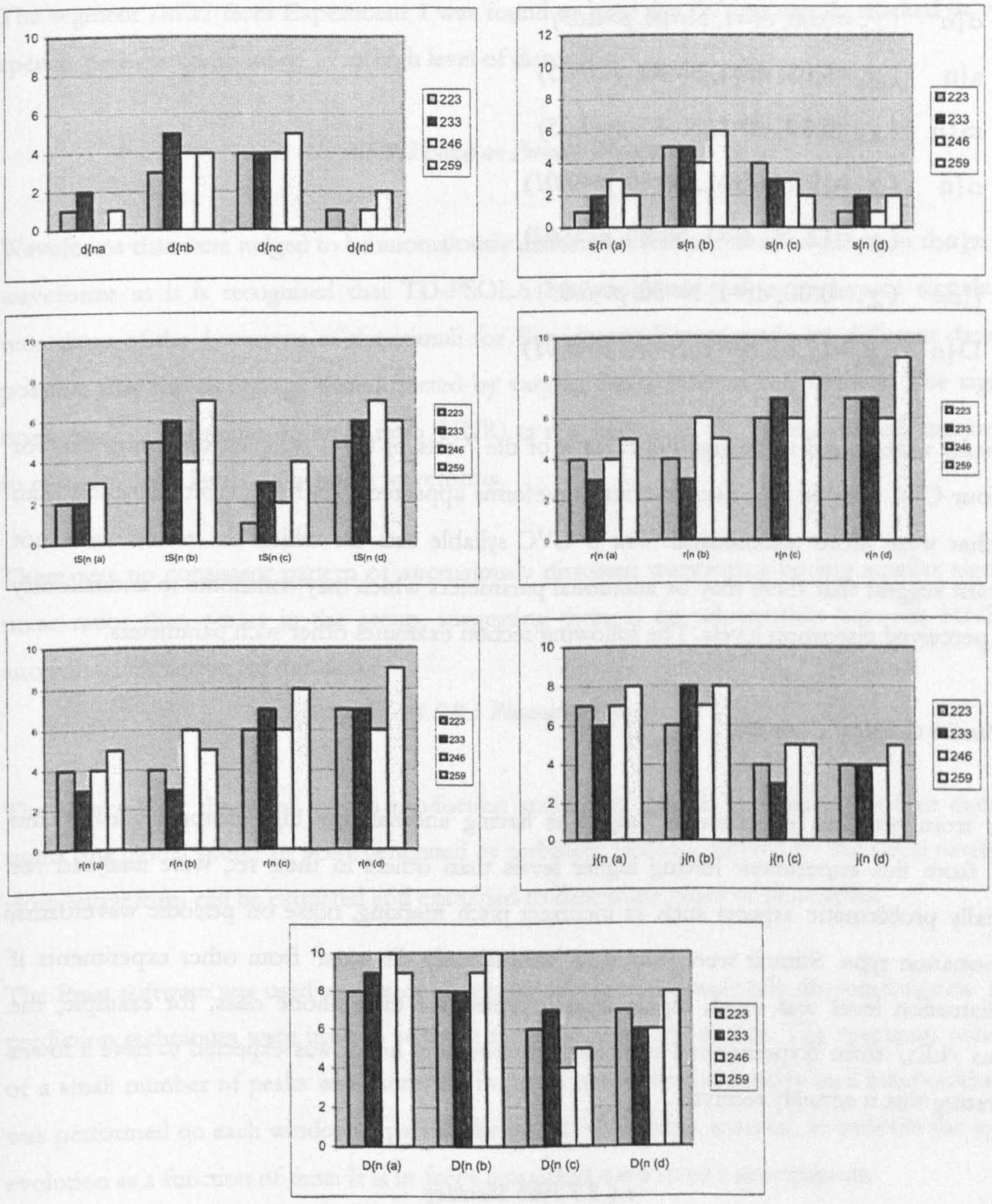


FIGURE 4.18 FOUR VERSIONS OF 7 CONSONANT STIMULI AND DISTORTION DETECTION LEVELS

The stimuli were grouped according to whether or not they exhibited asymmetry in their time-domain waveform, and a McNemar's test of change was performed on the data.

- d{n (χ_2 =0.00, df=1, N=80, *p>0.05*)
- s{n (χ_2 =8.03, df=1, N=80, *p<0.05*)
- tS{n (χ_2 =0.32, df=1, N=80, *p>0.05*)
- n{n (χ_2 =19.12, df=1, N=80, *p<0.01*)
- r{n (χ_2 =12.25, df=1, N=80, *p<0.01*)
- j{n (χ_2 =0.00, df=1, N=80, *p>0.05*)
- D{n (χ_2 =11.61, df=1, N=80, *p<0.01*)

The results were found to be significant for 4 of the 7 sets of CVC syllables, indicating that for these four CVC syllable sets, asymmetrical waveforms appeared to suffer greater distortion than those that were more symmetrical. The 3 CVC syllable sets for which the results were not significant suggest that there may be additional parameters which may contribute to anomalously higher perceived distortion levels. The following section examines other such parameters.

4.6.8 Possible Causes of Distortion

Stimuli from previous experiments judged as having anomalously high distortion levels and stimuli from this experiment having higher levels than others in their set, were analysed for potentially problematic aspects such as incorrect pitch marking, noise on periodic waveforms, and phonation type. Stimuli were judged as 'anomalously distorted' from other experiments if their distortion level was much higher than expected for their phone class, for example, the stimulus /kEt/ from Experiment 1 is a checked vowel and hence was expected to have a lower MOS rating that it actually received.

4.6.8.1 Pitch Marking

Pitch marking should be synchronised on glottal closure and be temporally consistent for minimal distortion (Kortekaas & Kohlrausch 1997b, Moulines & Charpentier 1990). The position of the pitch marking of anomalously distorted segments was checked using the Praat software. None of the stimuli were found to have offset pitch marks (jittered or shimmered pitch marking), suggesting there was no relationship between offset pitch marking and distortion for this data.

The segment /kEt/ from Experiment 1 was found to have the /t/ incorrectly marked as voiced speech, perhaps contributing to its high level of distortion.

4.6.8.2 Noise on Periodic Waveforms

Waveforms that were judged to be anomalously distorted were analysed for noise on the periodic waveforms as it is recognised that TD-PSOLA has problems with mixed-voice signals. The recordings of the 4 versions of the stimuli for Experiment 5 were made on different days; it is possible that the recordings were affected by varying hoarseness of the speaker. The signal-to noise ratio, or harmonics-to-noise ratio (HNR) as it is known in the Praat software, can be used to measure noise levels on periodic waveforms.

There was no consistent pattern of anomalously distorted waveforms having smaller signal-to-noise ratios than others in the group, suggesting there is no relationship between HNR and anomalous distortion for this data.

4.6.8.3 Phonation Type

The source-filter theory of speech production states that speech is composed of an excitation signal (the glottal source or noise generated by turbulent airflow), filtered by the vocal cavity. The glottal waveform can be extracted and examined to determine types of phonation.

The Praat software was used to extract the glottal source of anomalously distorted signals. Linear prediction techniques were used to separate the filter from the source. The spectrum consisting of a small number of peaks was estimated in terms of centre-frequencies and bandwidths. This was performed on each windowed part of the signal (short-term analysis) to provide the spectral evolution as a function of time. It is in fact a smoothed version of a spectrogram.

The peaks are associated with the formant resonances of the vocal tract. Analysis has shown that female voices typically have five formants between the ranges of 0 to 5500Hz. To implement this band-limiting, the original signal was resampled to 11kz.

LP analysis was then performed on this resampled sound. The LP Burg algorithm applied is described in Press *et al.* (1992). The result was a LP time function with 10 linear prediction coefficients in each time frame. To extract the source, the resampled sound and the LP object were inverse filtered. Given the filter (the LP coefficients) and the output (the resampled sound), the input (the glottal source) could be reconstructed. This signal represents everything in the speech signal that cannot be attributed to the resonating cavities. Figure 4.19 shows a glottal source for the production of /a/.

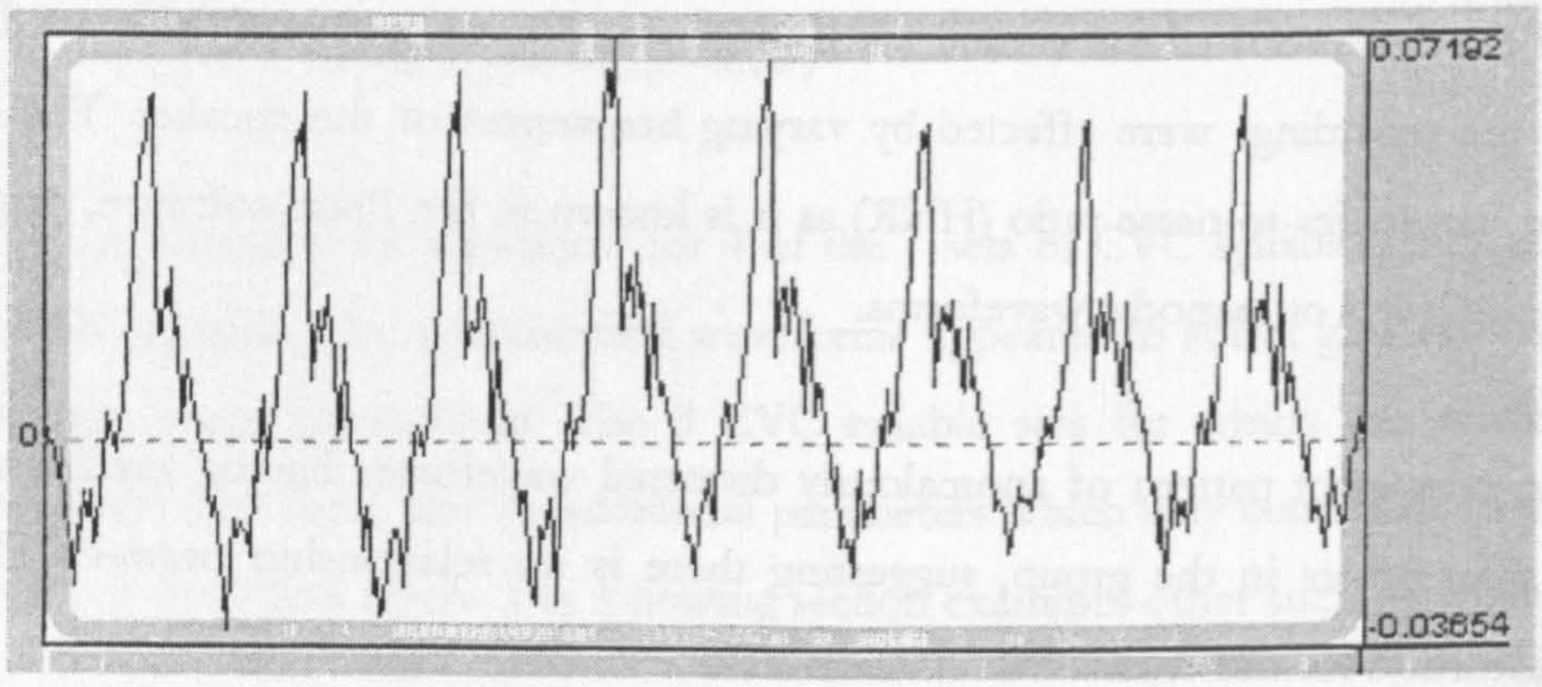


FIGURE 4.19 GLOTTAL SOURCE FOR THE PRODUCTION OF /A/

Glottal source extraction was used to determine phonation types e.g. breathy voice (characterised by longer fall phase, more symmetrical pulse and higher stochastic component) or creaky voice (also known as 'vocal fry', characterised by a very low f0, short rise time and irregular pulses), which are both known to be problematic for TD-PSOLA.

There were no stimuli exhibiting breathy voice, hence no evidence to suggest this may have been a factor in the large levels of distortion. One stimulus suffered creaky voice, having incorrectly low pitch-marking, which may have contributed to the large level of distortion.

4.6.9 Conclusions

During the course of the experiments, it was apparent that the success of the TD-PSOLA algorithm relied heavily on certain aspects of the original recordings of the speech to be manipulated. A parameter of waveform asymmetry that appeared to lead to anomalously high distortion levels was identified. This parameter was found to be statistically significant for 9 of

the 13 sets of CVC syllables, suggesting that there may be other factors additionally responsible for this anomalous distortion.

Other possible parameters of incorrect pitch marking, low HNR, and phonation type were investigated. There was no evidence to suggest that a low HNR, offset pitch-marking, or breathy phonation, may lead to the anomalously high distortion levels of the stimuli analysed. One stimulus suffered incorrect pitch-marking (unvoiced speech marked incorrectly as voiced), and a second suffered creaky voice phonation, both of which may have contributed to their unexpectedly high distortion levels.

It is possible that a combination of waveform asymmetry, creaky phonation type and incorrect pitch marking may have led to some of the anomalous distortions encountered during these experiments, and segments exhibiting such parameters should be removed from the speech corpus.

4.7 Investigative Experiments Conclusions

A set of investigative experiments was undertaken to investigate the effects of the TD-PSOLA algorithm on natural speech stimuli. Conclusions drawn from these experiments will be used to guide the development of a framework for reduced distortion when using TD-PSOLA for prosody modifications. The framework consists of a speech corpus design, a signal processing distortion measure, and a special selection process for voiced fricative segments.

The first experiment investigated the effect of pitch manipulation with TD-PSOLA on distortion levels in CVC stimuli. Significantly greater distortion was perceived for increasing levels of pitch modification. Modifications of as low as 1% were found to introduce distortions into the speech signal for certain phoneme stimuli. This indicates than in an ideal situation, the use of signal processing should be avoided completely, or more realistically, kept to a minimum. The individual phoneme identity was investigated post hoc and found to have a significant effect on resulting distortion levels. This will be used to determine the content of the speech corpus and develop the signal processing distortion measure.

The second experiment investigated the effect of the TD-PSOLA algorithm on distortion levels in positively and negatively manipulated speech. Over the range evaluated of +/-8% manipulation, similar distortion levels were found for both directions of manipulation. This result implies that there is no advantage, in terms of less distortion, in selecting segments that have either a higher or lower pitch than the target value. In fact, the correlation between individual stimulus distortion levels for both positive and negative directions was less significant indicating that individual stimuli may not respond similarly to positive and negative modifications. This requires further experimentation, and may be used to tune the design of the corpus and distortion measure, and is therefore discussed as further work in Section 7.2.

The third experiment investigated the effect of pitch manipulation with TD-PSOLA on distortion levels in synthetic speech at the sentence level. Two inventories were evaluated, the first requiring greater pitch modification than the second. There was no significant difference in distortion levels between stimuli synthesised from these inventories. This lack of significance may

be due to the large amount of TD-PSOLA modification that segments from both inventories underwent to achieve the target values. More importantly, the modification of static pitch contours at the word-level, investigated in previous experiments, may have provided the worst-case scenario. Effects seen in these experiments may not be as evident at the sentence level that consist of mainly dynamic contours. This suggests that the use of the new corpus design and signal processing distortion measure for sentence-level synthesis may have less effect in terms of reducing distortion when compared to a standard synthesis framework as the perception of distortion may be reduced for dynamic synthetic speech even using a standard approach.

The fourth experiment investigated the effect of TD-PSOLA pitch manipulation on distortion levels in speech for various voices. Different voices appeared to suffer different patterns and levels of distortion, which would be of great importance when selecting voices for recording a speech corpus to be used in conjunction with TD-PSOLA. The results were inconclusive as to whether female (higher f0) voices suffered more than male (lower f0) voices, although there was some evidence to suggest female voices suffered more. Again, the individual phoneme identity was investigated post hoc and found to have a significant effect on resulting distortion levels. This data will be used to determine the content of the speech corpus and develop the signal processing distortion measure.

During these experiments, it became apparent that aspects of the original recordings had a large effect on the success of the stimuli, when pitch modified using TD-PSOLA. A parameter of "waveform asymmetry" was isolated as a potential problem for TD-PSOLA. If such a parameter was found to contribute to perceived distortion, segments that exhibited this phenomenon could be removed from the corpus. Experiment 5 investigated this, and the results were found to be significant for 9 of 13 CVC syllable sets. Some additional possible parameters were suggested, although further investigation was outside the scope of this work.

To conclude, a number of issues raised in Chapter 2 have been addressed, although many other issues have been identified. These will be discussed in the recommendations for further work in Chapter 7.

The following chapter analyses the data collected during these experiments in an attempt to model the occurrence of perceived distortion, and analyses individual speech segment properties which may contribute to increased perceptible distortion levels. The results of this analysis will then be used to design the framework for use with the TD-PSOLA algorithm to allow a speech output with reduced perceptible distortion.

Chapter 5. Distortion Modelling and Development of a Novel Corpus Design and Signal Processing Distortion Measure

5.1 Introduction

Chapter 4 documented experiments undertaken to evaluate the effect of TD-PSOLA, in terms of perceived distortion, for pitch-modification of speech. The data collected during these experiments are analysed in this chapter in an attempt to model the occurrence of such distortion. It is impractical to carry out experiments that analyse every possible piece of data in every possible situation so this analysis relies on patterns of co-occurrence, or correlations, between sets of data. Parameters are identified and analysed to determine their effect on the amount of distortion introduced when speech is pitch-modified using TD-PSOLA.

The data analysis will be used to develop a framework for producing speech with reduced distortion when TD-PSOLA is applied for pitch modifications. To this end, existing speech corpus designs and unit selection processes are critically reviewed to examine how the development of such a framework could improve existing approaches. The concept of designing a corpus tailored to the use of TD-PSOLA as the signal-processing algorithm for imposing the target prosody on speech, is discussed. The corpus is not phonetically balanced, but balanced to the requirements of TD-PSOLA, containing more versions of adversely affected segments. The aim is to minimise the distortion of the output by allowing such segments to be selected with pitch contours closer to the target contours.

The concept of a signal processing distortion measure for inclusion in a unit selection process is also presented. This measure would enable speech segments that suffer most from pitch modification using TD-PSOLA to be identified. Then, using this measure as part of a standard unit selection process, optimal segments could be extracted from the corpus for TD-PSOLA modification with reduced distortion.

Previous experimentation has shown that segments of different phonetic types suffer different perceived amounts of distortion with voiced fricatives being especially problematic. A special

selection process for such segments is also developed as part of the framework to reduce this potential distortion.

5.2 Distortion Models

Experimental results suggested that the amount of distortion introduced when speech is TD-PSOLA pitch-modified varies for different phonetic identity speech segments. Chapter 2 discussed the inherent characteristics of speech sounds, when grouped according to manner of articulation. The possible effect of these characteristics on the perception of distortion is discussed below, and then the data will be analysed to determine their actual contribution to perceived distortion levels when the segments have been pitch-modified by TD-PSOLA.

- Duration: the effect of duration will be investigated as the duration of sound is known to affect perception – tones lasting less than one second cannot be viewed as infinite and auditory sensitivity is altered for durations much less than one second, such as 10ms or less (Gelfand, 1998).
- Intensity: auditory sensitivity is governed by intensity levels or loudness (Turner *et al.*, 1989). It will be investigated whether the mean intensity of individual phonemes may contribute to the perception of distortion.
- First formant (f1) value: the f1 value has been cited as having an effect on the perception of distortion (Blouin & Bagshaw, 2000). F1 is dependent on the size of the volume behind the tongue elevation, which increases as the elevated part of the tongue moves forward, lowering f1. F2 depends on the volume in front of the tongue elevation. Lip rounding lowers the first two formants by reducing the size of the mouth opening. Averagely lower f1 stimuli have been found to suffer less distortion than higher f1 stimuli and this will be investigated for this data.
- Voiced/unvoiced/mixed-voice status: The voiced/unvoiced/mixed-voice status of the speech sound has a large effect on the success of the algorithm (Moulines & Charpentier, 1990). Consonants may be voiced/unvoiced or a mixture of both. Vowel sounds are always voiced so this status will be analysed for consonants only. For pitch modification, Praat performs no manipulation of unvoiced parts of speech. Voiced speech and mixed

voiced/unvoiced speech undergo Short–Term (ST) repetition when increasing pitch, which may cause audible distortions.

The parameters common to phone classes grouped according to manner of articulation that are identified as perhaps contributing to increased distortion levels, can then be used to suggest a design for the speech corpus and signal processing distortion measure.

5.2.1 Vowels

5.2.1.1 Analysis of Vowel Data

Initially the stimuli from Experiment 5, consisting of four versions of each of six CVC syllables, were analysed. Only the vowel sound was altered in each of the six CVC structures. The stimuli were analysed in terms of

1. Phonetic type e.g. checked/free, monothong/diphthong
2. Duration
3. Intensity
4. F1 value

The averages of duration, intensity and f1 value were taken for each of the four versions of speech sounds. Whilst carrying out the analysis, there appeared to be a relationship between the shape of the f1 contour and resulting distortion, so f1 shape was included also. All four versions were used, because the waveform asymmetry parameter, as a potential cause of anomalously large distortion levels, was not statistically significant for all CVC sets. The results are summarised in Table 5.1. The % distortion detection is given in the final column, averaged for all four versions across each pitch modification level of 1, 5, 10 and 15%.

CVC identity	phonetic type	dur (s)	intensity (dB)	f1 value (Hz)	f1 shape	% distortion detection
kQt	checked	0.17	-22.6	690	Flat	20.0%
k{t	checked	0.18	-24.9	732	Flat	30.6%
kIt	checked	0.17	-22.0	530	Flat	41.9%
ku:t	free (mono)	0.28	-25.2	375	Flat	45.6%
kA:t	free (mono)	0.31	-23.5	770	Flat	46.9%
kaIt	free (dip)	0.38	-25.9	832 – 446	formant transition	50.0%

Table 5.1 Vowel Data (Experiment 5)

5.2.1.2 Results of Vowel Data

1. Phonetic type: The vowels can be grouped in terms of perceived distortion level into checked vowels and free vowels. The checked vowels performed well, whilst the free vowels suffered more from the application of the algorithm. This is illustrated in Figure 5.1, which shows the % distortion detection for the two groups of vowels. The free vowels can be further divided into monothongs and diphthongs. The data in Table 5.1 suggests that monothongs are more successful than diphthongs.

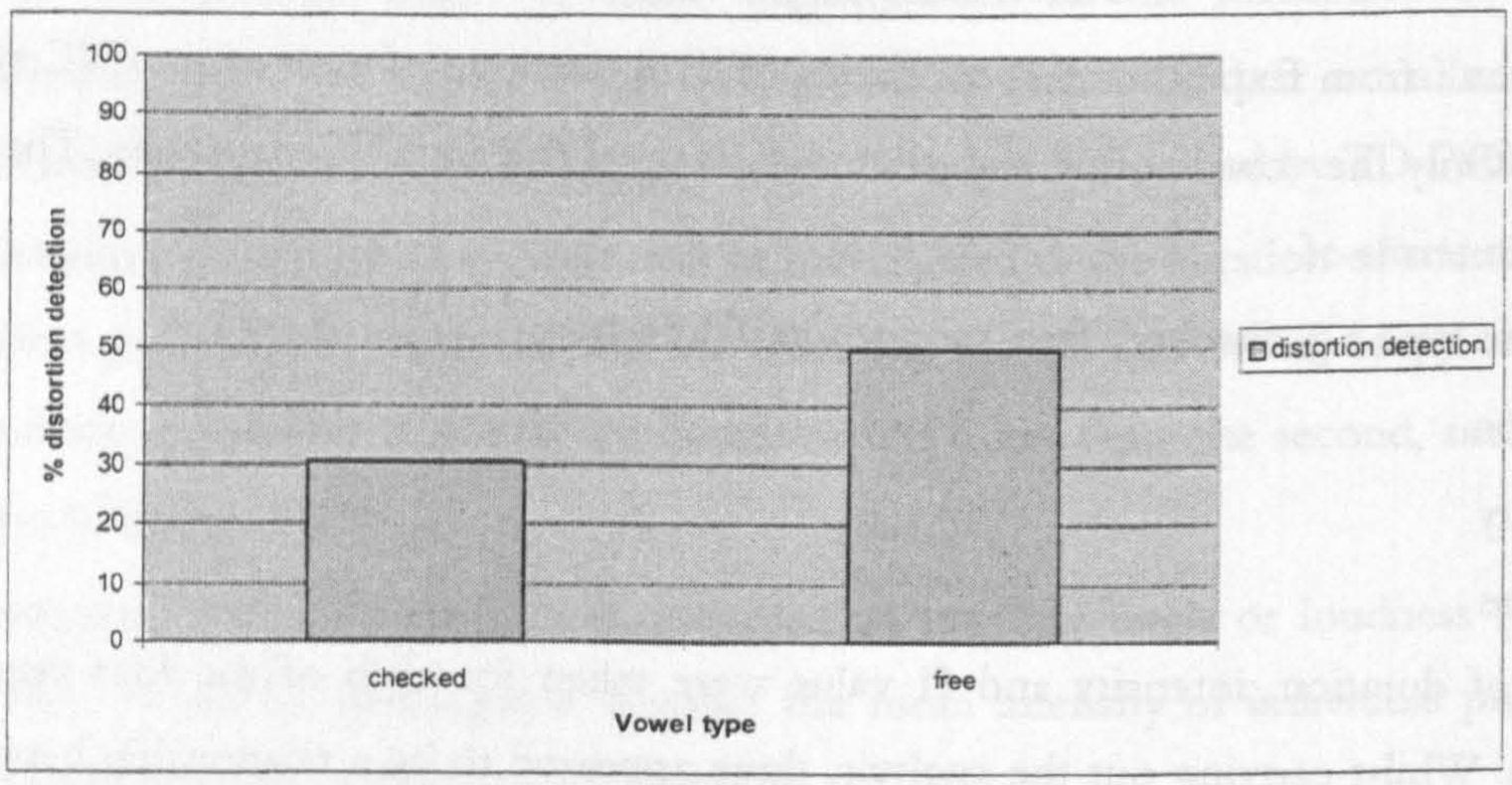


FIGURE 5.1 % DISTORTION DETECTION FOR CHECKED AND FREE VOWELS

2. Duration: the checked vowels, which are characterised by shorter durations, respond well to the algorithm, followed by the monothongs, which have a longer duration, and finally the diphthong with the longest duration.
3. Intensity: the mean intensity was calculated using the Praat software over the vowel part of the CVC stimuli. There is no evidence of a correlation between intensity and distortion for this data.
4. F1 value: vowel stimuli with lower f1 values were found to perform neither consistently better nor worse than those with higher f1 values. There is no evidence of a correlation between f1 values and distortion for this data.
5. F1 shape: the vowel sounds with flat f1 contours suffered less distortion than those with an internal formant transition such as often found in diphthongs.

From these results, it is hypothesised that the distortion detection levels are dependent upon the phonetic type of the vowel, its duration and f1 shape. Intensity and f1 value do not have an effect on distortion for this data.

5.2.1.3 Generalisation of Results to other Vowel Data

The data from Experiment 1 were analysed to determine whether these results can be generalised to other data collected for the same Voice. Table 5.2 shows the data from Experiment 1 in order of increasing distortion rating. The distortion ratings were calculated for each CVC syllable, averaged over each pitch manipulation level of 1, 5, 10 and 15%.

vowel	**phonetic type**	**duration (s)**	**Intensity (dB)**	**f1 value (Hz)**	**f1 shape**	**distortion rating**
kQt	checked	0.18	-29.7	822	Flat	1.82
kIt	checked	0.19	-20.5	470	Flat	2.02
kO:t	free (mono)	0.3	-18.5	506	Flat	2.07
kVt	checked	0.18	-22.1	758	Flat	2.22
k@t	unstressed	0.11	-20.37	673	Flat	2.28
kUt	checked	0.17	-18.7	525	Flat	2.40
k{t	checked	0.21	-24.8	963	Flat	2.40
kA:t	free (mono)	0.30	-21.9	720	Flat	2.43
k@Ut	free (mono)	0.26	-21.3	598-486	End trans.	2.47
kaUt	free (dip)	0.30	-22.9	1080-629	Transition	2.52
kOIt	free (dip)	0.24	-20.0	527-374	End trans.	2.58
keIt	free (dip)	0.27	-23.0	663-489	End trans.	2.60
ki:t	free (mono)	0.28	-24.9	439	Flat	2.68
k3:t	free (mono)	0.30	-22.0	662	Flat	2.88
ku:t	free (mono)	0.26	-23.0	429	Flat	2.97
kaIt	free (dip)	0.30	-23.8	903-487	Transition	2.97
kEt	checked	0.20	022.1	822	Flat	3.42
kU@t	free (dip)	0.37	-21.1	474-512	Transition	3.52
kI@t	free (dip)	0.38	-20.5	490-509	Transition	3.67
ke@t	free (dip)	0.29	-21.4	798-853	Transition	3.75

Table 5.2 Vowel Data (Experiment 1)

1. Phonetic type: the data from Experiment 1, shown in Table 5.2 may be grouped according to phonetic type. Diphthongs respond poorly to the application of the algorithm, the monothong vowels intermediately, with the checked vowels most successful in terms of least perceived distortion. This is illustrated in Figure 5.2.

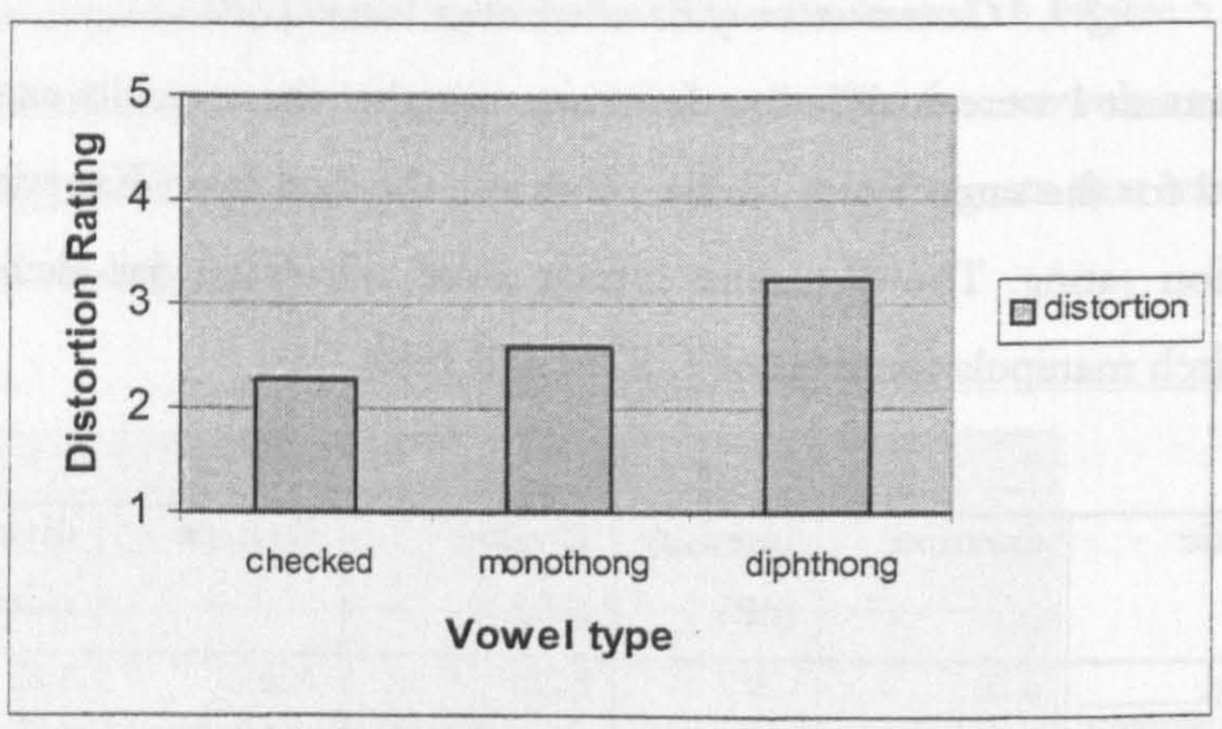


FIGURE 5.2 DISTORTION RATINGS FOR CHECKED, MONOTHONG AND DIPHTHONG VOWELS

2. Duration: the checked vowels, characterised by shorter durations are most successful, followed by monothongs that have longer durations and diphthongs with the longest durations.
3. Intensity: there is no evidence of intensity having an effect on the perception of distortion for this data.
4. F1 value: f1 value does not appear to have an effect on the perception of distortion. The results do not support Blouin & Bagshaw (2000) who state that vowels with low first formants respond better to the algorithm in terms of distortion.
5. F1 shape: The checked vowels with flat f1 contours respond best to the algorithm. The free vowels with either flat f1 contours or formant transitions at the end of the vowel sound perform averagely, whilst the diphthongs with internal formant transitions in the centre of the vowel sound perform worst.

An unexpected exception is the CVC stimuli /kEt/ which does not respond well to the algorithm. /E/ is a checked vowel, characterised by a short duration, and a flat f1 contour, and is therefore expected to have a low distortion rating. This waveform was examined and the

unvoiced phoneme /t/ was found to be incorrectly marked by the pitch detection algorithm as voiced, which may be one of the parameters that contributes to anomalous distortions.

5.2.1.4 Conclusions for Vowel Data

In summary, vowels may be categorised according to their phonetic categories of checked, monothong and diphthong vowels, which distinguish their success with the application of TD-PSOLA. Inherent characteristics of each of these groups are f1 shape and duration. Checked vowels are characterised by flat f1 contours and short durations and suffer least distortion. Monothongs have longer durations and flat f1 contours or contours that rise sharply at the end of the vowel sound and suffer intermediate levels of distortion. Diphthongs have the longest durations and formant transitions in the centre of the vowel sound, and suffer the greatest distortion. There is no evidence that intensity and f1 value affect the perception of distortion levels in speech for this data.

5.2.2 Consonants

5.2.2.1. Analysis of Consonant Data

Initially, the stimuli from Experiment 5, consisting of four versions of each of seven CVC syllables, were analysed. Only the initial consonant was altered in each of the seven CVC structures. The stimuli were analysed in terms of:

1. Phonetic type e.g. plosive, affricative etc.
2. Duration
3. Intensity
4. f1value (if applicable)
5. f1 shape (if applicable)
6. Voiced/ unvoiced/ mixed composition

The averages of duration, intensity, and f1 value were taken for each of the four versions of speech sounds. The results are summarised in Table 5.3. The % distortion detection is given in the final column, averaged for all four versions across each pitch modification level of 1, 5, 10 and 15%.

CVC identity	Phonetic type	voiced/ unvoiced	Duration (s)	mean intensity (dB)	f1 value (Hz)	f1 shape	% distortion detection
d{n	Plosive	voiced	0.02	-34.31	-	f1 rise	25.6
s{n	Fricative	unvoiced	0.18	-36.5	-	f1 rise	28.8
tS{n	Affricative	unvoiced	0.1	-30.93	-	f1 rise	39.4
n{n	Nasal	voiced	0.1	-32.55	376	flat – rise	49.4
r{n	Liquid	voiced	0.2	-28.73	-	flat – rise	56.3
j{n	Glide	voiced	0.15	-30.51	320	flat – rise	56.9
D{n	Fricative	voiced	0.16	-26.73	-	f1 rise	75.0

Table 5.3 Consonant Data (Experiment 5)

5.2.2.2 Results for Consonant Data

1. Phonetic type: the plosive sound responded best to the algorithm with respect to minimal % distortion detection, followed by the unvoiced fricative, the affricative, the nasal, the liquid, the glide and finally the voiced fricative. This is illustrated in Figure 5.3 which shows a barchart of the percentage distortion detection for each phoneme category.

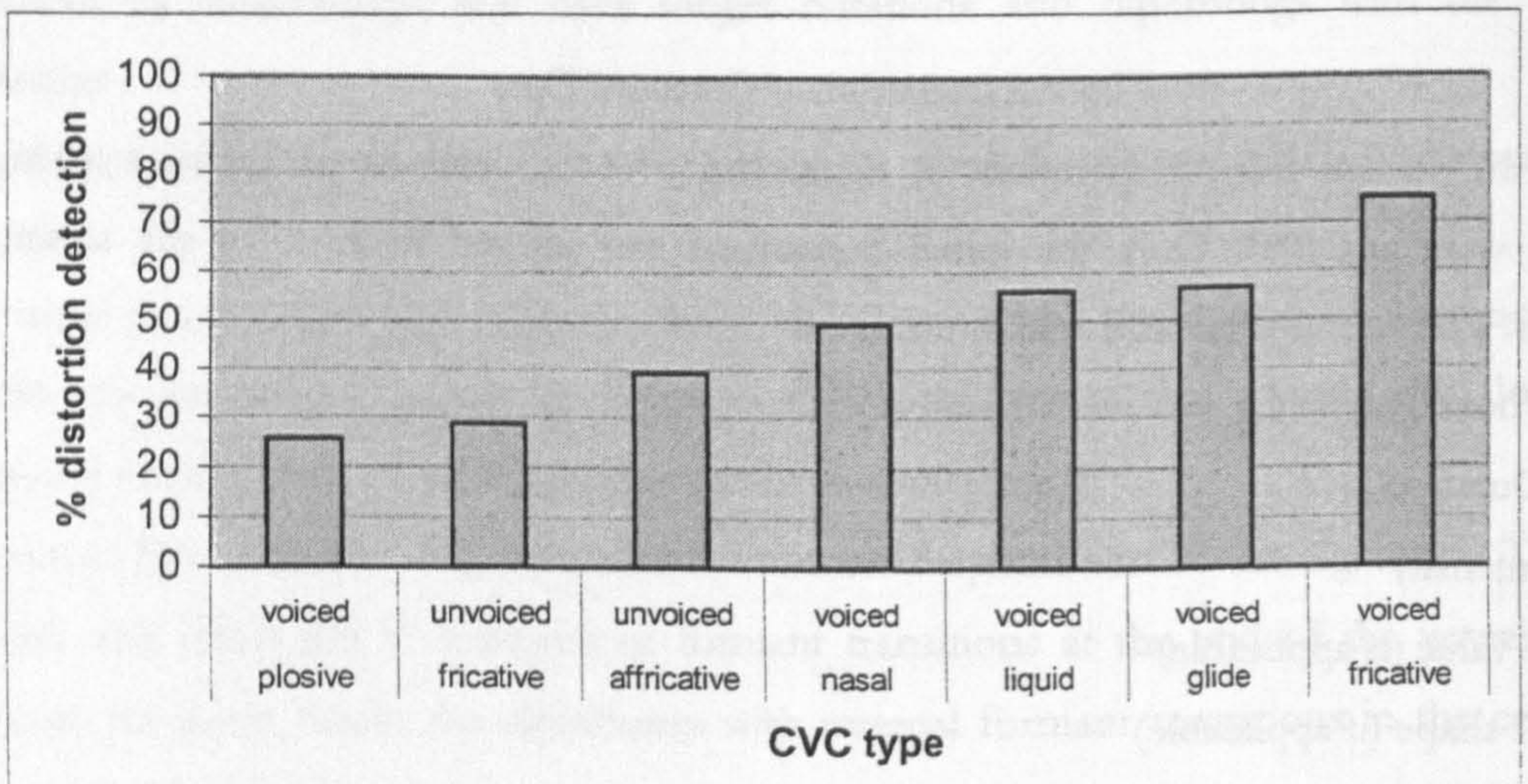


FIGURE 5.3 BARCHART OF % DISTORTION DETECTION FOR PHONEME CATEGORIES

2. Duration: The plosive has the shortest duration and responds best to the algorithm. The other consonants have similar durations. Consonants are characterised by much shorter durations than vowels so duration may not be as large a factor here as for vowels.

3. Intensity: there is no evidence to suggest that intensity has any bearing on the resulting % distortion detection for this data.
4. F1 value: An f1 value could only be measured for /n/ and /j/. Both of these had a flat f1 contour of 376 and 320Hz respectively before a rise in f1 at the beginning of the following vowel sound. With only two values, it is not possible to say whether f1 value has any significance. As it had no significance for vowel sounds, it will be assumed that for consonants, f1 value has little bearing on the resulting % distortion detection.
5. F1 shape: /d{n/, which performs extremely well in terms of least distortion detection, exhibits a very short voiced part, a silence and then a rise in f1 to the beginning of the following vowel. It was impossible to measure the f1 value of /d/ using Praat, but the effect of the plosive on the f1 contour of the vowel began at 457Hz and ended at 916Hz. /s{n/, and /tS{n/, which also perform well, exhibit a burst of noise, a silence, then a rise in f1 at the beginning of the voiced vowel. Again the f1 contour began at 484 for /s{n/ and 500 for /tS{n/ and became flat after the start of the vowel sound at 860 and 835 Hz respectively. /n{n/, /r{n/ and /j{n/ perform averagely to poorly and exhibit flat f1 contours across the initial consonant, then a transitional rise to the beginning of the /{/. /n{n/ had the lowest flat f1 contour value of 376Hz over the /n/ then exhibited a rise to 862 for the vowel sound. It was impossible to find the f1 value for /r/ as the liquids' formants are often not well defined. The f1 contour at the beginning of the vowel started at 592 Hz and rose to 808 Hz. /j{n/ had a flat f1 contour over the /j/ of 320Hz which rose to 780Hz at the beginning of the following vowel. /D{n/, which performs extremely poorly, exhibits an f1 rise from the beginning of the /D/ phoneme to the beginning of the /{/ from 500 to 860Hz. To summarise, segments exhibiting a rise in f1 at the beginning of the transition from C to V suffered less distortion in general than segments which exhibited a flat f1 shape before a rise at the CV transition.
6. Voiced/unvoiced composition: the unvoiced sounds and the voiced plosive respond well. The voiced nasal, liquid and glide perform less well but better than the voiced fricative that contains both voiced and unvoiced signals.

From these results it is hypothesised that % distortion detection is dependent upon the phonetic type of consonant, its duration, f1 shape, and voiced/unvoiced/mixed-voice composition. Intensity and f1 value do not appear to have an effect on the perceived distortion for these data.

5.2.2.3 Generalisation of Results to other Consonant Data

The data from Experiment 4 have been analysed for Voice 1 to determine whether these results can be generalised to other data for the same Voice. Table 5.4 shows the data from Experiment 4 for Voice 1 in order of increasing distortion rating. The distortion ratings were calculated for each CVC syllable averaged over each pitch manipulation level of 1, 5, 10 and 15%

CVC identity	Phonetic type	voiced/ unvoiced	duration (s)	mean intensity (dB)	f1 value (Hz)	f1 shape	distortion rating
k{n	Plosive	unvoiced	0.06	-31.2	-	-	1.90
s{n	Fricative	unvoiced	0.2	-32.8	-	-	2.04
t{n	Plosive	unvoiced	0.08	-29.5	-	-	2.05
n{n	Nasal	voiced	0.16	-28.3	430	flat – rise	2.13
dZ{n	Affricative	voiced	0.12	-32.5	-	-	2.19
f{n	Fricative	unvoiced	0.04	-24.9	-	-	2.23
h{n	Fricative	unvoiced	0.09	-35.2	-	-	2.24
d{n	Plosive	voiced	0.04	-32.0	-	f1 rise	2.33
tS{n	Affricative	unvoiced	0.14	-29.0	-	-	2.39
r{n	Liquid	voiced	0.17	-32.2	459	flat – rise	2.41
l{n	Liquid	voiced	0.15	-30.1	430	flat – rise	2.43
S{n	Fricative	unvoiced	0.2	-30.6	-	-	2.43
b{n	Plosive	voiced	0.04	-31.4	-	flrise	2.44
g{n	Plosive	voiced	0.04	-27.3	-	f1 rise	2.59
j{n	Glide	voiced	0.12	-29.2	267	flat – rise	2.73
T{n	Fricative	unvoiced	0.11	-29.1	-	-	2.89
z{n	Fricative	voiced	0.14	-33.4	420	flat – rise	3.00
p{n	Plosive	unvoiced	0.06	-32.8	-	-	3.13
m{n	Nasal	voiced	0.17	-28.2	376	flat – rise	3.21
D{n	Fricative	voiced	0.05	-31.3	240 - 720	f1 rise	3.33
w{n	Glide	voiced	0.15	-31.6	482	flat – rise	3.38
v{n	Fricative	voiced	0.10	-33.7	240	flat – rise	3.40
Z{n	Fricative	voiced	0.12	-32.8	320	flat-rise	3.53

Table 5.4 Consonant Data (Experiment 4)

1. Phonetic type: Figure 5.4 shows the levels of distortion for stimuli grouped according to phonetic type for Voice 1 averaged over all levels of pitch manipulation. It should be noted that the sample sizes are not balanced, but it only patterns of co-occurrence that are investigated during this analysis. The affricatives perform best in terms of least distortion, followed by plosives, unvoiced affricatives, nasals, liquids, glides and finally the voiced fricatives. The pattern of results are similar to those of the data set from Experiment 5, where the voiced fricative and glide performed least well, the liquid and nasal averagely, and the plosive, unvoiced fricative and affricative best.

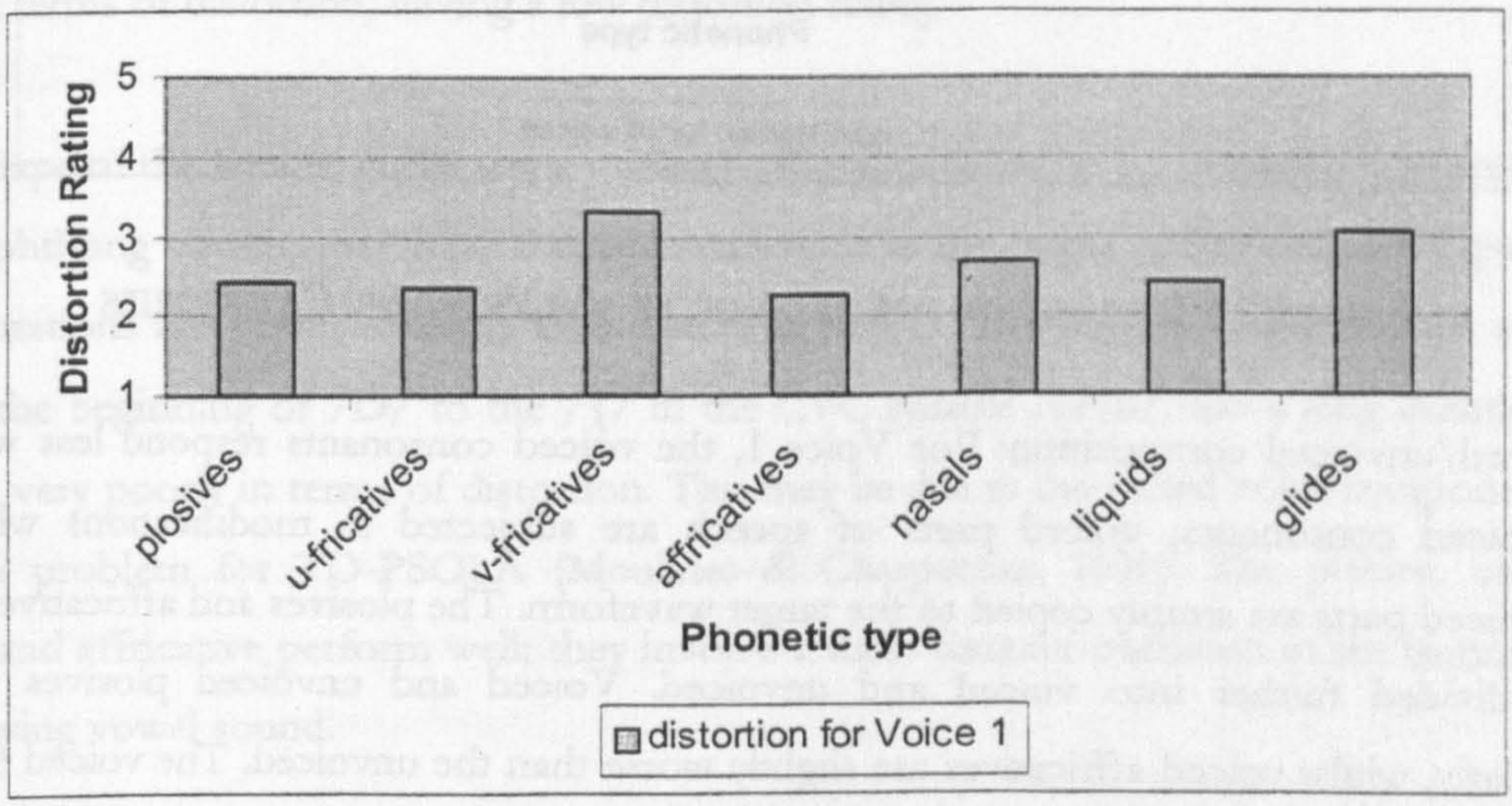


FIGURE 5.4 DISTORTION RATINGS FOR VOICE 1 PHONEME CATEGORIES

The data for the four voices investigated in Experiment 4 were also analysed in terms of average distortion ratings for each phonetic type. Figure 5.5 shows the distortion ratings across all pitch manipulation levels averaged for the four voices for each phonetic category. Again, the affricatives and plosives perform best, followed by the unvoiced affricatives. Next are the nasals and liquids, with glides and voiced fricatives performing less well.

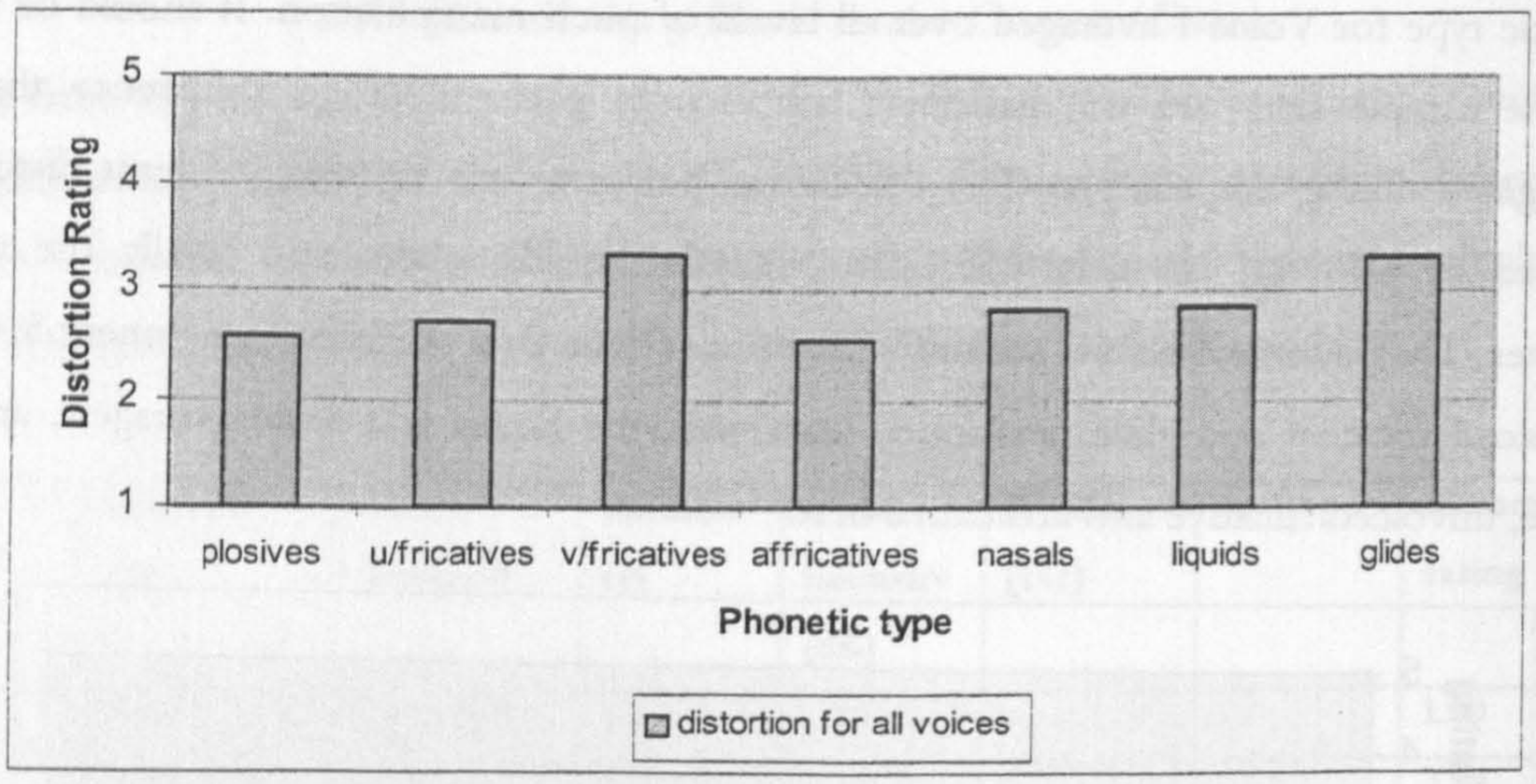


FIGURE 5.5 DISTORTION FOR ALL VOICES FOR PHONEME CATEGORIES

2. Voiced/unvoiced composition: For Voice 1, the voiced consonants respond less well than unvoiced consonants; voiced parts of speech are subjected to modifications whilst the unvoiced parts are simply copied to the target waveform. The plosives and affricatives can be sub-divided further into voiced and unvoiced. Voiced and unvoiced plosives perform similarly, whilst voiced affricatives are slightly worse than the unvoiced. The voiced fricatives suffer most due to their mixed stochastic and periodic composition.
3. Duration: the voiced plosives are as successful as the unvoiced plosives - this may be due to the short duration of plosives that makes distortion imperceptible for this pitch manipulation range.
4. F1 shape: the f1 shape of the liquids, glides (semi-vowels) and nasals have formant transitions in the centre of the voiced CV part and suffer high distortion levels. The voiced fricative exhibits a rise in f1 from the beginning of C to V and responds very poorly in terms of distortion. The plosive, unvoiced fricative and affricative perform well; they involve a short formant transition at the beginning of the following vowel sound.
5. There was no evidence to suggest that intensity and f1 value had any effect on the distortion ratings.

5.2.2.4 Conclusions for Consonant Data

In summary, consonants may be categorised according to their phonetic categories, voiced/ unvoiced composition, f1 shape and duration, which together distinguish the success of their response to the application of the TD-PSOLA algorithm.

Unvoiced types have lower distortion ratings than voiced due to no modification of the unvoiced parts of the signal during the TD-PSOLA pitch modification process. The exception is the voiced plosive, which is of such short duration that the voicing appears to have no perceptual effects in terms of distortion, having a low distortion rating.

The f1 shape of the liquids, glides (semi-vowels) and nasals followed by a vowel sound are similar to the diphthong vowels; they have formant transitions in the centre of the voiced CV part with longer durations and suffer similarly high distortion levels. The voiced fricative exhibits a rise in f1 from the beginning of /D/ to the /{/ in the CVC syllable /D{n/, has a long duration and responds very poorly in terms of distortion. This may be due to the mixed voice component that can be a problem for TD-PSOLA (Moulines & Charpentier, 1990). The plosive, unvoiced fricative and affricative perform well; they involve a short formant transition at the beginning of the following vowel sound.

These results may also be generalised to some extent to all voices tested in Experiment 1, when the data are grouped according to phonetic category.

5.2.2.5 Issues

An assumption has been made that it is the distortion in the consonant that is evaluated, not its effect of the following vowel. In reality, the vocal tract moves from the consonant to the position for the following vowel (coarticulation) where there are brief influences on the formants at the beginning of the vowel. It may be the perceived distortion in the consonant itself, and also the effect of the consonant on the following vowel formants, which is evaluated. Different levels of distortion may be perceived for different following vowels. Although not experimentally verified, an informal listening test, replacing /{n/ with /It/ as the following VC construct, indicates that the /j/ (glide) and /D/ (voiced fricative) respond least well, followed by /r/ (the liquid), and then /n/ (the nasal). /b/ (the plosive), /tS/ (the affricative), and /s/ (the unvoiced fricative) are

all unaffected. This pattern concurs with the results from Section 5.2.2.1 and indicates that it may be likely that the distortion perceived for the consonants is relatively independent of the following vowel, although further investigation would be advantageous. It should be noted that coarticulation is speaker specific (van den Heuvel *et al.*, 1996) so results for this voice reflect the coarticulation behaviour of this speaker only.

5.2.3 Summary

The consonant and vowel sounds have been categorised in terms of distortion levels and parameters have been identified which contribute to the success of segments when TD-PSOLA is applied for pitch modification. Both consonants and vowels sounds can be grouped according to their phonetic category determined by their manner of articulation. These categories have inherent characteristics of differing durations, f1 shape, and for consonants, voiced/unvoiced/mixed-voice composition.

In the following section, these results will be used to develop a novel speech corpus design tailored to the use of TD-PSOLA, and a signal processing distortion measure to be included in a segment selection process. During this analysis, voiced fricative phonemes were identified as being especially problematic, and a special selection process for these will also be developed. Before this, existing speech corpus designs and segment selection processes are reviewed.

5.3 Review of Existing Speech Corpus Techniques

5.3.1 Introduction

Speech signals contain more data than just f0 and duration. Simple diphone approaches to speech synthesis do not consider additional aspects of natural speech such as varying voice qualities. Corpus-based synthesis attempts to retain such details. Furthermore, speech segment quality is maintained by selecting segments closer to the target values, which minimises the amount of signal processing required. The most extreme corpus-based system CHATR (Black & Taylor, 1994) does not perform any prosodic modifications but relies on the fact that listeners are less

disturbed by discontinuities. This is often not the case and for speech to be generated with less discontinuity, Portele (1998) states that post-concatenation is still inevitable as finding all features in all combinations is not possible (van Santen & Buschbaum, 1997).

Portele (1998) claims there are three main factors that determine the quality of the synthetic speech output: the size and variety of the segments in the corpus, the choice of annotated features and accuracy of annotations, and the unit selection algorithm. Donovan & Woodland (1999) suggest it is also the number of concatenation points.

In the following section, the current speech corpus strategies of corpus design (size of corpus and variety of segments) and the unit selection process are reviewed. Then the concept of a novel TD-PSOLA balanced corpus is presented which is intended to retain much of the resulting speech quality by reducing serious signal processing degradation. This would potentially improve the resulting output quality of a TTS system that makes use of the TD-PSOLA algorithm.

5.3.2 Existing Corpus Designs: Size and Variety of Segments

A corpus may be designed to reflect the chosen application for limited or closed domains e.g. a speaking clock synthesiser. The synthesis is then usually robust and of a high quality (Black & Lenzo, 2000b). For open domain applications where arbitrary sentences are synthesised, a more general approach to corpus design is required.

Corpora may be phonetically rich containing every possible speech segment, or alternatively they may be phonetically balanced with specific additions needed to cover unusual phonetic sequences. Black & Lenzo (2001) addressed the concept of creating the smallest set of utterances for a speech corpus that would give optimal phonetic coverage. They take a large amount of intended output then greedily select sentences from Lewis Caroll's "Alice's Adventures" to give best diphone coverage with minimal redundancy.

Factors such as lexical stress, pitch, duration, and position in phrase are also taken into consideration. The addition of each feature increases the amount of data necessary and, as van

Santen & Buschbaum (1997) point out, getting all possible features is not feasible due to the complex variations and combinations of speech.

A current concern of corpus-based approaches is how to increase the emotional capacity (the ability to produce large prosody variations for arbitrary sentences) of such a system without greatly increasing the size of the corpus (Niimi *et al.*, 2001). Their database is prosodically balanced, containing segments which have emotional speech quality, and so any Japanese accent pattern may be generated with minimal signal processing modification to achieve more diverse pitch and duration targets without serious degradation.

As an example of typical size, the AT&T system uses 84,000 demiphones, which is equivalent to 1.8 billion unit pairs, plus 36 million mid-phone transitions (Möbius, 2000). Experiments have shown that a subset of 1.2 million unit pairs gave 99% coverage, and that such a subset was capable of producing sequences of speech almost identical (98.2%) to those using the optimal selection from the entire corpus (Beutnagel *et al*, 1999b).

Larger corpora require more time to record, leading to the problems associated with multiple recording sessions such as voice quality changes, which make unit selection more unreliable in terms of output quality. The data also need to be labelled, much of which still needs to be done by hand.

Obviously there must be a trade off between the variety of segments available and the size of the corpus. Signal processing is still required to keep the corpus a manageable size. As signal processing is necessary, the design of the corpus should take the requirements of the algorithm into account. This is addressed in the following section, which develops a novel speech corpus design that is balanced to the needs of the signal processing algorithm used to achieve the desired prosody.

5.3.2.1 A TD-PSOLA Balanced Speech Corpus

A TD-PSOLA balanced speech corpus would retain or even reduce the necessary size of the corpus; the need for certain speech segments to be represented in as many different contexts with varying pitch and durations may be reduced. Limiting the size of the corpus improves the

consistency of speech by allowing the recording to take place in one session and avoiding speaker fatigue.

Analysis of different speech sounds in Section 5.2 indicates that segments such as checked vowels, plosives, and affricatives respond well to pitch manipulation; fewer representations of these segments would be necessary in a TD-PSOLA balanced speech corpus. Black & Lenzo (2001) describe how to successfully prune a database to remove unnecessary versions of segments.

It would be advantageous to include more of the segments that do not respond well to TD-PSOLA, such as diphthongs and voiced fricatives. This would be especially necessary in a phonetically balanced corpus for adversely affected combinations of segments that are rare in the English language. Extending the database or corpus with problematic segments has already been achieved successfully for other purposes; Klabbers & Veldhuis (2001) included extra versions of highly coarticulated diphones in their database. The CHATR corpus contains not only a phonetically balanced corpus but also sets of isolated words and sets of isolated sentences to cover problematic texts.

Thus a speech corpus with optimal coverage for TD-PSOLA may be designed. Section 5.5 develops this concept by using the MOS scale ratings to predict which phoneme groups respond well to TD-PSOLA and hence allow material to be designed having more of certain segments and fewer of others, rather than being purely phonetically balanced.

5.3.3 Existing Unit Selection Procedures

Corpus-based systems enable the use of variable length units such as phones, diphones, triphones (or even longer) rather than just diphones or syllables. Non-uniform synthesis was first suggested by Sagisaka (1988) of ATR (Advanced Telecommunications Research Institute) and Takeda *et al.* (1990). Also at ATR, Black & Campbell (1995) and Hunt & Black (1996) began parallel research that also included prosodic measures (duration and pitch) in the unit selection process. Unit selection aims to select the longest available string of phonemes therefore minimising the number

of concatenation points. Möbius (2000) states that several hours worth of speech in a corpus-based system will produce utterances constructed of units that are considerably longer than diphone or demi-syllable synthesis.

Unit selection involves finding the best unit in the corpus to fit a target utterance. 'Best' is in terms of closest phonetic, acoustic and prosodic features. Unit selection minimises two cost functions: one for unit distortion and one for concatenation distortion. Unit distortion or *target cost* measures the distance from the target values to the values of a candidate unit. Concatenation cost measures the distance between two adjacent segments at the concatenation point, hence measures the quality of the segment join.

Segments in the corpus are labelled with features consisting of both segmental and prosodic properties. The feature vector of these properties is computed for the target segment at runtime for values predicted from text. Hence the unit distortion cost can only make use of those that can be predicted from text, namely the phonetic and prosodic features of phonetic context, duration, pitch, and position in syllable, word and phrase.

Continuity distortion can use all features, as it compares features of two actual segments from the corpus, making use of using spectral properties such as mel-cepstrum vectors, local f0, and local power. A weighted sum of distances is used to calculate the continuity cost and also for optimal coupling (Conkie & Isard, 1996). Optimal coupling determines the best place to join two units by varying the cut point of the speech segments.

Target and corpus units are labelled with the same set of features. For each feature a distance measure (e.g. absolute, equal/non-equal, squared difference) is selected. The target cost may be calculated as the weighted sum of the distances of these features.

Unit selection occurs as follows. Each unit in the corpus is represented by a state in a state transition network with state occupancy represented by the unit distortion value, and state transitions by the continuity distortion values. Hunt & Black (1996) select the optimal units by finding a path through the states that minimises both costs.

The size of unit selected is usually phone sized which does not actually minimise concatenation points, although the method does encourage longer unit selection, as the concatenation cost would be zero between segments found together in the corpus.

The difficulty in unit selection is determining the distance measures themselves and their weightings. Currently, there are two main approaches: weight space search (Black & Campbell 1995, Campbell & Black 1996) and multiple linear regression (Hunt & Black, 1996).

Weight space search uses analysis-by-synthesis to determine the weights. An utterance is synthesised from the first choice best set of units and its distance from the natural utterance is measured. This is then repeated for various weight settings and utterances until the best set of weightings is found.

Multiple linear regression is used to find the weightings for the target cost separately and the weight space search is used to train the concatenation cost. This allows separate weights for different phone classes to be used e.g. high vowels, nasals etc. It is also computationally less expensive. For each occurrence of a type of unit in the corpus, the acoustic distance between each is calculated. The target cost distances for each feature are also calculated producing a large table of acoustic costs and feature distances as shown in Equation 5.1.

$$AverageCost = w_1 D_1 + w_2 D_2 + ...w_n D_n \qquad \text{Eqn 5.1}$$

w_n can then be estimated, for each phone each class of unit, using linear regression. This method is used in CHATR (Campbell *et al.*, 1989).

Some new more efficient weight training methods have been proposed (Meron & Hirose, 1999) that divide the analysis-by-synthesis into two processes: selection and scoring, meaning only candidate segments are scored. Additionally, they apply regression training to target and concatenation costs simultaneously, which is an improvement as the two costs are not independent of one another. It also takes into account prosodic costs at synthesis time. Holzapfel & Campbell (1998) use fuzzy logic to determine the cost of a candidate unit and also define a range of acceptable distances for each feature, which prevents large mismatches (fuzzy logic allows small deviations to exist assuming they are not be perceptually relevant). Wouters &

Macon (1998) and Macon *et al.* (1998) evaluated the use of distance measures to predict optimal unit selection, which provided a correlation of 0.66 with perceptual differences, indicating there is much work needed here. It may be that some important features are missing or that weights have not been trained accurately.

The following section discusses the development and potential addition of a TD-PSOLA distortion measure to current unit selection processes, which may be one of the features missing from current processes.

5.3.3.1 A TD-PSOLA Distortion Measure

A TD-PSOLA signal-processing measure may be added to existing target cost calculations. This would not be just a simple measure of the absolute distance in Hz for all segments, but weighted according to the phonetic identity of the segment. This would allow segments adversely affected by the algorithm to be selected closer to the target in terms of signal processing cost, as they would have greater signal processing cost weightings, lessening the importance of other features in the selection vector, such as position in phrase, when necessary. The signal–processing measure is developed in Section 5.6 where the MOS scale ratings are converted into weights for the individual phoneme groups, using linear regression. This TD-PSOLA distortion measure, if included as part of a target cost estimation to select units from a corpus would take into account the effects of TD-PSOLA on final pitch modification and select segments that would suffer minimal TD-PSOLA distortion.

5.3.4 Context Clustering

An alternative process to unit selection is context clustering using a decision tree (Nakajima & Hamada 1988, Nakajima 1994, Itoh *et al.*, 1994, Wang *et al.* 1993). Context clustering is used in the Microsoft Whistler system (Hon *et al.* 1998, Huang *et al.* 1996). The idea is to take all units of each type found in the corpus and to define an acoustic distance between them. All units of a particular type, having the same segmental phonetic context, are grouped together. Using features used for target cost calculation, such as metrical and prosodic context, the cluster is split by minimising the acoustic distance between members until some threshold is attained. Using CART trees (Brieman *et al,* 1984), which allow the most significant factor to be selected by using a

greedy algorithm (a greedy algorithm makes optimal decisions at each stage without regard for subsequent stages), the result is a decision tree available at synthesis time with each leaf on the tree represented by a segment and its features. TTS systems often use binary trees: at each node, clusters are split by questions about features of units that require a yes/no answer. The target cost is the distance of a unit to its cluster centre. A subset of phones from each cluster is selected and stored. The selection process computes statistics for duration, energy and pitch and removes tokens not near the average. A small number are selected based on an objective function that measures the match with the cluster.

The number of leaf nodes therefore determines the number of synthesis units, and the depth of the decision tree determines the size of the inventory. This provides a trade-off mechanism between unit inventory size and specificity. A small inventory will have a shallow tree and the clusters will be less context-specific, and will probably require more signal processing.

Syllable stress, word accent and position relative to the phrase boundary, all influence pitch and duration of phonemes. Surrounding phonemes can cause coarticulation phenomena and must be taken into consideration during the selection process. This method inherently determines the importance of contextual (syllable position, word/phrase position) and coarticulatory effects using questions such as "is the previous phoneme a /U/?".

Unit selection can then be performed using unit distortion (measured as the Euclidean distance from the cluster centroid (Iwahashi *et al.*, 1992)) and continuity distortion to minimise the global cost over the utterance to be synthesised. Further developments (Iwahashi & Sagisaka 1995, Sagisaka & Iwahashi 1995) simultaneously minimise both costs, although this is very computationally expensive. This version also added prosodic properties such as pitch, duration and intensity to the selection criteria.

Black & Taylor (1997) combined context clustering and unit selection and implemented it in the Festival system (Black *et al.*, 1999). The segments are clustered according to their phonetic and prosodic contexts with each cluster containing more than one segment (usually 10 to 15 examples). The main advantage is that most of the computation is performed offline so the search effort at runtime is reduced; the target cost computation is moved from synthesis time

into the training process. Rather than having to select between all available units, the most appropriate cluster is chosen using the decision tree, then the best unit is chosen from the cluster. Möbius (2000) suggests that the cost of signal processing could be included in the scoring during unit selection from the cluster. The tree is grown until the desired number of terminal or leaf nodes is reached, or if the number of phones in the leaf node falls below a threshold. This approach finds larger matches with less computation and does not depend on the prediction of duration and f0, which is notoriously difficult.

Prosodic features are used in the Eloquens speech system when large mismatches are encountered between candidate and target values (Balestri *et al.*, 1999) as they argue that it is still necessary to impose model-based artificial prosody to attain enough flexibility. Their unit selection takes into account prosodic labels as well as acoustic correlates of f0 and duration. Results suggest that unit selection by categorical prosodic features is better than by matching numerical prosodic values.

Synthesis for either approach uses the target costs (as either weighted distances, or distances from the cluster centre for members of the optimal cluster) and the continuity costs between the possible selections of units to calculate the minimal total cost. The optimal route through these units is found using a Viterbi decoding algorithm.

The following section looks at the concept of context clustering guided by the requirements of the TD-PSOLA algorithm which may be used for final prosodic modifications.

5.3.4.1 TD-PSOLA Guided Context Clustering

The tree does not have to be grown evenly as long as there are enough segments in the corpus. Phone clusters that respond well to TD-PSOLA would not need to be grown as far as there can be fewer variations; their prosodic characteristics of duration and pitch can be modified to target values without introducing unacceptable levels of distortion. Phone clusters that do not respond well would have a deeper tree structure with more questions and hence more context specific clusters.

Voiced fricatives, diphthongs and glides would be grown furthest giving more variety of pitch and duration values, therefore it would be more likely to find ones closer to target values, hence less modification would be required. Checked vowels, affricatives and plosives would have shallow branches with monothongs, nasals, and liquids having moderate branches. The novel signal processing cost would also be included in the unit selection from the cluster as suggested by Möbius (2000). The use of this in conjunction with a TD-PSOLA balanced corpus would allow the tree to be grown to different depths to facilitate the choice of TD-PSOLA problematic segments having prosodic values nearer the target values.

Additionally, the decision tree approach would allow generalisation to new contexts not encountered during training by backing-off to broader categories for the neighbouring contexts. This may be used for bad response TD-PSOLA candidates from the chosen context; if the prosodic modification will be large, segments in other contexts that may not be correct linguistically may have closer prosodic values. The generalisation to new contexts has been implemented successfully by Chou *et al.* (1999). The development of a TD-PSOLA guided context clustering approach is outside the scope of this work, although the concept may provide a successful solution to reducing distortion introduced by the TD-PSOLA algorithm.

5.4 Summary

This section reviewed existing corpus design and speech unit selection methods. A TD-PSOLA balanced corpus and signal processing distortion measure has been proposed. This may be implemented in a TTS system as either unit selection from a whole corpus or as context clustering and as unit selection, to select from the clusters. The following section develops the TD-PSOLA balanced corpus and signal processing distortion measure based on the data gathered during the investigative experiments in Chapter 4.

5.5 Development of a TD-PSOLA Balanced Corpus

The data from Experiment 5 are used to design the TD-PSOLA balanced corpus; segments that suffered high percentage distortion detection will be represented in greater numbers in the corpus. The data are in a categorical yes-no format depending on whether any distortion was detected. Klabbers & Veldhuis (2001) suggest using a 'majority score' to reduce the variability of judgements between participants. At each pitch modification level, a stimulus is marked as distorted if 5/10 participants judge it as so. A score of 5/10 was chosen to give a spread of values and avoid any ceiling effect. The percentage distortion detection was then calculated averaging the distortion at each pitch manipulation level and for each of the four recording versions. The results are shown in Table 5.5.

plosive	**u/ fricative**	**affricative**	**nasal**	**liquid**	**glide**	**v/ fricative**
13%	19%	44%	56%	69%	69%	94%
checked	**monothong**	**diphthong**				
27%	50%	69%				

Table 5.5 Percentage Distortion Detection for Different Phonetic Categories

Voiced fricatives were nearly always detected as containing perceptible distortion when the majority score was set at 5/10. Conversely, plosives were rarely detected as containing distortion.

It is this measure that will be used to develop the novel speech corpus design, containing more of the adversely affected segments such as voiced fricatives, glides, liquids and diphthongs, and less of the segments that respond well to the algorithm such as plosives, unvoiced fricatives and checked vowels.

5.6 Development of a Signal Processing Distortion Measure

In this section a signal processing distortion measure is developed, which may be used as an objective measure to predict the occurrence of distortion. Previous segment selection processes have either ignored the effect of signal processing by not including a prosodic measure, whereas others have included a prosodic measure as simply the absolute distance of the f0 between the

candidate segment and the target. This work proposes a signal processing distortion measure that comprises an absolute distance from the target, weighted for individual phonemes according to how they respond to the TD-PSOLA algorithm in terms of perceived distortion levels.

The weights were trained using the MOS scale ratings gained from experimentation. It is widely accepted in the speech community that MOS scale ratings provide a popular subjective measure of naturalness, although running MOS experiments is time consuming. MOS experiments cannot be carried out for every parameter change or voice change. This work aims to show that a small set of training data can be used to derive the weights for a signal processing cost.

The signal-processing cost C_{sig_proc} may be defined as shown in Equation 5.2.

$$C_{sig-proc} = W_{phoneme} \times D_{f0} \qquad \text{Eqn 5.2}$$

where C_{sig_proc} is the cost of necessary signal-processing, D_{f0} is the absolute distance of the fundamental frequency of the candidate segment selected to the target fundamental frequency, and $W_{phoneme}$ is the weighting of the importance of the measure determined by the identity of the phoneme. In other words, if the distance between the target and the selected segment is zero, the signal processing cost is zero. C_{sig_proc} is believed to be correlated with the amount of distortion present in the segment, therefore a lower cost would mean less resulting distortion after TD-PSOLA pitch modification.

If no consideration is given to the effect of TD-PSOLA in terms of distortion for individual phoneme identities during selection, $W_{phoneme}$ can be set to 1, then the cost is simply the absolute distance in Hz between the target f0 and the segment f0. Initially, the weights $W_{phoneme}$ were set to 1 for all phonemes.

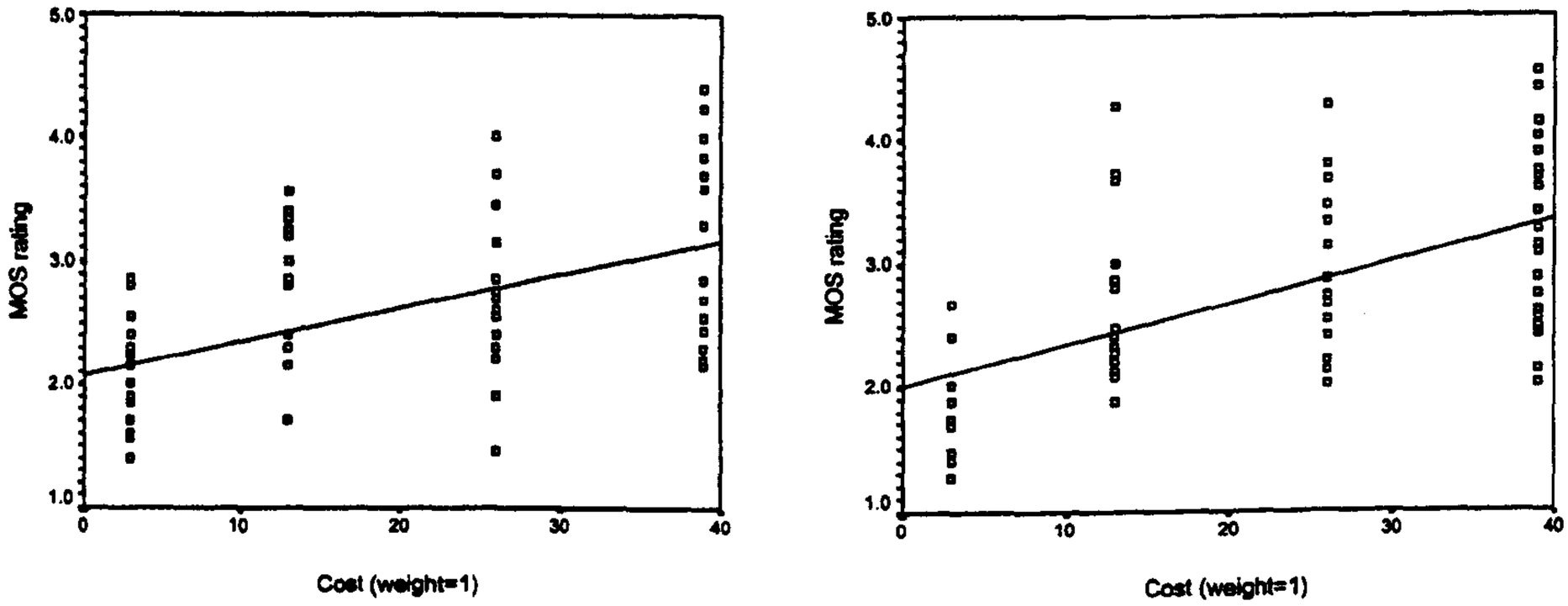
The vowel data from Experiment 1 and the consonant data from Experiment 3 were used to train the weightings for the phonemes. Weightings were developed for categories of phonemes grouped according to manner of articulation. Categories were used as analysis has indicated that phonemes grouped according to manner of articulation were affected similarly in terms of

distortion. In addition, this lessens the effect of anomalous occurrences of distortion in segments, the cause of which is still uncertain. It may also allow the generalisation of weightings to other voices. The categories are:

- vowels: checked vowels, monothong vowels, and diphthong vowels
- consonants: plosives, unvoiced fricatives, voiced fricatives, nasals, liquids, glides and affricatives.

In Figure 5.6 the MOS score data for the individual stimuli are plotted against the initial signal-processing cost at 1, 5, 10 and 15% pitch modification (3, 13, 26, 39Hz absolute distance) when weightings $W_{phoneme}$ were set to 1. Figure 5.6 shows scatterplots of (a) consonant data and (b) vowel data. Each point represents the MOS rating, averaged for 20 participants at each cost.

A linear regression trend line for each scatterplot was estimated by calculating the least squares fit through the data. The equation of the line for consonant data is $y = 0.026x + 2.07$, and for vowel data is $y = 0.033x + 2.00$.



(a) Consonant data

(b) Vowel data

FIGURE 5.6 SCATTERPLOT OF COST AND MOS SCORES

The regression line in Figure 5.6 is used as an approximation to show the trend in the data and as an intermediate stage to achieve the training of the weights. The data for the interval scale Cost are confined to categories due to the design of the experiment, and it would require further

experimentation with greater amounts of data for a complete model. To this end, the correlations quoted are purely indicative of the trend and the success of the training of the weightings, by comparing the values at this intermediate stage to those found after the application of the weights. A Spearman's rho correlation was performed between the cost and the MOS ratings for the consonant and vowel data. A high correlation would indicate that the absolute value of cost gives reasonable perceptual correlations.

- consonants: (rho=0.530, N=92, $p<0.01$, *one-tailed*).
- vowels: (rho=0.605, N=80, $p<0.01$, *one-tailed*).

The correlations are significant, although the use of different weightings for individual phonemes may provide larger correlations.

The linear regression lines on the scatterplots in Figure 5.6 were used as an intermediate representation to estimate the signal-processing cost (denoted by $C_{weighted}$) needed to predict the MOS rating for each data point.

The optimised weights $W_{phoneme}$ were then calculated using Equation 5.3.

$$W_{phoneme} = C_{weighted} / D_{f0} \qquad \text{Eqn 5.3}$$

where $C_{weighted}$ is the new estimated cost predicted from the linear regression lines and D_{f0} is the absolute distance of the segment from the target f0 value in Hz.

The average MOS ratings of original unmanipulated waveforms were calculated to provide the lower bounds for MOS ratings of synthetic stimuli. A MOS score of 1.93 for consonants and 1.77 for vowels was calculated as the average score for unmodified stimuli. It was decided that ratings below this value should not be used in weight training. In fact, 48% of consonants at 1% modification and 55% of vowels at 1% modification level had MOS ratings below this value, so the 1% level was not used during the weight training. Inclusion of the 1% level was found to yield a lower correlation due to the small amounts of distortion introduced at the 1% level and hence varied response amongst participants.

Average weights were calculated for each group of phoneme category. The optimised weights in the cost function for each group of phoneme category are given in Table 5.6.

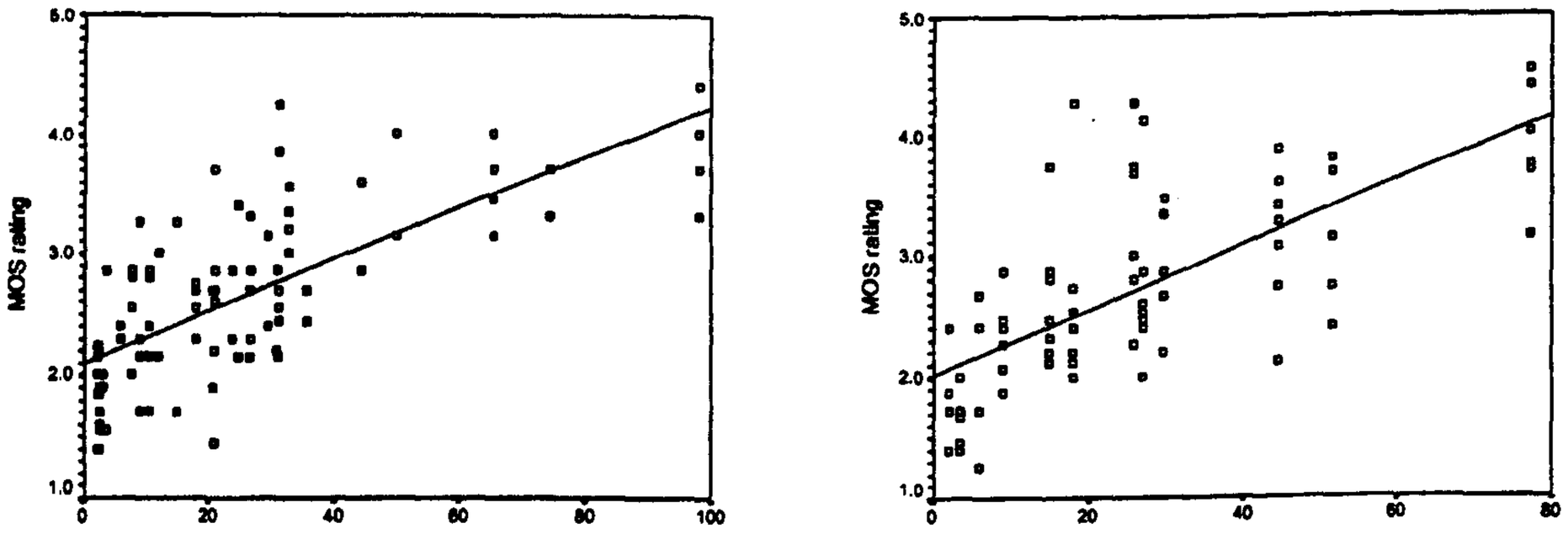
affricatives	**plosives**	**u/frics**	**v/frics**	**glides**	**liquids**	**nasals**
0.79	0.80	0.68	2.25	1.91	0.91	1.13
checked	**monothong**	**diphthong**				
0.69	1.14	1.98				

Table 5.6 Weights for Different Phonetic Categories

The correlation between MOS scale ratings and the new signal-processing cost $C_{weighted}$ was calculated:

- consonants: (rho=0.710, N=92, $p<0.01$, one-tailed).
- vowels: (rho=0.731, N=80, $p<0.01$, one-tailed).

The use of weighted distances in the calculation of the cost of signal-processing using TD-PSOLA has increased the correlation between this cost and the subjective MOS scale ratings of distortion levels. This can be seen in the scatterplots in Figure 5.7 which show the average MOS scale rating versus the new weighted cost for consonants and vowel data. Several outliers are visible, which would reduce the value of the correlation.



(a) Consonant Data

(b) Vowel data

FIGURE 5.7 SCATTERPLOT OF WEIGHTED COST AND MOS RATINGS

The correlation between the weighted cost derived for this voice and the other voices from Experiment 4 was then measured.

- Voice 2 (rho=0.672, N=92, *p<0.01, one-tailed*)
- Voice 3 (rho=0.528, N=92, *p<0.01, one-tailed*)
- Voice 4 (rho=0.711, N=92, *p<0.01, one-tailed*)

This suggests that the weightings can be generalised to other voices to a certain extent, but would probably need to be retrained for individual voices using their respective data.

5.6.1 Minimum Distortion for Pitch Modification of Voiced Fricatives

From the analysis in Section 5.2.2.2, the worst distortion is apparent when raising the pitch of voiced fricatives as this involves time-scale modifications; repeating Short Term (ST) signals causes local periodicity which is perceived as buzzyness. It is possible to avoid this for purely unvoiced parts of speech, by reversal of every other ST-signal. Unfortunately, voiced fricatives contain both voiced and unvoiced parts, making ST-signal reversal impossible. A new segment selection process has been devised which can be used in a corpus-based system containing multiple versions of such segments. This repetition of ST signals when increasing pitch could be avoided, by selecting a signal of longer duration than required. Thus by combining pitch and duration manipulation i.e. decreasing duration, there is no need for ST-repetitions. For a flat f0 contour, the necessary duration of the voiced part of the candidate segment is given in Equation 5.4.

$$duration_{segment} = (duration_{target} \times f0_{target}) / f0_{segment} \qquad \text{Eqn 5.4}$$

where $duration_{segment}$ and $f0_{segment}$ refer to the physical duration and f0 of the candidate segment in the corpus, and $duration_{target}$ and $f0_{target}$ refer to the required duration and f0 target values for synthesis.

For instance, the voiced fricative /Z/ with a static pitch contour may be required to have a target f0 of 220Hz and a target duration of 0.12. The segment from the corpus that would give least distortion in terms of repeated ST-signals when increasing the pitch of the segment would have a product of its duration and its f0 similar to the product of the target duration and the target f0. This is expressed mathematically in Equation 5.5.

$$duration_{segment} \times f0_{segment} \cong duration_{target} \times f0_{target} \quad \text{Eqn 5.5}$$

In other words, the product of the target values of 220 x 0.12 equals 26.4, so a segment in the corpus having a similar product value may suffer less distortion than a segment having a pitch perhaps closer to the target value but with a shorter duration, hence requiring ST-repetition. A segment of 210Hz and duration 0.12 seconds has a product of 25.2, indicating ST-repetition would be necessary, but a segment in the corpus having a pitch further from the target of perhaps 200Hz but with duration 0.132 seconds, would have the same product value as the target segment and no ST-repetition would be necessary.

This approach may only be applied for moderate modifications. Large duration modifications introduce the issue of whether the characteristic content of the phoneme would be altered by removal of major amounts of ST-signals. Therefore the candidate segment should have f0 and duration values relatively close to the target values.

This may be implemented in the Praat software by adding a duration point that corresponds to the new synthesis pitch point. The position of the duration point is given in Equation 5.6 as a ratio.

$$Position = duration_{target} / duration_{segment} \quad \text{Eqn 5.6}$$

For the worked example above, where $duration_{target}$=0.12s and $duration_{segment}$=0.132s, the duration point would be positioned at 0.91. This is illustrated in Figure 5.8 which shows a signal being increased in pitch and decreased in duration in the Praat editor window.

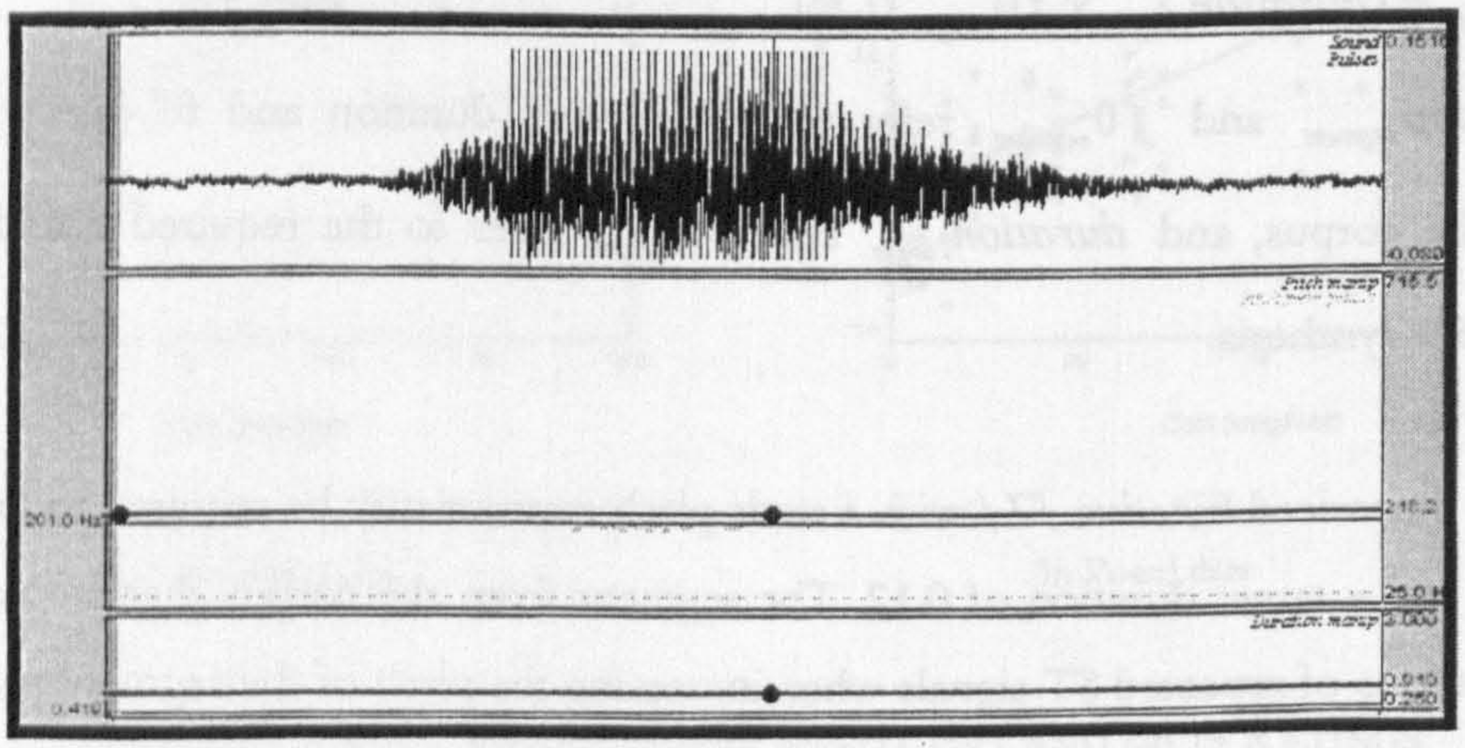


FIGURE 5.8 VOICED FRICATIVE MODIFICATION

Informal listening indicates that much of the buzzyness encountered when imposing static pitch contours on voiced fricatives is eliminated, but its success is evaluated formally in Chapter 6. Dynamic pitch contours require a more complex algorithm, where corresponding duration contour points are calculated for every synthesis pitch point. Informal listening suggests that the use of this process is not as critical for dynamic contours, as it is the inherent local periodicity of repeating similar ST signals that may contribute to the distortion. The repetition of ST-signals does not seem to have as adverse an effect on dynamic f0 contours, which is in accordance with the work of Blouin & Bagshaw (2000) who found that static pitches provide the worst-case scenario for TD-PSOLA distortion discrimination. The illustration of this algorithm at the diagnostic static-pitch word-level ensures that the worst-case scenario has been considered and any applications during dynamic speech can only introduce less distortion.

This method could also be extended to purely voiced speech sounds for both increasing and decreasing pitch, to avoid or minimise ST repetition and deletion. ST repetition and deletion is a form of signal manipulation which could introduce some distortion and avoiding this may provide better quality speech output with TD-PSOLA. Further research would be necessary, and this is discussed as a recommendation in Chapter 7.

5.7 Summary

Data from the investigative experiments in Chapter 4 have been analysed in an attempt to model the occurrence of perceptible distortion when speech is pitch-manipulated using the TD-PSOLA algorithm.

Similar distortion levels were found for phonemes from the same phonetic category, when grouped according to manner of articulation. Such phonemes from the same phonetic category possess inherently similar characteristics of duration, f1 shape (if applicable) and voiced/unvoiced/mixed composition (if applicable).

The results of the analysis were used to design a novel speech corpus balanced to the needs of the signal processing algorithm. Segments that suffered large amounts of perceived distortion will

be represented in greater numbers in the corpus, therefore providing more prosodic variations. The aim is to reduce the amount of signal processing required for such segments to achieve the target f0 of the new constructs to be synthesised. This may thus reduce the level of perceptible distortion introduced.

The results of the analysis were then used to develop a signal processing distortion measure by calculating individual weightings for the phonetic categories. These weights indicate the levels of perceived distortion that may be potentially introduced in to the speech signal when pitch-modified to the target f0 value. The product of the individual weighting for a phoneme and the distance (in Hz) of the modification results in a signal processing cost, which may be used to determine the amount of distortion that may occur in that speech segment after modification by TD-PSOLA. This signal-processing distortion measure may be used as part of an existing segment selection process when selecting segments from a speech corpus which uses TD-PSOLA for final prosody modifications. The use of such a measure would ensure that the cost of signal processing is taken into consideration besides other criteria such as 'position in word or phrase', when selecting the appropriate segment from the corpus to synthesise a new utterance.

In addition, a special selection process was developed for voiced fricatives which are very problematic for TD-PSOLA. Segments are chosen in terms of their f0 and duration and the target f0 and duration to avoid or minimise ST-signal repetition, which causes local periodicity and may be perceived as buzzyness.

Together, these three developments of a speech corpus, a signal processing distortion measure, and a selection process for voiced fricatives, provide a framework for reducing perceptible distortion in speech that is pitch-modified using TD-PSOLA. The use of this framework is illustrated in Chapter 6 and its success is evaluated.

Chapter 6. Evaluation of the novel corpus design and signal processing measure

6.1 Introduction

For current text-to-speech systems the use of a signal processing algorithm, such as TD-PSOLA, to modify the pitch and duration of existing speech segments is unavoidable. Previous experiments have shown that certain segments remain relatively unaffected in terms of introduced distortion when pitch-manipulated using TD-PSOLA. In Chapter 5, a novel speech corpus design, and a signal processing distortion measure, for use with TD-PSOLA were developed. In addition, in Section 5.6.1, a special pitch selection method for voiced fricatives was presented. In this chapter, a listening experiment is carried out to evaluate the success of

1. the novel speech corpus, designed to retain the naturalness of synthetic speech by reducing potential distortion,
2. the signal-processing distortion measure, and hence illustrate its potential usefulness for inclusion as part of a target cost estimation in a unit selection process,
3. the rules for the selection process for voiced fricatives.

6.2 Design

This section states the experimental hypotheses and details the structure of the experiment.

6.2.1 Hypotheses

H1: The signal processing costs of test stimuli and their resulting distortion levels when TD-PSOLA pitch-modified will be positively correlated.

H2: Stimuli synthesised using segments from the novel corpus will suffer significantly less distortion than stimuli synthesised using segments from a phonetically balanced corpus.

H3: Stimuli containing voiced fricatives and synthesised using the voiced fricative selection method will suffer significantly less distortion than those synthesised using the standard selection method.

6.2.2 Structure of Experiment

The listening test was designed to evaluate the ability of the signal-processing measure to predict the resulting amount of distortion in individual stimuli, depending upon their phonetic content at sentence level. The success of the signal processing measure was evaluated by calculating the signal processing cost of each sentence level stimulus and comparing this to its resulting distortion level.

The dependent variable *distortion* was measured on a MOS scale, which had been used successfully in the previous experiments, to facilitate the measurement of various amounts of distortion for different values of signal processing cost.

A test was required to describe the relationship between signal processing cost and distortion. The data were ordinal, so a non-parametric Spearman's rho correlation was performed. A one-tailed test was used as the hypothesis predicts a positive correlation.

The listening test was also designed to evaluate the success of the voiced fricative pitch selection method. Sentences containing voiced fricatives were resynthesised using the special selection method and the dependent variable *distortion* was judged on a MOS scale. The IV *synthesis method* had two levels of *v-fricative selection* and *standard selection.* A test was required to compare the differences between the medians of the two IV levels to determine whether stimuli synthesised using this novel method suffered significantly less distortion than those synthesised by the standard method. The data were ordinal, so a within-subjects Wilcoxon Signed Ranks test was performed. A one-tailed test was used, as the hypothesis was directional.

The experiment also evaluated the success of the novel speech corpus design, tailored to the requirements of TD-PSOLA, compared to the performance of a phonetically balanced corpus. The aim of the novel design is to provide segments for TD-PSOLA pitch modification that result in a less distorted output.

The corpora were simulated for this experiment. Synthesising stimuli from the different corpora, one representing the TD-PSOLA balanced corpus and one representing the phonetically balanced design, allowed the resulting distortion levels in the stimuli to be compared. The independent variable *corpus design* has two levels of 'TD-PSOLA-balanced' and 'phonetically-balanced' design. The dependent variable *distortion* was measured on a MOS scale. A statistical test was required to compare the differences between the medians of the two IV levels to determine whether speech synthesised using the novel corpus suffered less distortion than speech synthesised from the phonetically balanced corpus. The data were ordinal, so a within-subjects Wilcoxon Signed Ranks test was performed on the non-parametric data. The hypothesis predicted that the TD-PSOLA balanced corpus would perform better than the phonetically balanced corpus, so a one-tailed test was used.

For all parts of the experiment, participants rated the variable *distortion* for each stimulus using the definitions from previous experiments.

6.3 Stimuli

Initially, the test sentences were designed. Parts of sentences were used to minimise the influence of incorrect prosody on the experimental results. They were designed with fundamental intonation contours (Dutoit, 1997), covering types of sentences such as questioning, ordering, echoing, wh-questioning (who, what, where, why questions), exclamations, implications, and finality, which are defined on a grid of four levels covering approximately one octave. These contours can be distinguished by their slope and curvature, and initial and final pitch levels. This ensured that most of the pitch levels within the speaker's vocal range and various contours were tested.

These sentences were then recorded and analysed for pitch using the Praat software. The pitch contours of the sentences were stylised, as not all pitch movements are perceptible. The results are shown in Table 6.1.

SENTENCE	SEGMENTS AND PITCH CONTOURS						
Take it.	teI 355-299	eIk 299-276	kI 190	It 190			
My cat?	maI 360-430	aIk 430-300	k{ 180-238	{t 238-353			
Prove it.	pr -	ru: 338-338	u:v 338-223	vI 195	It 195		
Look Here!	lU 359	Uk 425	hI@ 341-383 383-341	I@ 341-224 224-185			
Evidently....	Ev 196	vI 383	Id 383	dE 406-213	En 213-163	nt -	li: 158
That's okay.	D{ 272	{t 272	ts -	s@U 210-190	@uk 190-172	keI 239-193	
Who's there?	hu: 265-289	u:z 289-309	De@ 309-255	e@ 255-173			
Measure them?	mE 258	EZ 258-180	Z@ 200	DE 200-342	Em 342-444		
What dog?	wQ 378-411	Qt 411-325	dQ 218	Qg 183			
Three fish?	Tr 439-423	ri: 423-222	fI 168-189	IS 189-292			

Table 6.1 Segments and Target F0 Values

It was not practical to test every phoneme or combination of phonemes in each word position in the sentences due to the unfeasibly large set of stimuli, which may introduce uncontrolled variables due to listener fatigue and boredom during the test. The sentences were designed to include the majority of consonants and vowel phonemes in a random word and phrase position. The intonation contours were randomly assigned to each sentence and hence each phoneme had a random target pitch and contour.

6.3.1 Simulating the Corpora

The TD-PSOLA tailored corpus and the phonetically balanced corpus were simulated for this experiment to minimise the effect of variables present in a real speech corpus. Variables such as necessary duration modification, phoneme context, spectral mismatches etc. may affect the results and mask the effects under investigation. The phonetically balanced corpus design and the TD-PSOLA balanced corpus design were simulated by recording two segment inventories, each consisting of segments recorded at 1, 5 and 10% from the target pitches of the segments in the sentences to be synthesised.

The first corpus (representing the phonetically balanced corpus) consisted of sentences having f0 contours at 1, 5, or 10% below the target f0 contour, depending on their frequency of appearance in the English language. An assumption is made that increased frequency of occurrence increases the variability of segments with different pitch contours (an assumption which the corpus-based approach itself relies on). For example, /@/ the most common phoneme (Crystal, 1995) may be selected with an f0 of 1% lower than the target, whereas /U@/, the least common phoneme may be selected with an f0 of 10% lower.

The second corpus (representing the TD-PSOLA balanced corpus) consisted of segments having f0 values of 1, 5, and 10% below the target values depending on the values given in Table 6.3. For example, affricatives may be selected at 10% below target f0 and voiced fricatives at 1% below target.

A third corpus was simulated to evaluate the success of the signal processing distortion measure and the voiced fricative selection method. It consisted of segments having pitch contours 5% below the target f0 contours.

Prior to the recording of each of the CV, VC, and CC segments, pitch prompts consisting of the same pitch contours as the segments in the original sentences were generated. These were modified using the Praat software to 1, 5 and 10% below the target contours and played to guide the pitch of the speaker when recording the segments to be included in the corpora. All segments were recorded in the same phonetic contexts as they appear in the test sentences to reduce the

effect of variables, such as spectral mismatches between joining segments, or pronounced differences in duration between segment and target values. Such variables may otherwise affect the results of the experiment.

6.3.2 Sentence-level Stimuli

Using segments from these inventories, four sets of sentences were synthesised:

- Set1: using segments based on their phonetic balance at 1, 5 or 10% below the target f0, representing a phonetically balanced corpus. TD-PSOLA was applied to the sentences to achieve the target pitch contour.
- Set2: using segments of 1, 5 or 10% below target f0 based on their TD-PSOLA distortion rating, representing a TD-PSOLA balanced corpus. TD-PSOLA was applied to the sentences to achieve the target pitch contour.
- Set3: using segments all at 5% below target f0. These were used to evaluate the validity of the signal processing distortion measure. The average costs were calculated for each of the sentence level stimuli. The variable *average signal-processing cost* was calculated using Equation 6.1.

$$AC = \frac{1}{N}\sum W_{phoneme} D_{f0} \qquad \text{Eqn 6.1}$$

where N is the number of segments in the sentence, $W_{phoneme}$ is the weighting for the individual phonemes, and D_{f0} is the absolute distance in Hz of the segment f0 to the target f0. The average cost was calculated to allow comparisons between stimuli with differing numbers of segments.

- Set4: using segments all at 5% below target. The rules (described in Section 5.6.1) for the selection of voiced fricatives were applied to 5/10 of the test sentences that contained voiced fricatives, to evaluate the success of these rules.

The experimental stimuli consisted of short sentences, synthesised by concatenation of CV, VC and CC waveforms from each of the inventories. For each set of sentences, the concatenation point for each waveform was chosen by an iterative synthesis/adjustment method. The experimenter created each sentence using subjective optimal segment concatenation points to

produce the most natural sound in terms of concatenation smoothness and natural sentence rhythm. The synthesis methods for each individual set of sentences is described below:

Set 1. This set of sentences represents a phonetically balanced corpus. Table 6.2 shows the frequency of occurrence of phonemes in spoken text (taken from Fry, 1947), and the corresponding percentage f0 below the target f0 the phoneme may be selected for this set of stimuli.

PHONEME	FREQ IN SPOKEN TEXT	CORPUS REPRESENTATION
VOWELS		SET 1
@	10.74%	1%
I	8.33%	1%
E	2.97%	5%
aI	1.83%	10%
V	1.75%	10%
eI	1.71%	10%
i:	1.65%	10%
@U	1.51%	10%
{	1.54%	10%
Q	1.37%	10%
O:	1.24%	10%
u:	1.13%	10%
U	0.86%	10%
A:	0.79%	10%
aU	0.61%	10%
3:	0.52%	10%
e@	0.34%	10%
I@	0.21%	10%
OI	0.14%	10%
U@	0.06%	10%
CONSONANTS		
n	7.58%	1%
t	6.42%	1%
d	5.14%	1%
s	4.81%	5%
l	3.66%	5%
D	3.56%	5%

r	3.51%	5%
m	3.22%	5%
k	3.09%	5%
w	2.81%	5%
z	2.46%	5%
v	2.00%	10%
b	1.97%	10%
f	1.79%	10%
p	1.78%	10%
h	1.46%	10%
N	1.15%	10%
g	1.05%	10%
S	0.96%	10%
j	0.88%	10%
dZ	0.60%	10%
tS	0.41%	10%
T	0.37%	10%
Z	0.10%	10%

Table 6.2 Frequencies of Occurrence of Phonemes in Spoken Text and Corpus Representation

The phoneme representations in the simulated corpus were chosen at 10% below target frequency for phonemes that occur less than 2% of the time, 5% below target for phonemes that occur between 2 and 5%, and 1% below target for phonemes that occur more than 5%.

The sentence level stimuli were created using the two-stage pitch modification process described in Experiment 3, Section 4.4.3. Joining segments were first modified to an intermediate f0, midway in absolute value between the two, using TD-PSOLA. They were then concatenated by cutting the segments at the stable parts of the phonemes, and abutted. The final target pitch contour was then applied using TD-PSOLA.

Set 2. This set of sentences represents the TD-PSOLA balanced corpus. Table 6.3 shows the phoneme representations in the TD-PSOLA corpus, based on the analysis in Section 5.5, with the percentage f0 below the target f0 that the phonemes may be selected at for this set of stimuli.

The two-stage pitch modification process used for Set 1 sentences was used to create the Set 2 stimuli.

PHONEME CATEGORY	% DISTORTION DETECTION	CORPUS REPRESENTATION
CHECKED VOWEL	27%	10%
MONOTHONG VOWEL	50%	10%
DIPHTHONG VOWEL	69%	5%
PLOSIVE	13%	10%
UNVOICED FRICATIVE	14%	10%
AFFRICATIVE	44%	10%
NASAL	56%	10%
LIQUID	69%	5%
GLIDE	69%	5%
VOICED FRICATIVE	94%	1%

Table 6.3 Phoneme Representation in the TD-PSOLA Balanced Corpus

The distribution of the corpus representation between the frequencies of 1, 5 and 10% below target were chosen to reflect the number of phonemes in each % group in the simulated phonetically balanced corpus. More specifically, for the phonetically balanced corpus and the TD-PSOLA balanced corpus respectively, 5 and 4 phonemes are represented at 1% below target, 9 and 10 at 5% below target, and 30 and 30 at 10% below target.

Set 3. This set was used for the evaluation of the signal processing distortion measure. As all segments were recorded at 5% below the target values, a one-stage pitch modification process was employed; segments were concatenated, taking care to avoid pitch mismatches when joining dynamic pitch contours, and then the final pitch contour modification was made using TD-PSOLA.

For the sentences in Set 3, the signal processing cost for each sentence was calculated using the weights determined in Section 5.6 and shown in Table 5.6. The weights were multiplied by the absolute f0 of 5% or 13Hz for each phoneme. The average cost for each sentence was calculated

to allow comparisons between sentences having different numbers of phonemes. Table 6.4 shows each sentence and its average signal processing cost.

Sentence	Signal processing cost
"Take it."	11.00
"My cat?"	14.04
"Prove it."	14.30
"Look here!"	13.16
"Evidently..."	13.13
"That's okay."	13.94
"Who's there?"	21.58
"Measure them?"	16.41
"What dog?"	11.62
"Three fish?"	10.35

Table 6.4 Stimuli and Signal Processing Costs

Set 4. This set was used to evaluate the success of the voiced fricative selection method. The five sentence-level stimuli from Set 3 that contained voiced fricatives were used, and duration modifications were made to the voiced fricative segments in the sentences using TD-PSOLA. The duration modifications were calculated using Equations 5.4 and 5.6 in Section 5.6.1, and applied as described there. As the duration of the segment will remain essentially the same, i.e. the actual duration of the segment and the duration of the target value is the same, the equation to calculate the duration positions simplifies to the ratio expressed in Equation 6.2.

$$position = f0_{segment} / f0_{target} \qquad \text{Eqn. 6.2}$$

The values of the duration positions were rounded down to two decimal places to ensure no ST-signal repetition would occur, which may not be the case if the numbers were rounded up. Table 6.5 shows the five sentences containing the voiced fricatives, the identity of the voiced fricatives, their f0 contours, the target f0 contours, and the duration point positions for each phoneme. No value was assigned to the phoneme /Z/ in the sentence "measure them?" as the pitch detection algorithm in Praat did not detect this phoneme as voiced; no pitch modification would be performed on that part of speech, hence no necessary duration modification.

Sentences	"Evidently..."	"Measure them?"		"Prove it."	"That's ok."	"Who's there?"	
V. fricatives	/v/	/Z/	/D/	/v/	/D/	/z/	/D/
F0 contours	368	X	187-252	237-191	259	291-296	296-242
Target f0	383	X	200-265	250-204	272	304-309	309-255
Duration pts.	0.96	X	0.93-0.95	0.94-0.93	0.95	0.95-0.95	0.95-0.94

Table 6.5 F0 Contours and Duration Points for Voiced Fricatives

6.4 Procedure

Prior to the experiment, participants were given a set of instructions (Appendix C), which were explained to them. The participants were then given a short training session before the experiment began. The training session involved the presentation of four CVC syllable examples, two that had not been manipulated by the TD-PSOLA algorithm and two that had been judged to be very distorted in a previous experiment. Participants were told the change in voice quality, in terms of buzzyness perceived, between the two pairs of segments was called *distortion*. This also provided a range for the distortion they may encounter in the test stimuli.

Participants were then collectively tested using a standardised procedure. The experiment was carried out using the C++ automated interface (Appendix A).

Thirty-five short sentences were presented in a random order via headphones. After each presentation, a judgement concerning the level of distortion was made using the 5-point MOS scale.

6.5 Participants

Nine participants took part in this experiment and all were university staff. This restricted sample number and population was due to the constraints of cost and availability. Participants ranged from 25 – 52 years of age and were from both genders (5 male, 4 female). All had self-reported normal hearing. Some had experience of previous experiments performed during the course of this research. This experiment was performed several months after the previous ones, so it is assumed participants experienced minimal 'training effects'. They were not made aware of the purpose of the experiment but were asked to judge solely the amount of perceived distortion

present in each sentence. The participants were familiarised with speech test procedures and the definition of distortion prior to the test. They were not paid to participate in the test.

6.6 Test Conditions

Test conditions were controlled for as in previous experiments.

6.7 Results

6.7.1 Results of the Evaluation of the Signal Processing Measure

Table 6.6 shows the summary statistics for the evaluation of the signal processing measure. The sentences are given with their signal processing cost estimation and their average MOS ratings.

Sentence	Signal processing cost	MOS rating
"Take it."	11.00	1.33
"My cat?"	14.04	1.89
"Prove it."	14.30	2.44
"Look here!"	13.16	2.00
"Evidently..."	13.13	2.33
"That's okay."	13.94	3.56
"Who's there?"	21.58	1.89
"Measure them?"	16.41	3.11
"What dog?"	11.62	1.56
"Three fish?"	10.35	1.78

Table 6.6 Summary Statistics: Sentences, Costs, and MOS Ratings

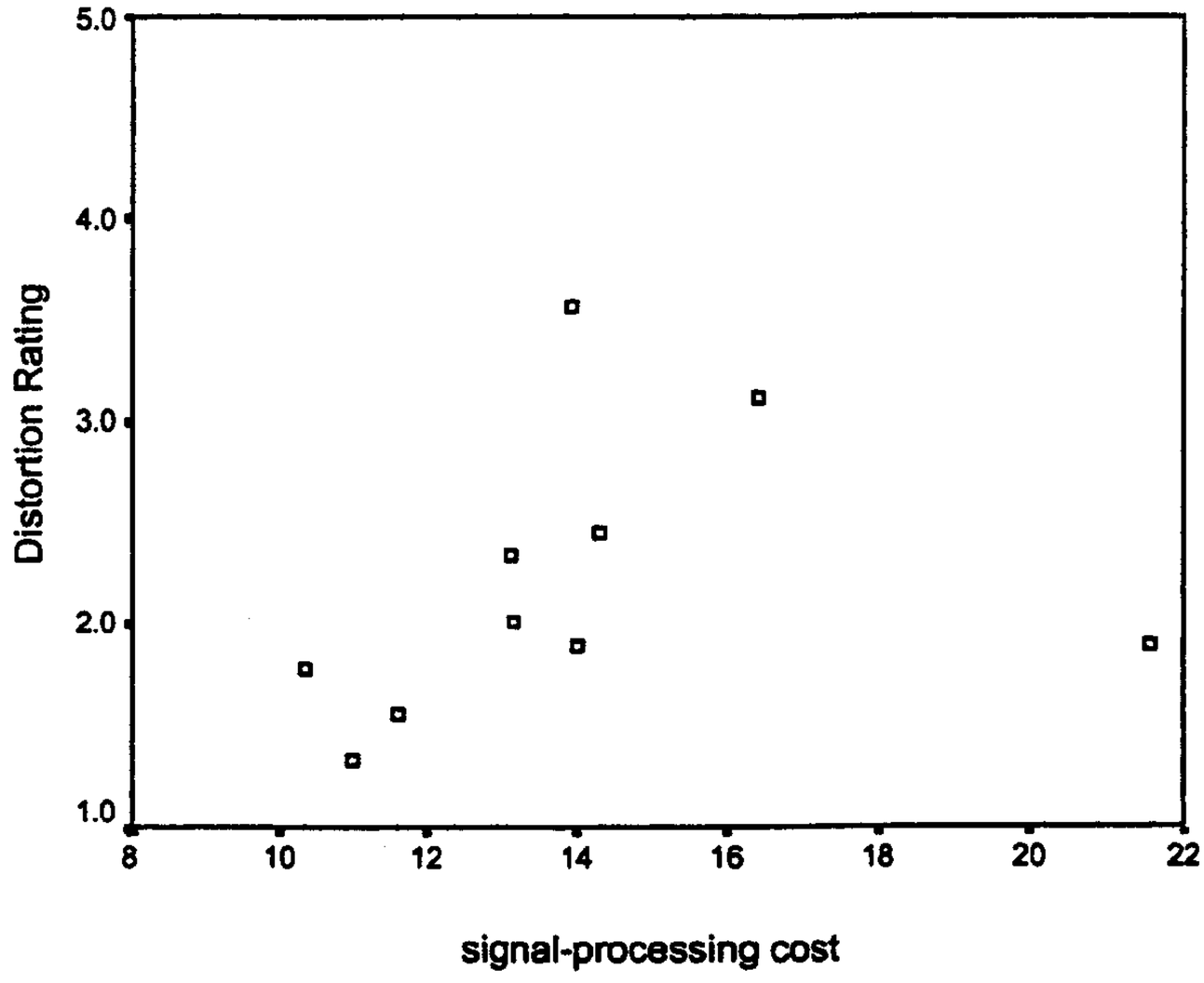


FIGURE 6.1 SCATTERGRAM OF MOS RATINGS AND SIGNAL PROCESSING COSTS

Figure 6.1 shows the scattergram illustrating the relationship between the MOS scale ratings and the signal processing cost calculated for each sentence. A Spearman's rho was calculated for the data and found to be significant (rho=0.58, N=90, $p<0.05$, *one-tailed test*), supporting the hypothesis H1, which stated that the signal processing cost was positively correlated with distortion levels. The scattergram shows that there may an outlier in the data, which would affect the value of rho. The outlier is the sentence "Who's there?" which was given a high signal processing cost but was judged as relatively undistorted. Potential reasons for this are discussed in Section 6.8.

6.7.1.1 Results for the Evaluation of the Voiced Fricative Selection Method

Table 6.7 shows the summary statistics of each sentence containing a voiced fricative, and the corresponding MOS ratings for synthesis using the proposed special-case selection method, and for standard selection.

Sentence	MOS rating (v. fricative selection)	MOS rating (standard selection)
"Prove it."	2.11	2.44
"Evidently..."	2.11	2.33
"That's okay."	2.44	3.56
"Who's there?"	1.67	1.89
"Measure them?"	3.00	3.11

Table 6.7 Summary Statistics: MOS Rating for Voiced Fricative Selection Methods

These data are illustrated in Figure 6.2, which shows the median distortion ratings, averaged over each sentence, for the voiced fricative selection method and the standard selection method.

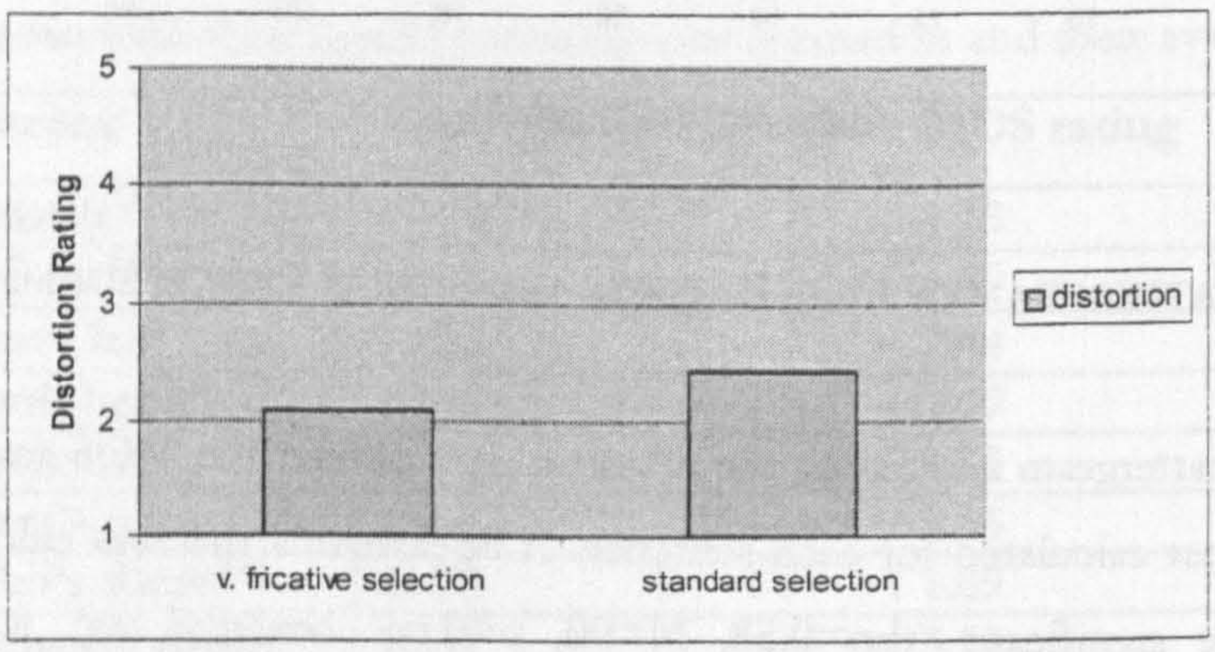


FIGURE 6.2 BARCHART OF DISTORTION FOR VOICED FRICATIVE SELECTION METHODS

A Wilcoxon Signed Rank test was performed, which indicated there was a significant difference between the medians of the distortion levels for the two selection methods ($Z=-2.325$, $N=9$, $p<0.05$, *one-tailed test*), supporting hypothesis H3, which stated that stimuli containing voiced fricatives, synthesised using the special voiced fricative selection method, suffer less distortion than those synthesised using the standard method.

6.7.2 Results of the Evaluation of the TD-PSOLA Balanced Corpus

Table 6.8 shows the average MOS ratings of distortion levels in sentence level stimuli synthesised using segments from either the phonetically balanced corpus or the TD-PSOLA balanced corpus.

If confidence limits were set on the data, of a difference of 0.3 being noteworthy, the TD-PSOLA balanced corpus produces a less perceived distorted output than the phonetically balanced corpus for 6 samples. Of the remaining 4 samples, there is little difference between 3 and the phonetically balanced corpus performs better for 1 sample. These results are discussed in Section 6.8.

Sentence	MOS rating (phonetically balanced corpus)	MOS rating (TD-PSOLA balanced corpus)
"Take it."	3.00	1.33
"My cat?"	1.67	1.56
"Prove it."	2.89	2.67
"Look here!"	3.33	3.00
"Evidently..."	2.67	2.33
"That's okay."	4.56	4.00
"Who's there?"	2.89	2.11
"Measure them?"	4.00	3.33
"What dog?"	2.67	2.56
"Three fish?"	2.78	3.44

Table 6.8 Summary Statistics: MOS Ratings for Phonetically and TD-PSOLA Balanced Corpus Stimuli

The results are illustrated in the barchart in Figure 6.3 which shows the median distortion ratings, averaged over the sentences, for the stimuli synthesised using segments from the phonetically balanced and the TD-PSOLA balanced corpora.

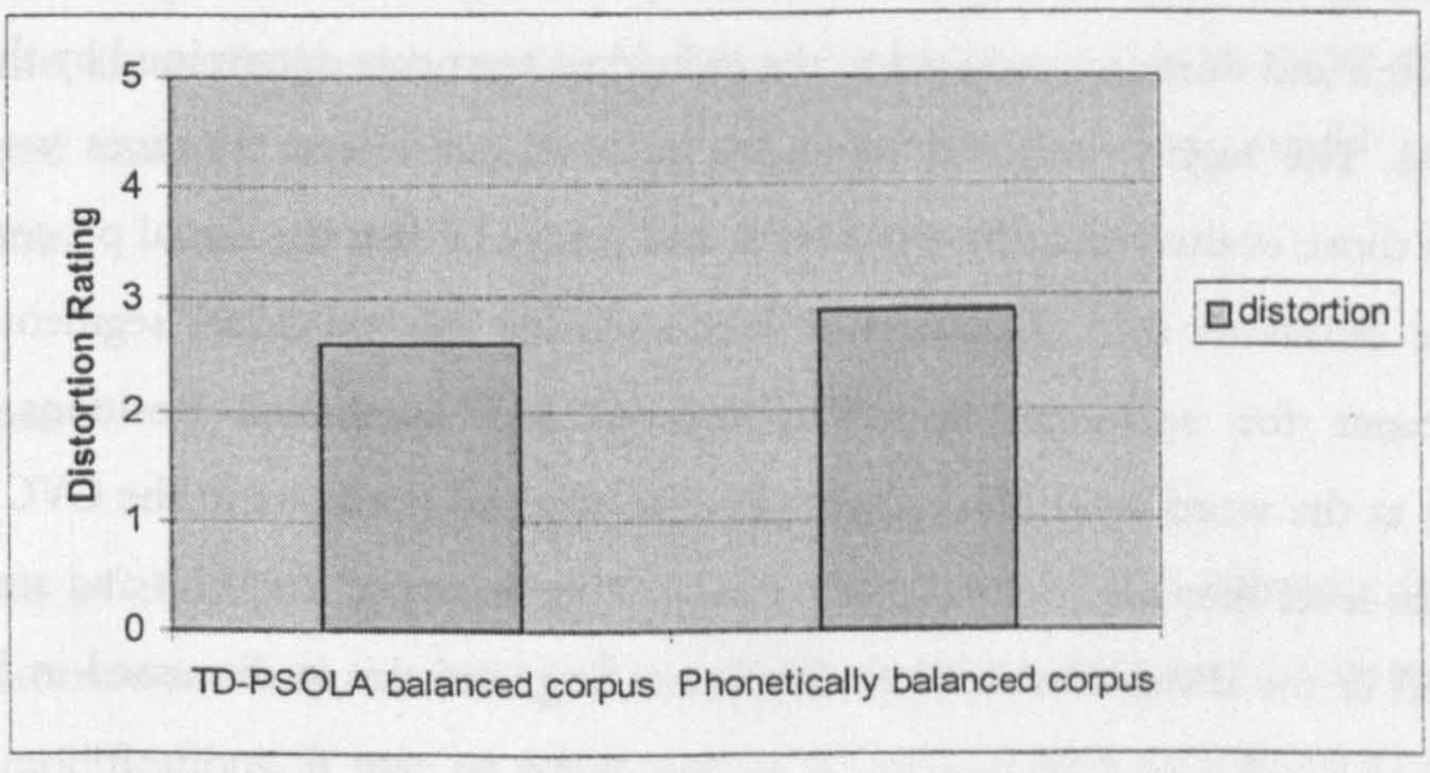


FIGURE 6.3 DISTORTION LEVELS FOR STIMULI SYNTHESISED FROM TWO CORPORA

A Wilcoxon Signed Rank test indicated that medians of the distortion levels were significantly different for the two corpora (Z=-2.670, N=9, $p<0.05$, *one-tailed test*), supporting hypothesis H2 that there is a significant difference between the distortion levels for the phonetically balanced corpus and the TD-PSOLA balanced corpus.

6.8 Discussion

Concerning the evaluation of the signal processing distortion measure, the correlation of 0.58 illustrates that there is a relationship between the signal processing cost and the resulting perceived distortion levels when the sentences are pitch-modified using TD-PSOLA. The correlation value has dropped from 0.71 for consonants and 0.73 for vowels, determined for word-level CVC syllables in Section 5.6. This may be due to the synthesis of dynamic pitch contours for sentence level stimuli which appear to be less problematic for TD-PSOLA. It may also be due to the effect of duration at sentence level. For example, /r/ in "Prove it" is given a score determined from the MOS scores for the initial consonant of the CVC segment "ran". The duration of /r/ in the sentence is of much shorter duration than in the CVC syllable, suggesting that the measure perhaps should be adjusted accordingly.

The data were found to contain an outlier (the sentence "Who's there?"), which had a high cost but was judged as relatively undistorted. The unexpected result for this stimulus may be due to the random pitch and duration assigned to the individual segments determined by the sentence to be synthesised. The highly weighted segments, such as the voiced fricatives were of shorter duration than those evaluated at the word level, and it may be that the signal processing measure needs to take durations into consideration when scoring the individual segments. There is a similar argument for segments found in stressed and unstressed positions. During the investigations at the word-level, all segments were in stressed positions in the CVC syllables, but at the sentence-level they can be in either position. The factors of duration and stress may need to be included in the distortion measure for fine-tuning, and this is discussed in Section 7.2 as further work.

The evaluation of the TD-PSOLA balanced corpus illustrated that sentences synthesised using the TD-PSOLA balanced corpus were significantly less distorted than those synthesised using the phonetically balanced corpus. Having set confidence limits on the data, 6 out of 10 samples suffered less perceived distortion, 3 showed little difference, and 1 showed more distortion. The results may be affected by the type of sentence and hence different stress and intonation contours. Segments assigned stressed positions may not respond similarly as when found in unstressed positions. The analysis in Section 5 was performed on phonemes in stressed positions, which was assumed to give the worst case scenario for the introduction of perceptible distortion. This may reduce the effect seen when using sentence level stimuli with segments in both stressed and unstressed positions, and would require further work to determine its importance.

It should also be noted that this simulation represents a best-case scenario where segments are available from the corpus at near target f0 contours. In reality, having more of certain segments in a corpus may not necessarily mean all pitch contours are represented as near as 1% from the target, but this result has illustrated the potential of the approach. Although the effect using a real corpus may not reduce distortion by such a large amount, it is still expected to provide a significantly less distorted output than when using a phonetically balanced corpus.

The voiced fricative selection method significantly reduced distortion at the sentence level for such phonemes. This approach may be extended for use when increasing and decreasing the pitch of all voiced segments that require the repetition and deletion of ST-signals. In a speech corpus, segments are available with varying pitches and duration, so it is probable that a segment may be selected that, by combining pitch and duration modification, would lead to a less distorted output.

6.9 Conclusions

The correlation of the signal processing cost and MOS scale ratings suggests that the signal processing distortion measure is a valid indictor of the distortion that may be introduced during certain pitch modifications. It may be advantageous to include such a measure in target costing during a segment selection process, to take into account the effect of signal processing when selecting segments, to reduce potential distortion. The measure requires fine-tuning, possibly by

taking into account durations and stress of segments when appearing in different contexts, which may require different costs.

The results illustrated that sentences synthesised using the voiced fricative selection method (requiring simultaneous duration modification) were perceived as significantly less distorted than sentences synthesised using the standard method (applying no duration modification).

The results also illustrated that sentences synthesised using segments from the TD-PSOLA balanced corpus were perceived to be significantly less distorted than sentences synthesised from the phonetically balanced corpus.

Overall, the use of a TD-PSOLA balanced corpus, and voiced fricative selection method provides a framework for generating TD-PSOLA modified speech with reduced distortion. The signal processing distortion measure was able to predict distortion to some certain extent, and the results indicate its potential use in a segment selection process, when selecting segments for pitch modification with TD-PSOLA.

Chapter 7. Conclusions and Further Work

7.1 Conclusions

In this thesis, a framework for the generation of pitch-modified speech with reduced distortion has been implemented and evaluated. This work was motivated by Kortekaas & Kohlrausch (1997a) who investigated the perceptual effects of the TD-PSOLA algorithm on single formant stimuli, and van Santen (1997) who stated that signal processing is still unavoidable even with a large speech corpus.

The first part of this work consisted of a set of investigative experiments to determine the effect of the TD-PSOLA algorithm on natural speech stimuli, in terms of the distortion that is perceived as buzzyness. The first experiment evaluated the effect of greater pitch manipulation on the perceived amount of distortion, concluding that greater pitch manipulation may lead to significantly greater distortion. This indicated the necessity of keeping the amount of signal processing applied to a minimum. In addition, modifications of as low as 1% may have led to perceptible distortion in certain stimuli.

The effect of positive versus negative manipulation over a small pitch modification range, such as often required in a speech corpus system, was then investigated. It was found to be significantly similar; segments can be selected which are either below or above the pitch of the target value with no expected difference in introduced distortion levels. There was some indication that negative modifications may be slightly more problematic with respect to the overall average for all stimuli, which may be contrary to some past research. In addition, individual stimuli appeared to respond differently to positive and negative modifications.

The third experiment looked at the effect of pitch manipulation using TD-PSOLA on distortion levels in synthetic speech at the sentence level. Sentences were synthesised from two inventories, one containing segments with monotone f0 values, and the other containing segments with f0 values closer to the target values. The results were not significant indicating little difference between the two sets of stimuli in terms of distortion. The lack of significance was thought to be

due to the smaller effect size of these data, and that the small effect size was due to the application of dynamic pitch contours for sentence-level stimuli as opposed to static pitch contours, which had been investigated previously at the word level. Dynamic contours appear to lessen the adverse effect of pitch modification, in terms of introducing less perceptible distortion.

The fourth experiment investigated the effect of pitch manipulation in speech for various voices. Certain voices were found to respond better to the application of the algorithm than others with some evidence to suggest that female voices suffered more. This highlighted the need for careful speaker selection when recording a speech inventory or corpus for synthesis using TD-PSOLA.

In Experiments 1 and 4, a significant effect of phoneme identity on distortion levels was found, suggesting that the composition of the speech sound may be responsible for some of the distortion introduced. In addition, there was a significant correlation between distortion levels of individual phonemes for certain voices in Experiment 4; results for one voice may be generalised to an extent to other voices, especially ones of the same gender or of similar neutral f0.

Throughout the course of the experiments, some anomalously high distortion levels occurred, suggesting that aspects of the original recording were a factor in the resulting success of the TD-PSOLA modified stimuli. The fifth experiment investigated this issue. A parameter of "waveform asymmetry" was identified, which may have led to significantly higher distortion for 9 of 13 sets of stimuli. Other possible causes of this anomalous distortion, such as phonation type, low HNR and incorrect pitch marking, were investigated. Although not experimentally verified, creaky voice and incorrect pitch detection (the marking of unvoiced speech as voiced) may have contributed to some of the anomalous occurrences of distortion.

The data gathered during these experiments were analysed for patterns of co-occurrence and correlations. Groups of speech sounds determined by their manner of articulation were found to respond similarly to the algorithm, in terms of perceived distortion. The inherent characteristics of phonemes in the groups were similar duration, f1 shape (if applicable), and voiced/unvoiced/mixed composition (if applicable).

Existing corpus designs and speech segment selection processes were then reviewed, highlighting the fact that the effect of the signal processing algorithm is not taken into consideration in the design of the corpus. The TD-PSOLA algorithm affects certain types of speech sounds more than others, but such effects are often not taken into account during existing segment selection processes. A signal processing distortion measure was developed, which was weighted for different phonemes, trained using experimental data. The signal-processing measure developed here could be included in a unit selection measure to allow candidate segments to be chosen that would significantly reduce distortion when using the TD-PSOLA algorithm to modify the speech. Segment selection using this measure would encourage selection of adversely affected segments closer to the target values, with less weighting given to other selection features, where *necessary*.

A novel corpus was designed, tailored to the use of TD-PSOLA. Purely phonetically balanced corpora do not take into consideration the needs of the signal processing algorithm that may be applied for fine prosody modifications. Experimental data were used to determine which segments required greater representation, and in more varied contexts in the corpus to reduce potential distortions.

A major problem for TD-PSOLA was increasing the pitch of voiced fricatives where the repetition of ST-signals created a buzzy characteristic. A selection process for voiced fricatives was developed to prevent repetition of ST-signals and retain speech segment quality. The selection involved choosing speech sounds of longer duration to avoid ST-repetition, which may be especially pertinent for static pitch signals, where the effect of ST-repetition appears to contribute to the largest amounts of perceptible distortion.

The validity of the signal processing distortion measure, and the success of the voiced fricative selection method and the TD-PSOLA balanced corpus, were evaluated in a final listening test. The TD-PSOLA balanced corpus was found to produce a significantly less distorted output than a phonetically-balanced corpus design. Sentence-level stimuli synthesised using the voiced fricative selection method were significantly less distorted than sentences synthesised using the standard method. The signal processing distortion measure was able to predict the resulting

distortion, illustrating its potential use as part of a target cost estimation in a unit selection process.

In a TTS system, the use of a signal processing distortion measure as part of a segment selection process, used in conjunction with a TD-PSOLA balanced corpus, would provide an approach to concatenative synthesis using TD-PSOLA, which could significantly reduce distortion levels in the speech output.

7.2 Further Work

The framework for minimising the distortion introduced for TD-PSOLA pitch-modification of speech requires further development. Although the small set of training data has been successful in developing the signal processing distortion measure and the novel corpus design, it may be possible to fine-tune these with greater amounts of, and more specific, training data.

Allophonic variations of speech segments have not been explicitly investigated; only initial consonants and mid-vowels in a CVC structure have been investigated and these results generalised to other word positions. It was assumed that the results could be extended to other contexts, as the analysis was based on the inherent characteristics of speech sounds such as voiced/unvoiced/mixed composition.

It was uncertain as to whether participants were judging solely the perceived distortion of the C or V phoneme under investigation, or whether it was additionally the distortion caused by the influences of C and V phonemes on each other in the CVC syllables. An informal experiment was carried out to determine the effect of plosives on different following vowels, and although indications suggested that they had little effect, a more extensive investigation is needed.

The effect of segment identity on distortion levels was investigated for the worst-case scenario; the phonemes evaluated at the word level had long durations, were in stressed positions and with minimal coarticulation, so any other contexts encountered are expected to be less problematic. When the results for static pitch stimuli at the word level are compared to dynamic contour stimuli at the sentence level, the effect of pitch manipulation on distortion is lessened, which

supports this. It should be noted though, that static pitch contours often occur during natural speech alongside dynamic contours.

Closer investigation into the effect on individual phonemes may be undertaken to tune the signal processing distortion measure and corpus design further. Individual segments could be analysed rather than segments grouped according to manner of articulation, as there may be variations in characteristics within the groups. Larger numbers of individual segments would need to be investigated in terms of their responses at each of the pitch manipulation levels so they can be more accurately described for each level of manipulation. For example, a certain phoneme may be able to withstand 5 % manipulation, whereas another may be able to sustain 15%. The measure and corpus design could be refined in this way.

Experiment 2 found that increasing and decreasing pitch manipulation had a similar effect on distortion levels in vowels sounds over the small range measured. It is necessary to test this over a larger pitch manipulation range to investigate whether there is a general loss in voice quality for individual identity phonemes when decreasing pitch rather than increasing or vice versa. For larger modifications the cues of distortion become less important but the perception of loss of voice quality becomes more evident. Determining an allowable maximum level of manipulation in both modification directions, whilst retaining natural voice quality, would be advantageous.

For the development of a complete signal processing measure of distortion and speech corpus design, a duration modification distortion measure is also required. This was partly addressed in the investigation of the effect of pitch manipulation, which also involves increased duration when raising the pitch of voiced parts of speech and decreased duration when lowering the pitch of speech. The pitch manipulation and duration manipulation costs for each segment could then be combined into the overall framework.

The development of a special-case voiced fricative selection process was found to produce significantly less distortion in the output than the standard algorithm. This method avoided the repetition of ST-signals, which may cause local periodicity that is perceived as buzzyness. The algorithm may be extended to the pitch modification of all voiced segments for both positive and

negative modifications to avoid or reduce the repetition and deletion of ST-signals. Further investigation into the use of this and issues such as limits of pitch and duration modifications would be required. An assessment of the trade-off between the complexity of implementing this algorithm and the reduction in distortion could be undertaken. Perhaps it would only be advantageous for the more adversely affected segments such as voiced fricatives and diphthongs.

During this study, one implementation of the TD-PSOLA algorithm (the Praat implementation) has been investigated. Some of the distortions may be peculiar to this particular algorithm, such as deficiencies in the pitch-marking algorithm, or perhaps choice and size of analysis window. It may be necessary to test more implementations to allow the results from this work to be generalised.

During the course of this research, issues were raised that were unfortunately outside the scope of the thesis. Factors for successful voice selection for synthesis systems, and determination of aspects of segment recordings that affect the success of TD-PSOLA, provide interesting areas for further work.

To conclude, the framework developed during this research provides a method to generate pitch-modified speech using the TD-PSOLA algorithm, with reduced distortion. Whilst distortion still occurs in the signals, there exists aspects of this work that may be improved further. In fact, whilst signal processing algorithms are still necessary in speech synthesis systems, the goal of reducing signal processing distortion remains a challenging and important area of research for the speech community as a whole.

Appendices

Appendix A. Code and Interface

The following Figures show screenshots of the C++ software used to automate the listening experiments undertaken during this work. Figure A.1 shows the initialisation of the software, enabling the experimenter to select the number of stimuli for each experiment.

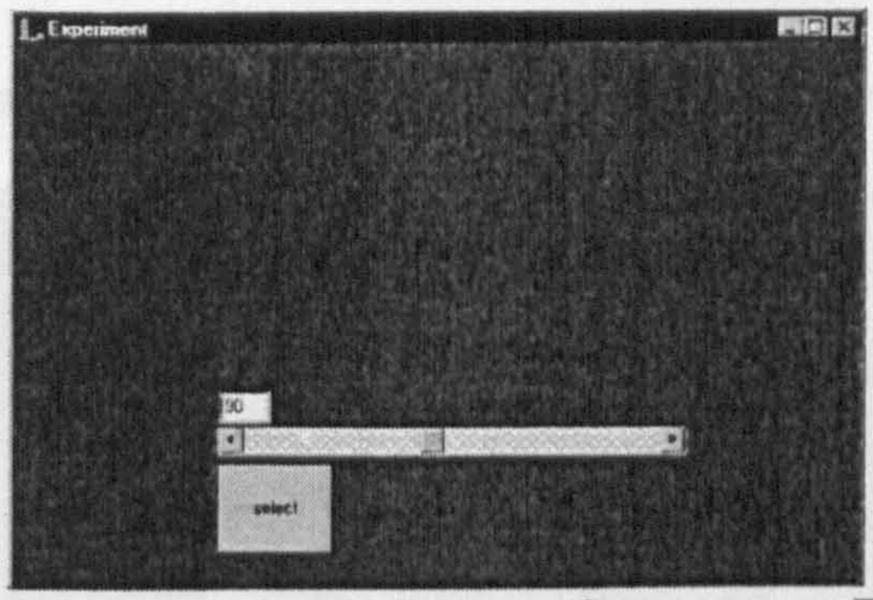


FIGURE A.1 SELECTION OF NUMBER OF STIMULI FOR EXPERIMENT

Figure A.2 shows the interface for Experiments 1, 2, 3, 4 & 6. The Play Stimulus button must be selected by the participant to hear each stimulus. The MOS scale buttons, labelled 1 to 5, are then enabled, allowing the participant to record their judgement of that stimulus. Once a judgement has been made (by selecting a button between 1 and 5), the Play Stimulus button is re-enabled, allowing the next stimulus to be played.

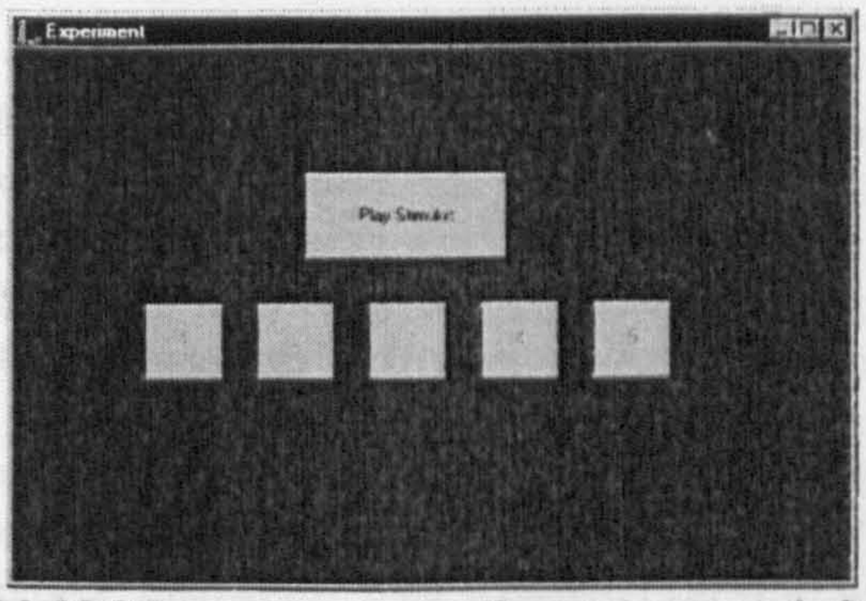


FIGURE A.2 MOS INTERFACE FOR EXPERIMENTS 1, 2, 3, 4 AND 6

Figure A.3 shows the interface for Experiment 5. The Play button must be selected by the participant to hear a stimulus. The Yes/ No buttons are then enabled, allowing the participant to record whether or not they perceived any distortion in that stimulus. Once a judgement has been made (by selecting either button), the Play button is re-enabled, allowing the next stimulus to be played.

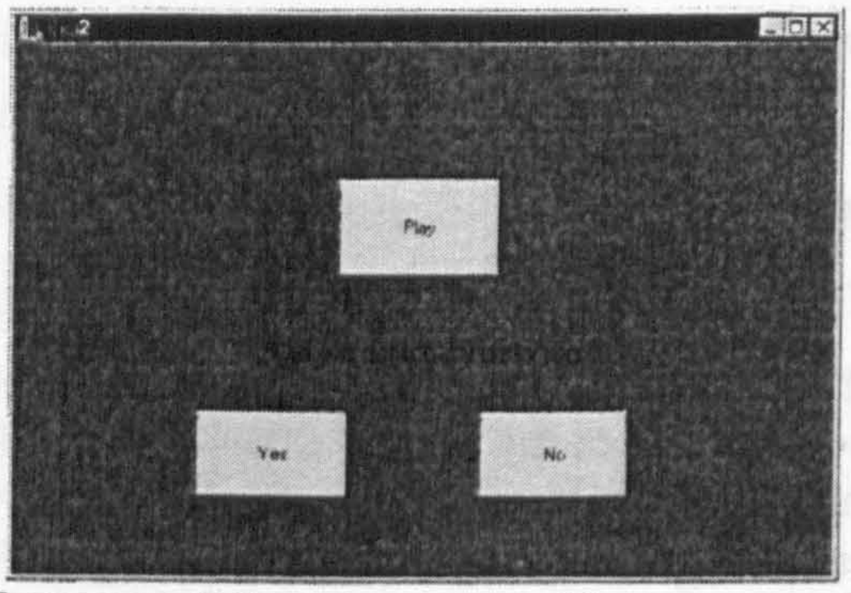


FIGURE A.3 INTERFACE FOR EXPERIMENT 5

Figure A.4 shows the Form "Experiment" used to create the visual interface for the C++ code.

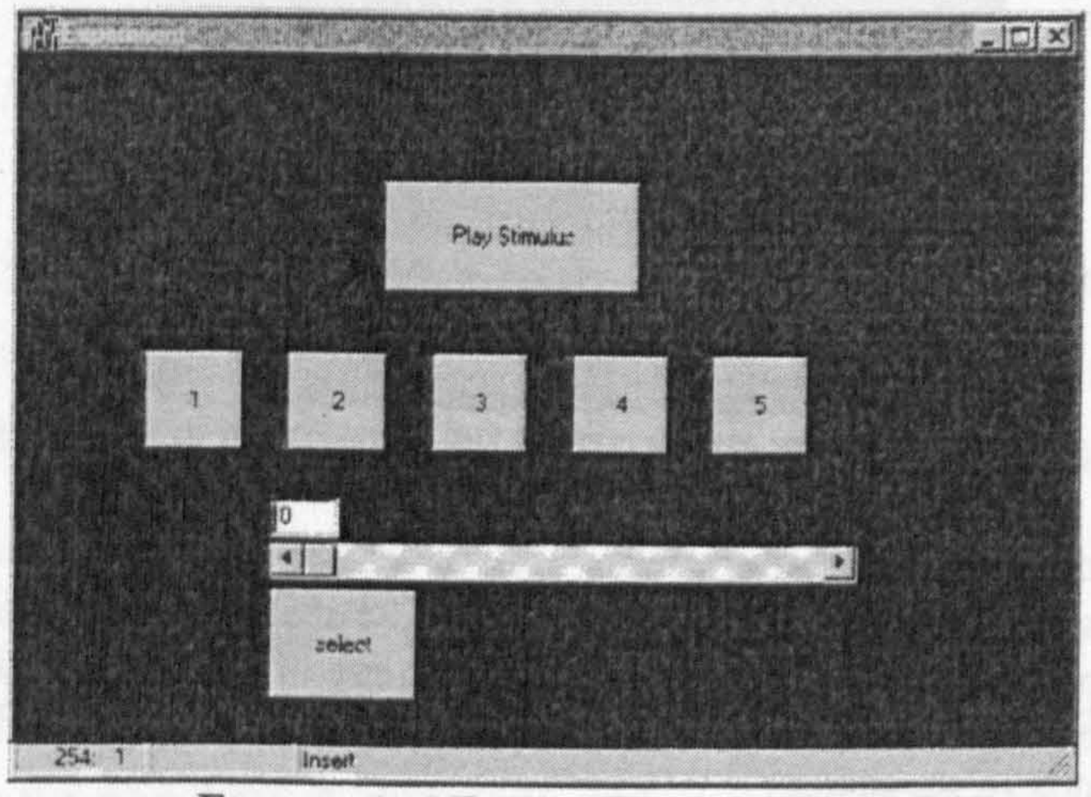


FIGURE A.4 FORM "EXPERIMENT"

The following C++ code was written using the Borland C++ Builder 3 environment. It provides the experimenter with the ability to initialise the system with the number of stimuli for each experiment. The names of the .wav files of each stimulus are read in from a floppy disk and stored in a playlist. The participant is then presented with the interface shown in Figure A.2 and must click on the Play Stimulus button to hear the first stimulus. A random number is then generated to select a random stimulus form the playlist. This ensures that for each test run, the order of presentation of the stimuli is different. The stimulus is played once only. The Play Stimulus button is then disabled, preventing the user from playing another stimulus at this stage. The MOS scale, consisting of buttons labelled 1 to 5, is enabled. The participant must then make a judgement about the stimulus just played by clicking one of the buttons. When this has been done, the number of the button clicked (the MOS rating) is then associated with the stimulus just played to be output to a results file when the experiment is completed. The MOS buttons are then disabled to prevent another judgement to be made at this stage. The Play Stimulus button is re-enabled to allow the participant to play the next stimulus and so on.

File: listening_experiment.cpp

```
//---------------------------------------------------------------------------
#include <vcl.h>
#pragma hdrstop
USERES("listening_experiment.res");
USEFORM("exp.cpp", Experiment);
USEUNIT("exp_engine.cpp");
//---------------------------------------------------------------------------
WINAPI WinMain(HINSTANCE, HINSTANCE, LPSTR, int)
{
  try
  {
    Application->Initialize();
    Application->CreateForm(__classid(TExperiment), &Experiment);
    Application->Run();
  }
  catch (Exception &exception)
  {
    Application->ShowException(&exception);
  }
  return 0;
}
//---------------------------------------------------------------------------
```

File: exp.h

```
//---------------------------------------------------------------------------
#ifndef expH
#define expH
//---------------------------------------------------------------------------
#include <Classes.hpp>
#include <Controls.hpp>
#include <StdCtrls.hpp>
#include <Forms.hpp>
#include "exp_engine.h"
//---------------------------------------------------------------------------
class TExperiment : public TForm
{
__published:       // IDE-managed Components

   TScrollBar *stim_num;          //Scroll bar to choose number of stimuli for experiment
   TEdit *display;                //Display of number of stimuli determined by position of scroll bar
   TButton *Select_stim_num;      //Button to select number of stimuli for experiment

   TButton *MOS_1;                //Buttons to select MOS rating of 1 to 5
   TButton *MOS_2;
   TButton *MOS_3;
   TButton *MOS_4;
   TButton *MOS_5;

   TButton *play_stim;            Button to play stimulus

   void __fastcall FormCreate(TObject *Sender);

   //methods performed when scroll bar position altered or buttons clicked

   void __fastcall stim_numChange(TObject *Sender);
   void __fastcall Select_stim_numClick(TObject *Sender);
   void __fastcall play_stimClick(TObject *Sender);
   void __fastcall MOS_1Click(TObject *Sender);
   void __fastcall MOS_2Click(TObject *Sender);
   void __fastcall MOS_3Click(TObject *Sender);
   void __fastcall MOS_4Click(TObject *Sender);
   void __fastcall MOS_5Click(TObject *Sender);

private:  // User declarations

 Experimental Exp;          //instantiate object Exp from class Experimental, providing
                            //stimuli attributes and methods

public:             // User declarations
   __fastcall TExperiment(TComponent* Owner);
};
//---------------------------------------------------------------------------
extern PACKAGE TExperiment *Experiment;
//---------------------------------------------------------------------------
#endif
```

File: exp.cpp

```
//---------------------------------------------------------------------------
#include <vcl.h>
#pragma hdrstop

#include "exp.h"
//---------------------------------------------------------------------------
#pragma package(smart_init)
#pragma resource "*.dfm"
TExperiment *Experiment;
//---------------------------------------------------------------------------
__fastcall TExperiment::TExperiment(TComponent* Owner)
   : TForm(Owner)
{
}
//---------------------------------------------------------------------------

void __fastcall TExperiment::FormCreate(TObject *Sender)
{
   play_stim->Visible=false;          //create form and initialise buttons used during
   MOS_1->Visible=false;              //experiment to non-visible. Visible buttons
   MOS_2->Visible=false;              //used for selection of number of stimuli
   MOS_3->Visible=false;              //for experiment
   MOS_4->Visible=false;
   MOS_5->Visible=false;
}
//---------------------------------------------------------------------------

void __fastcall TExperiment::stim_numChange(TObject *Sender)
{                                                    //when scroll bar is altered to select number of stimuli
   display->Text = IntToStr(stim_num->Position);     //display is updated to show number depending on
}                                                    //position of scroll bar
//---------------------------------------------------------------------------

void __fastcall TExperiment::Select_stim_numClick(TObject *Sender)
{                                    //function called when number of stimuli for experiment is selected
  int no_of_stims;
  no_of_stims=stim_num->Position;
  Exp.initialise(no_of_stims);       //read in stimuli list from floppy disk

   Select_stim_num->Visible=false;//form components not needed for experiment made non-visible
   display->Visible=false;
stim_num->Visible=false;

play_stim->Visible=true;           //button to play stimulus visible and enabled

MOS_1->Visible=true;               //MOS buttons visible (only enabled after each stimulus played)
MOS_1->Enabled=false;
MOS_2->Visible=true;
MOS_2->Enabled=false;
MOS_3->Visible=true;
MOS_3->Enabled=false;
```

```
   MOS_4->Visible=true;
   MOS_4->Enabled=false;
   MOS_5->Visible=true;
   MOS_5->Enabled=false;
}
//---------------------------------------------------------------------------

void __fastcall TExperiment::play_stimClick(TObject *Sender)
{                                       //when stimulus play button is clicked.....
int stim_num;
stim_num=Exp.get_stim_num();     //get random stimulus number from list of stimuli

Exp.play(stim_num);        //play stimulus

    play_stim->Enabled=false;        //play stimulus button disabled and MOS rating buttons enabled
    MOS_1->Enabled=true;
    MOS_2->Enabled=true;
    MOS_3->Enabled=true;
    MOS_4->Enabled=true;
    MOS_5->Enabled=true;
}
//---------------------------------------------------------------------------

void __fastcall TExperiment::MOS_1Click(TObject *Sender)
{                                       //when MOS rating button 1 is clicked.....
  int number,stim_total;
  Exp.store_num(1);        //associate MOS rating of 1 with current stimulus just played

number=Exp.get_counter();        //get count of stimuli already played
stim_total=Exp.get_stim_total(); //get total number of stimuli in experiment

if(number>=(stim_total-1))       //if end of experiment, make MOS buttons non-visible
{                                //and save all data to floppy disk
    MOS_1->Visible=false;
    MOS_2->Visible=false;
    MOS_3->Visible=false;
    MOS_4->Visible=false;
    MOS_5->Visible=false;

    Exp.save_stuff();
}
else            //if not end of experiment, enable button to play next stimulus and disable MOS buttons
{
    play_stim->Enabled=true;
    MOS_1->Enabled=false;
      MOS_2->Enabled=false;
      MOS_3->Enabled=false;
      MOS_4->Enabled=false;
      MOS_5->Enabled=false;

      Exp.inc_counter();     //increment the number of stimuli played so far during experiment run
   }
}
//---------------------------------------------------------------------------
```

```
void __fastcall TExperiment::MOS_2Click(TObject *Sender)
{
  int number,stim_total;

  Exp.store_num(2);
  number=Exp.get_counter();
  stim_total=Exp.get_stim_total();
  if(number>=(stim_total-1))
  {
     MOS_1->Visible=false;
     MOS_2->Visible=false;
     MOS_3->Visible=false;
     MOS_4->Visible=false;
     MOS_5->Visible=false;

     Exp.save_stuff();
}
else
{
     play_stim->Enabled=true;
     MOS_1->Enabled=false;
     MOS_2->Enabled=false;
     MOS_3->Enabled=false;
     MOS_4->Enabled=false;
     MOS_5->Enabled=false;

     Exp.inc_counter();
}
}
//---------------------------------------------------------------------------

void __fastcall TExperiment::MOS_3Click(TObject *Sender)
{
  int number,stim_total;

  Exp.store_num(3);

  number=Exp.get_counter();
  stim_total=Exp.get_stim_total();
  if(number>=(stim_total-1))
  {
     MOS_1->Visible=false;
     MOS_2->Visible=false;
     MOS_3->Visible=false;
     MOS_4->Visible=false;
     MOS_5->Visible=false;

     Exp.save_stuff();
}
else{
     play_stim->Enabled=true;
     MOS_1->Enabled=false;
     MOS_2->Enabled=false;
     MOS_3->Enabled=false;
     MOS_4->Enabled=false;
     MOS_5->Enabled=false;
```

```
      Exp.inc_counter();
   }
}
//---------------------------------------------------------------------------

void __fastcall TExperiment::MOS_4Click(TObject *Sender)
{
  int number,stim_total;
  Exp.store_num(4);

  number=Exp.get_counter();
  stim_total=Exp.get_stim_total();
  if(number>=(stim_total-1))
  {
     MOS_1->Visible=false;
     MOS_2->Visible=false;
     MOS_3->Visible=false;
     MOS_4->Visible=false;
     MOS_5->Visible=false;

     Exp.save_stuff();
  }
  else{
     play_stim->Enabled=true;
     MOS_1->Enabled=false;
     MOS_2->Enabled=false;
     MOS_3->Enabled=false;
     MOS_4->Enabled=false;
     MOS_5->Enabled=false;

     Exp.inc_counter();
   }
}
//---------------------------------------------------------------------------

void __fastcall TExperiment::MOS_5Click(TObject *Sender)
{
  int number,stim_total;
  Exp.store_num(5);

  number=Exp.get_counter();
  stim_total=Exp.get_stim_total();

  if(number>=(stim_total-1))
  {
     MOS_1->Visible=false;
     MOS_2->Visible=false;
     MOS_3->Visible=false;
     MOS_4->Visible=false;
     MOS_5->Visible=false;

     Exp.save_stuff();
  }
  else{
```

```
        play_stim->Enabled=true;
        MOS_1->Enabled=false;
        MOS_2->Enabled=false;
        MOS_3->Enabled=false;
        MOS_4->Enabled=false;
        MOS_5->Enabled=false;

        Exp.inc_counter();
    }
}
//---------------------------------------------------------------------------
```

File: exp_engine.h

```
#ifndef exp_engineH
#define exp_engineH

class Experimental
{
private:
 struct              //List holds stimuli waveform paths and filenames and respective MOS rating
  {
    char Filename [50];   //path and name of wav files
    int answer;           //MOS rating
  } List [1000];  //max of 1000 stimuli may be used in experiment

   int counter;           //current number of stimuli played so far in experiment
   int random_check [1000];       //used in random number generation.
                                  //Ensures each stimulus played only once
   int stimulus;          //current stimulus identity number
   int total;             //total number of stimuli in experiment

public:
   void initialise (int no_of_stimuli);        //initialise stimuli playlist
   void play (int num);              //play stimulus
   int get_counter (void);                     //get current number of stimuli played so far
   void inc_counter (void);          //increment current number of stimuli played so far
   int get_stim_num (void);          //generate random stimulus identity number
   void save_stuff(void);            //save MOS data to floppy disk
   void store_num(int score);        //store MOS rating with respective stimulus filename
   void set_stim_total(int no_of_stims);       //set total number of stimuli in experiment
   int get_stim_total(void);         //get total number of stimuli in experiment
};

//---------------------------------------------------------------------------
#endif
```

File: exp_engine.cpp

```
#include <vcl.h>
#pragma hdrstop

#include "exp_engine.h"
#include <mmsystem.h>

#include <stdlib.h>
#include <conio.h>
#include <iostream.h>

#include <fstream.h>
//---------------------------------------------------------------------------

void Experimental::initialise(int no_of_stimuli)        //initialises stimuli playlist
{
int i;
counter=0;

for (i=0; i<no_of_stimuli; i++)       //initialise random checker used during random num generation
    random_check[i]=i;

ifstream inFile("a:\\stim.dat",ios::in);          //open file of stimuli path and filenames

if(!inFile)
{
  cerr<<"File could not be opened";
  exit(1);
}

i=0;                  //read in wav file path and filenames and store in List structure, Filename field
while(inFile>>List[i].Filename)
   i++;

   inFile.seekg(0);
   inFile.close();

set_stim_total(no_of_stimuli);        //set the number of stimuli in the experiment
}
//---------------------------------------------------------------------------

void Experimental::play(int num)
{
PlaySound(List[num].Filename,NULL,SND_SYNC);          //play stimulus
}

int Experimental::get_counter(void)
{
  return counter;             //get current count of stimuli played so far
}
//---------------------------------------------------------------------------
```

```
void Experimental::inc_counter(void)
{
  counter++;                    increment current count of stimuli played so far
}
//---------------------------------------------------------------------

int Experimental::get_stim_num(void)          //generated random stimulus identity number
{
  int num, i=0, flag=0;
  srand(time(NULL));        //seed random number generator from PC clock
  num=rand()%total;         //random number between 0 and (total number of stimuli in experiment-1)

do
  {
   if (i==total)     //if random number has been generated before, generate a new number
   {
    i=0;
    num=rand()%total;
   }

   if (num==random_check[i])       //if stimulus identity number has not been generated before.....
   {
     random_check[i]=999999;       //set element in array to 999999 so it cannot be chosen again
     flag=1;
   }
   i++;

  }while (flag==0);          //repeat until stimulus identity number not previously chosen is generated

  stimulus=num;

  return num;       //return stimulus identity number to be played
}
//---------------------------------------------------------------------

void Experimental::save_stuff(void)
{
   ofstream outfile("a:\\results.dat");            //save stimuli filenames and respective MOS ratings to floppy disk
   int i;
   for (i=0;i<total;i++)
     outfile<<List[i].Filename<<" "<<List[i].answer<<endl;
   outfile.close();
  }
//---------------------------------------------------------------------

void Experimental::store_num(int score)
{
   List[stimulus].answer=score;       //store MOS rating for stimulus just played in structure List
}
//---------------------------------------------------------------------

void Experimental::set_stim_total(int no_of_stims)
{
   total=no_of_stims;          //sets the total number of stimuli in the experiment
}
//---------------------------------------------------------------------
```

```
int Experimental::get_stim_total(void)
{
   return total;                //gets the total number of stimuli in the experiment
}
//-------------------------------------------------------------------
#pragma package(smart_init)
```

Appendix B. String Lists for Experiments

B.1 CVC Syllables with Varying Central Vowel for Experiment 1 and 2

/k{t/	/kEt/	/kIt/	/kQt/	/kUt/	/kVt/
/k@t/	/ki:t/	/keIt/	/kaIt/	/kOIt/	/ku:t/
/k@Ut/	/kaUt/	/k3:t/	/kA:t/	/kO:t/	kI@t/
/ke@t/	/kU@t/				

B.2 CVC Syllables with Varying Initial Consonant for Experiment 4

/p{n/	/b{n/	/t{n/	/d{n/	/k{n/	/g{n/
/tS{n/	/dZ{n/	/f{n/	/v{n/	/T{n/	/D{n/
/s{n/	/z{n/	/S{n/	/Z{n/	/h{n/	/m{n/
/n{n/	/r{n/	/l{n/	/w{n/	/j{n/	

B.3 Sentence-Level Stimuli for Experiment 3

"My cat?"
"Prove it."
"No way!"
"Look here....."

B.4 CVC Syllables for Experiment 5

B.4.1 Vowel Stimuli

/k{t/	/kIt/	/kQt/	/kA:t/	/kaIt/	/ku:t/

B.4.2 Consonant Stimuli

/d{n/	/s{n/	/tS{n/	/n{n/	/r{n/	/j{n/
/D{n/					

B.5 Sentence-Level Stimuli for Experiment 6

"Take it."
"My cat?"
"Prove it."
"Look here!"
"Evidently....."
"That's okay."
"Who's there?"
"Measure them?"
"What dog?"
"Three fish?"

Appendix C. Instructions for Experiments

Instructions for Listening Experiment No. ....

Do you to the best of your knowledge, have normal hearing?

In this experiment you will hear 96 mono-syllabic syllables via headphones, and will be asked to give your opinion of the speech you hear.

Prior to the start of the experiment, you will be provided with examples of the criterion under investigation and familiarised with the testing procedure.

Experiment Procedure

Using the computer interface provided, press the 'Play Stimulus' button to hear the speech. You will hear the stimulus only once, then you must press the appropriate button on the interface (1-5) to indicate your opinion using the following scale:

AMOUNT OF DISTORTION PRESENT IN STIMULI

1 no perceived distortion

2 quite undistorted

3 distorted

4 quite distorted

5 very distorted

After a short pause, you will be able to play the next stimulus.

Thank you for your participation in this experiment.

BLANK IN ORIGINAL

Appendix D. Experimental Data

	% PITCH MODIFICATION			
PARTICIPANTS	0%	5%	10%	15%
1	2.15	3.60	3.35	4.15
2	2.15	2.75	2.70	3.15
3	1.45	2.25	2.60	3.10

D.1 EXPERIMENT 1: PILOT STUDY DATA (positive modification only)

	% PITCH MODIFICATION						
PARTICIPANTS	-15%	-10%	-5%	0%	+5%	+10%	+15%
1	3.60	4.00	3.75	2.15	3.60	3.35	4.15
2	3.00	3.00	3.00	2.15	2.75	2.70	3.15
3	2.75	2.80	2.00	1.45	2.25	2.60	3.00

D.2 EXPERIMENT 1: PILOT STUDY DATA

	PITCH MODIFICATION (HZ)				
PARTICIPANTS	220	223	233	246	259
1	1.70	1.85	2.85	2.95	3.35
2	2.00	2.10	3.55	3.30	4.15
3	2.05	2.10	2.65	2.60	3.05
4	1.35	1.55	2.15	2.60	2.90
5	2.10	2.25	3.15	3.30	3.45
6	1.90	2.05	2.95	3.00	3.15
6	1.75	1.65	2.35	2.55	2.90
8	1.25	1.45	2.35	2.85	2.90
9	1.35	1.35	2.15	2.55	3.00
10	1.40	1.55	2.25	2.25	2.80
11	1.70	1.55	2.40	2.75	2.85
12	1.75	1.95	2.90	3.05	3.25
13	1.75	2.00	2.80	2.65	3.20
14	2.40	2.50	3.35	3.40	3.75
15	2.15	2.35	3.15	3.50	3.70

D.3 EXPERIMENT 1: PARTICIPANTS' MOS RATINGS OF DISTORTION AT EACH PITCH MODIFICATION LEVEL

	PITCH MANIPULATION (HZ)				
CVC SYLLABLE	220	223	233	246	259
ki:t	1.73	1.73	2.47	3.47	3.07
kA:t	1.47	1.47	2.33	2.67	3.27
kO:t	1.20	1.73	2.20	2.20	2.13
ku:t	1.40	1.40	3.73	2.87	3.87
k3:t	1.73	1.73	2.87	3.33	3.60
klt	1.60	1.40	2.07	2.20	2.40
kEt	2.00	2.40	2.87	4.27	4.13
k{t	1.87	1.87	2.47	2.73	2.53
k@t	1.53	1.73	2.27	2.53	2.60
kVt	1.73	1.87	2.40	2.13	2.47
kQt	1.40	1.40	1.87	2.00	2.00
kUt	1.73	1.87	2.47	2.40	2.87
kelt	1.40	2.00	2.80	2.87	2.73
kalt	2.13	2.40	3.00	2.73	3.73
kOlt	1.27	1.27	2.80	3.13	3.13
k@Ut	1.53	1.67	2.13	2.67	3.40
kaUt	1.87	1.73	2.27	2.40	3.67
kl@t	2.67	2.67	3.67	3.80	4.53
ke@t	2.67	2.67	4.27	3.67	4.40
kU@t	2.53	2.67	3.73	3.67	4.00

D.4 EXPERIMENT 1: MOS RATINGS FOR INDIVIDUAL CVC STIMULI

	PITCH MODIFICATION (HZ)				
PARTICIPANTS	200	210	220	230	240
1	1.70	1.40	1.10	1.35	2.10
2	2.60	1.80	1.50	1.85	2.00
3	3.05	2.60	1.60	2.85	2.80
4	1.55	1.45	1.15	1.45	2.00
5	2.05	1.80	1.30	1.75	2.15
6	2.50	1.75	1.60	2.00	2.40
7	2.60	1.85	1.30	1.65	2.05
8	2.80	2.45	1.80	2.15	2.85
9	2.40	1.95	1.35	1.90	2.45
10	2.95	2.25	1.70	2.35	2.75

D.5 EXPERIMENT 2: PARTICIPANS' MOS RATINGS AT EACH PITCH MODIFICATION LEVEL

	% PITCH MODIFICATION				
CVC SYLLABLE	-8%	-4%	0%	4%	8%
kA:t	1.40	1.20	1.20	1.30	1.70
k{t	2.30	1.20	1.20	1.20	2.60
k@t	2.10	1.20	1.20	1.30	2.40
kO:t	3.00	1.70	1.20	2.00	3.00
k@Ut	3.30	2.00	1.20	2.30	3.30
kOlt	2.70	1.70	1.20	2.00	3.00
ku:t	3.00	2.60	1.30	2.00	3.30
kUt	2.40	1.70	1.40	2.60	2.00
kVt	2.20	1.60	1.50	2.00	2.10
ke@t	2.30	2.00	1.10	1.70	1.40
kelt	3.00	3.00	1.70	2.00	1.40
ki:t	3.00	2.60	2.00	2.00	3.30
kEt	2.60	2.30	2.00	2.10	2.30
kU@t	1.60	1.90	1.20	2.00	2.30
kI@t	2.70	2.30	1.60	1.60	2.30
kQt	1.90	1.60	1.10	1.70	2.00
klt	2.00	1.70	1.10	1.30	1.70
kalt	2.30	2.30	2.00	2.30	1.70
kaUt	2.60	2.00	2.00	3.20	3.30
k3:t	2.00	2.00	1.60	2.00	2.00

D.6 EXPERIMENT 2: CVC STIMULI AND MOS RATINGS

	INVENTORY 1		INVENTORY 2	
PARTICIPANTS	DISTORTION	HUMANNESS	DISTORTION	HUMANNESS
1	2.50	3.25	2.00	4.25
2	3.25	4.25	2.75	3.50
3	3.25	2.50	2.50	4.00
4	2.50	3.00	2.25	3.75
5	1.75	3.75	2.25	2.75
6	3.00	2.00	3.50	2.00
7	3.00	3.25	2.75	3.50
8	2.50	3.75	2.50	4.25
9	2.25	3.50	1.00	4.75
10	3.25	4.00	3.25	3.50

D.7 EXPERIMENT 3: PARTICIPANTS' MOS RATINGS OF DISTORTION AND HUMANESS FOR TWO INVENTORIES

SENTENCES		
INVENTORY 1	**DISTORTION**	**HUMANNESS**
"My cat?"	2.5	3.7
"No way!"	2.9	3.5
"Look here..."	3.1	2.5
"Prove it."	2.4	3.6
INVENTORY 2	**DISTORTION**	**HUMANNESS**
"My cat?"	2.2	3.8
"No way!"	2.8	3.2
"Look here..."	2.2	3.8
"Prove it."	2.7	3.7

D.8 EXPERIMENT 3: SENTENCE STIMULI AND MOS RATINGS

	PITCH MODIFICATION (HZ)				
PARTICIPANTS	**220**	**223**	**233**	**246**	**259**
1	2.09	2.13	2.26	2.52	2.83
2	1.78	2.13	3.13	3.26	3.35
3	2.57	2.39	2.83	2.74	3.00
4	1.78	1.87	2.91	3.65	3.91
5	2.17	2.35	2.91	2.91	3.17
6	1.57	1.52	2.04	2.17	2.17
7	1.65	1.61	1.78	2.09	2.39
8	1.37	1.59	2.07	2.45	2.82
9	1.43	1.48	1.91	2.48	2.78
10	1.83	1.96	2.48	2.61	2.78
11	1.87	2.09	2.57	2.91	3.35
12	1.43	1.78	2.17	2.35	2.74
13	2.00	2.26	2.78	3.00	3.00
14	2.61	2.30	3.13	3.26	3.39
15	1.78	1.91	2.52	2.78	3.09
16	1.83	1.91	2.26	2.65	2.96
17	1.96	1.96	2.61	2.78	2.96
18	1.74	2.04	2.48	2.57	2.87
19	2.57	2.39	3.09	3.17	3.22
20	2.52	2.48	3.26	3.30	3.39

D.9 EXPERIMENT 4: PARTICIPANTS' MOS RATINGS AT EACH PITCH MANIPULATION LEVEL FOR VOICE 1

	% PITCH MODIFICATION			
CVC SYLLABLE	1%	5%	10%	15%
b{n	2.20	2.15	2.85	2.55
k{n	1.55	1.70	2.20	2.15
d{n	1.60	2.40	2.60	2.70
f{n	2.15	2.30	2.30	2.15
g{n	1.40	2.40	2.70	3.85
h{n	1.85	1.70	2.55	2.85
l{n	2.00	2.15	2.85	2.70
m{n	2.85	3.25	3.15	3.60
n{n	1.55	1.70	2.40	2.85
p{n	1.70	2.85	3.70	4.25
r{n	1.90	3.00	2.30	2.45
s{n	1.40	2.15	2.30	2.30
t{n	1.90	2.40	1.45	2.45
v{n	2.00	3.20	4.00	4.40
w{n	2.40	3.40	4.00	3.70
z{n	2.55	3.00	3.15	3.30
tS{n	1.85	2.80	2.70	2.20
dZ{n	1.60	2.40	1.90	2.85
T{n	2.25	3.25	2.75	3.30
D{n	2.80	3.35	3.45	3.70
S{n	2.00	2.30	2.70	2.70
Z{n	2.85	3.55	3.70	4.00
j{n	2.30	2.15	3.15	3.30

D.10 EXPERIMENT 4: STIMULI AND MOS RATINGS FOR VOICE 1

	PITCH MODIFICATION (HZ)				
PARTICIPANTS	**200**	**203**	**212**	**223**	**235**
1	1.48	2.57	2.83	3.13	3.22
2	1.65	2.43	3.17	3.35	3.39
3	1.96	3.48	4.17	4.17	3.83
4	1.22	2.26	3.26	3.96	4.17
5	1.39	2.26	3.13	3.48	3.22
6	1.26	2.96	3.74	3.96	3.83
7	1.09	2.43	3.35	3.43	3.35
8	1.04	1.91	2.30	2.83	3.09
9	1.26	2.13	3.09	3.26	3.48
10	1.48	2.70	3.30	3.48	3.39
11	1.26	2.65	3.09	3.74	3.39
12	1.22	2.30	3.13	3.61	3.70
13	1.61	2.91	3.57	3.78	3.87
14	1.74	3.22	3.91	4.26	4.04
15	1.22	2.26	3.17	3.61	3.39
16	1.48	2.43	3.17	3.48	3.57
17	1.52	2.48	3.30	3.78	3.57
18	1.30	2.13	2.91	3.17	3.17
19	1.57	3.35	4.00	3.96	3.83
20	1.78	3.30	4.22	4.22	4.09

D.11 EXPERIMENT 4: PARTICIPANTS' MOS RATINGS AT EACH PITCH MANIPULATION LEVEL FOR VOICE 1

	% PITCH MODIFICATION			
CVC SYLLABLE	1%	5%	10%	15%
b{n	2.00	1.85	3.00	2.30
k{n	3.00	3.40	3.85	4.05
d{n	1.85	3.10	3.70	3.15
f{n	2.70	2.70	2.10	2.40
g{n	3.00	3.80	3.55	3.70
h{n	2.80	3.00	3.30	3.30
l{n	3.50	4.65	4.80	4.85
m{n	3.00	3.60	4.05	4.00
n{n	1.70	3.30	3.30	3.70
p{n	2.50	4.25	3.40	4.55
r{n	2.00	2.70	3.00	3.30
s{n	2.00	2.30	3.10	3.15
t{n	3.00	3.40	3.85	3.85
v{n	3.20	3.35	3.85	4.05
w{n	2.70	3.85	4.70	4.25
z{n	2.50	3.30	3.85	3.30
tS{n	2.50	3.00	3.15	3.15
dZ{n	1.70	3.40	3.85	3.74
T{n	1.85	2.40	2.70	3.05
D{n	4.30	4.40	4.05	3.85
S{n	2.40	2.45	3.05	2.60
Z{n	2.50	4.50	4.70	4.20
j{n	3.30	4.15	4.65	4.00

D.12 EXPERIMENT 4: STIMULI AND MOS RATINGS FOR VOICE 2

	PITCH MODIFICATION (HZ)				
PARTICIPANTS	130	133	137	144	152
1	2.17	2.22	2.52	2.65	2.78
2	2.83	2.43	2.96	3.13	3.61
3	2.17	2.13	2.30	2.57	2.65
4	2.09	2.17	2.65	3.13	3.17
5	2.26	2.30	2.83	2.61	2.39
6	2.00	1.91	2.04	2.04	2.26
7	1.91	1.91	2.26	2.35	2.70
8	1.87	1.83	2.35	2.57	2.70
9	2.35	2.48	2.43	2.83	2.83
10	2.13	2.22	2.48	2.74	2.83
11	2.17	2.17	2.48	2.52	2.83
12	2.65	2.74	2.87	2.87	3.17
13	2.74	2.87	2.78	3.30	3.17
14	1.96	1.83	2.35	2.39	2.43
15	2.17	2.13	2.57	2.70	2.70
16	2.13	1.78	2.43	2.65	2.83
17	2.13	1.87	2.26	2.57	2.65
18	2.22	2.22	2.70	2.52	2.70
19	2.74	2.74	3.04	3.04	3.17
20	2.70	2.61	2.96	2.96	3.17

D.13 EXPERIMENT 4: PARTICIPANTS' MOS RATINGS AT EACH PITCH MANIPULATION LEVEL FOR VOICE 3

	% PITCH MODIFICATION			
CVC SYLLABLE	1%	5%	10%	15%
b{n	2.30	2.55	2.50	2.50
k{n	1.95	2.10	1.95	2.40
d{n	1.70	1.70	2.35	1.95
f{n	2.40	2.80	2.80	2.50
g{n	1.40	1.40	2.40	2.25
h{n	2.40	2.40	2.20	1.85
l{n	2.10	2.40	2.60	3.20
m{n	2.00	2.40	2.80	2.80
n{n	2.40	2.40	2.80	2.80
p{n	1.70	2.80	2.20	1.80
r{n	2.20	2.40	3.00	3.40
s{n	2.20	2.00	2.00	3.00
t{n	2.00	2.80	2.00	2.80
v{n	2.00	2.40	2.40	2.20
w{n	2.60	3.00	3.05	2.50
z{n	2.60	3.60	3.60	4.20
tS{n	2.25	2.20	2.40	2.80
dZ{n	2.40	2.40	2.60	2.20
T{n	2.40	2.80	3.20	2.50
D{n	2.65	3.20	3.40	3.00
S{n	2.10	2.20	3.75	4.60
Z{n	2.80	3.60	3.05	3.80
j{n	2.70	3.40	3.20	4.20

D.14 EXPERIMENT 4: STIMULI AND MOS RATINGS FOR VOICE 3

	PITCH MODIFICATION (HZ)				
PARTICIPANTS	**120**	**122**	**127**	**133**	**140**
1	1.61	2.13	2.65	2.96	3.22
2	2.00	2.26	2.57	3.22	3.30
3	1.70	1.96	2.65	3.17	3.61
4	1.52	2.09	3.43	3.35	3.78
5	2.13	2.57	3.39	3.78	3.91
6	1.78	1.83	1.83	2.04	2.30
7	1.43	1.61	2.30	2.65	3.00
8	1.43	1.57	2.22	2.83	3.39
9	1.96	2.30	2.61	3.09	3.57
10	1.91	2.35	3.00	3.26	3.43
11	1.57	1.91	2.48	2.91	3.13
12	2.00	2.65	3.22	3.52	3.52
13	2.22	2.83	3.57	3.96	4.22
14	1.83	2.09	2.78	3.22	3.39
15	1.78	2.04	2.74	2.96	3.30
16	1.74	1.87	2.39	2.83	3.04
17	1.70	2.13	2.61	2.87	3.43
18	1.74	2.04	2.52	3.00	3.09
19	2.35	2.61	3.26	3.43	3.83
20	2.17	2.57	3.26	3.57	3.43

D.15 EXPERIMENT 4: PARTICIPANTS' MOS RATINGS AT EACH PITCH MANIPULATION LEVEL FOR VOICE 4

	% PITCH MODIFICATION			
CVC SYLLABLE	1%	5%	10%	15%
b{n	2.20	2.40	3.00	2.50
k{n	2.15	2.40	2.40	3.20
d{n	1.75	2.00	2.80	3.00
f{n	2.70	2.80	2.40	2.80
g{n	2.10	2.40	3.00	3.80
h{n	1.80	2.80	2.65	3.00
l{n	1.80	2.60	3.20	3.60
m{n	2.00	2.40	2.40	3.00
n{n	1.90	3.00	4.00	3.45
p{n	2.00	3.00	2.80	2.60
r{n	2.05	2.60	3.00	3.25
s{n	3.00	3.80	4.05	4.40
t{n	1.40	2.00	2.60	3.20
v{n	2.75	2.80	3.60	4.20
w{n	2.65	3.40	3.60	3.25
z{n	2.35	2.80	4.00	4.20
tS{n	1.60	2.20	2.40	2.80
dZ{n	2.00	2.80	2.80	3.20
T{n	2.00	3.00	3.80	3.60
D{n	1.60	2.80	3.05	3.40
S{n	2.50	3.40	3.00	3.60
Z{n	3.50	3.80	4.05	4.05
j{n	2.10	2.60	3.40	4.00

D.16 EXPERIMENT 4: CVC STIMULI AND MOS RATINGS OF VOICE 4

PITCH MODIFICATION		CVC SYLLABLE					
	VERSION	/k{t/	/kA:t/	/kQt/	/ku:t/	/klt/	/kalt/
223Hz	1	0	0	0	1	0	0
	2	0	0	0	1	0	1
	3	0	0	1	0	1	0
	4	0	1	0	0	0	1
233Hz	1	0	0	0	1	0	0
	2	0	0	0	1	0	1
	3	0	0	1	0	1	0
	4	0	1	0	0	1	1
246Hz	1	0	1	0	1	0	0
	2	0	0	0	1	0	1
	3	1	1	0	0	1	1
	4	0	1	0	0	1	1
259Hz	1	0	1	0	1	0	1
	2	1	0	0	1	0	1
	3	1	1	1	0	1	1
	4	0	1	0	0	1	1

D.17 EXPERIMENT 5: DISTORTION DETECTION WITH MAJORITY SCORE 5/10 FOR VOWEL STIMULI

PITCH MODIFICATION		CVC SYLLABLE						
	VERSION	/d{n/	/s{n/	/tS{n/	/n{n/	/r{n/	/j{n/	/D{n/
223Hz	1	0	0	0	0	0	1	1
	2	0	1	1	1	0	1	1
	3	0	0	0	1	1	0	1
	4	0	0	1	0	1	0	1
233Hz	1	0	0	0	0	0	1	1
	2	1	1	1	1	0	1	1
	3	0	0	0	1	1	0	1
	4	0	0	1	0	1	0	1
246Hz	1	0	0	0	0	0	1	1
	2	0	0	0	1	1	1	1
	3	0	0	0	1	1	1	0
	4	0	0	1	0	1	0	1
259Hz	1	0	0	0	0	1	1	1
	2	0	1	1	1	1	1	1
	3	1	0	0	1	1	1	1
	4	0	0	1	1	1	1	1

D.18 EXPERIMENT 5: DISTORTION DETECTION WITH MAJORITY SCORE 5/10 FOR CONSONANT STIMULI

		PITCH MODIFICATION (HZ)				
CVC SYLLABLE	VERSION	223	233	246	259	ASYMMETRY
/k{t/	1	1	1	2	3	NO
	2	2	3	3	6	NO
	3	4	4	6	6	YES
	4	3	1	1	3	YES
/kA:t/	1	3	4	5	5	YES
	2	1	2	2	3	NO
	3	4	4	6	5	NO
	4	7	8	7	9	YES
/kQt/	1	0	1	0	0	NO
	2	0	1	0	1	NO
	3	5	5	4	5	YES
	4	2	3	3	2	YES
/ku:t/	1	6	6	5	6	YES
	2	6	5	7	7	YES
	3	4	4	3	4	NO
	4	3	2	1	4	NO
/klt/	1	2	1	2	1	NO
	2	2	1	2	2	NO
	3	8	9	8	8	YES
	4	3	5	6	7	YES
/kalt/	1	4	3	4	5	NO
	2	5	5	7	9	YES
	3	4	3	5	5	NO
	4	6	9	8	9	YES

D.19 EXPERIMENT 5: WAVEFORM ASYMMETRY AND NUMBER OF DISTORTION DETECTIONS FOR VOWEL STIMULI

		PITCH MODIFICATION (HZ)				
CVC SYLLABLE	VERSION	223	233	246	259	ASYMMETRY
/d{n/	1	1	2	0	1	YES
	2	3	5	4	4	YES
	3	4	4	4	5	NO
	4	1	0	1	2	NO
/s{n/	1	1	2	3	2	NO
	2	5	5	4	6	YES
	3	4	3	3	2	YES
	4	1	2	1	2	NO
/tS{n/	1	2	2	3	0	NO
	2	5	6	4	7	NO
	3	1	3	2	4	YES
	4	5	6	7	6	YES
/n{n/	1	2	2	3	4	NO
	2	7	6	7	7	YES
	3	8	6	7	6	YES
	4	3	2	3	6	NO
/r{n/	1	4	3	4	5	NO
	2	4	3	6	5	NO
	3	6	7	6	8	YES
	4	7	7	6	9	YES
/j{n/	1	7	6	7	8	YES
	2	6	8	7	8	NO
	3	4	3	5	5	NO
	4	4	4	4	5	YES
/D{n/	1	8	9	10	9	YES
	2	8	8	9	10	YES
	3	6	7	4	5	NO
	4	7	6	6	8	NO

D.20 EXPERIMENT 5: WAVEFORM ASYMMETRY AND NUMBER OF DISTORTION DETECTIONS FOR CONSONANT STIMULI

STIMULI	WEIGHTINGS			
CVC SYLLABLE	1%	5%	10%	15%
ki:t	-2.73	1.10	1.71	0.83
kA:t	-5.35	0.77	0.78	0.99
kO:t	-2.73	0.47	0.23	0.10
ku:t	-6.06	4.03	1.01	1.45
k3:t	-2.73	2.03	1.55	1.24
kIt	-6.06	0.16	0.23	0.31
kEt	4.04	2.03	2.65	1.66
k{t	-1.31	1.10	0.85	0.41
k@t	-2.73	0.63	0.62	0.47
kVt	-1.31	0.93	0.15	0.37
kQt	-6.06	-0.30	0.00	0.00
kUt	-1.31	1.10	0.47	0.68
keIt	0.00	1.86	1.01	0.57
kaIt	4.04	2.33	0.85	1.34
kOIt	-7.37	1.86	1.32	0.88
k@Ut	-3.33	0.30	0.78	1.09
kaUt	-2.73	0.63	0.47	1.30
kI@t	6.77	3.89	2.10	1.97
ke@t	6.77	5.29	1.95	1.86
kU@t	6.77	4.03	1.95	1.55

D.21 CHAPTER 5: VOWEL WEIGHTINGS FOR EACH PITCH MANIPULATION LEVEL

	WEIGHTINGS			
CVC SYLLABLE	1%	5%	10%	15%
b{n	1.67	0.24	1.15	0.47
k{n	-6.67	-1.09	0.19	0.08
d{n	-6.03	0.98	0.78	0.62
f{n	1.03	0.68	0.34	0.08
g{n	-8.59	0.98	0.93	1.76
h{n	-2.82	-1.09	0.71	0.77
l{n	-0.90	0.24	1.15	0.62
m{n	10.00	3.49	1.60	1.51
n{n	-6.67	-1.09	0.49	0.77
p{n	-4.74	2.31	2.41	2.15
r{n	-2.18	2.75	0.34	0.37
s{n	-8.59	0.24	0.34	0.23
t{n	-2.18	0.98	-0.92	0.37
v{n	-0.90	3.34	2.86	2.30
w{n	4.23	3.93	2.86	1.61
z{n	6.15	2.75	1.60	1.21
tS{n	-2.82	2.16	0.93	0.13
dZ{n	-6.03	0.98	-0.25	0.77
T{n	2.31	3.49	1.01	1.21
D{n	9.36	3.79	2.04	1.61
S{n	-0.90	0.68	0.93	0.62
Z{n	10.00	4.38	2.41	1.90
j{n	2.95	0.24	1.60	1.21

D.22 CHAPTER 5: CONSONANT WEIGHTINGS AT EACH PITCH MANIPULATION LEVEL

	INVENTORY	
PARTICIPANTS	TD-PSOLA BALANCED	PHONETICALLY BALANCED
1	2.70	3.20
2	2.20	2.60
3	2.30	2.50
4	2.00	2.70
5	2.30	3.10
6	2.90	3.40
7	3.00	3.20
8	3.00	3.30
9	3.30	3.40

D.23 EXPERIMENT 6: PARTICIPANTS' MOS RATINGS FOR EACH INVENTORY

	INVENTORY	
SENTENCES	TD-PSOLA BALANCED	PHONETICALLY BALANCED
"Evidently...."	2.33	2.67
"That' okay...."	4.00	4.56
"Look here"	3.00	3.33
"Three fish?"	3.44	2.78
"Prove it"	2.67	2.89
"My cat?"	1.56	1.67
"What dog?"	2.56	2.67
"Measure them?"	3.33	4.00
"Take it"	1.33	3.00
"Who's there?"	2.11	2.89

D.24 EXPERIMENT 6: MOS RATINGS FOR SENTENCE-LEVEL STIMULI FROM EACH INVENTORY

PARTICIPANTS	V/FRICATIVE SELECTION	STANDARD SELECTION
1	1.80	2.40
2	1.60	1.60
3	2.40	2.20
4	2.60	2.80
5	2.60	3.00
6	2.00	2.40
7	2.40	2.80
8	2.60	3.40
9	2.40	3.40

D.25 EXPERIMENT 6: PARTICIPANTS' MOS RATINGS FOR VOICED FRICATIVE AND STANDARD SELECTION METHODS

SENTENCES	V/FRICATIVE SELECTION	STANDARD SELECTION
"Evidently..."	2.11	2.33
"Measure them?"	3.00	3.11
"Prove it."	2.11	2.44
"That's okay."	2.44	3.56
"Who's there?"	1.67	1.89

D.26 EXPERIMENT 6: MOS RATINGS FOR SENTENCES SYNTHESISED USING VOICED FRICATIVE AND STANDARD SELECTION METHODS

Acronyms and Abbreviations

CART	Classification and Regression Trees
CELP	Code Excited Linear Prediction
CLID-test	Cluster Identification test
C	Consonant
DFT	Discrete Fourier Transform
DMOS	Degradation MOS
DRT	Diagnostic Rhyme Test
DSP	Digital Signal Processing
DV	Dependent Variable
EC	Error rate per Contrast
FD-PSOLA	Frequency-Domain PSOLA
FFT	Fast Fourier Transform
F0	Fundamental frequency; the frequency of vibration of the vocal folds
F1, F2, F3....	Formant frequencies
HMN	Harmonic plus Noise Model
HNR	Harmonics-to-Noise Ratio
H/S	Harmonic/ Stochastic model
IPA	International Phonetic Association
IV	Independent Variable
JND	Just Noticeable Differences
LPC	Linear Predictive Coding
LP-PSOLA	Linear Prediction PSOLA
MOS	Mean Opinion Score
MBR-PSOLA	(MBROLA) Multi-band Resynthesis PSOLA
MLPC	Multipulse Linear Predictive Coding
MRT	Modified Rhyme Test
NLP	Natural Language Processing
PC	Paired Comparison
PSOLA	Pitch-Synchronous Overlap-Add
RELP	Residual Excited Linear Prediction
SAMPA	Speech Assessment Methods Phonetic Alphabet
SNR	Signal-to-Noise Ratio
SPL	Signal Pressure Level
ST-signal	Short Term signal
STFT	Short Term Fourier Transform
SUS	Semantically Unpredictable Sentences
TD-PSOLA	Time-Domain Pitch-Synchronous Overlap-Add
ToBI	Tones and Break Indices
TTS	Text-to-Speech
V	Vowel
V/UV	Voiced/Unvoiced

BLANK IN ORIGINAL

References

Allen, J. B. (1977). Short-Term Spectral Analysis, Synthesis and Modification by Discrete Fourier Transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-25, 235 – 238.

Allen, J., Hunnicut, S. & Klatt, D. (1987). *From Text to Speech, The MITALK System.* Cambridge UK: Cambridge University Press.

Balestri, M., Pacchiotti, A., Quazza, P. J. S. & Sandri, S. (1999). Choose the Best to Modify the Least: A New Generation Concatenative Synthesis System. In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Budapest*, 5, 2291-2294.

Bartkova, K. & Sorin, C. (1987). A Model of Segmental Duration for Speech Synthesis in French. *Speech Communication*, 6 (3), 245 – 260.

Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y. & Syrdal, A. (1999a). The AT&T Next-Gen TTS System. In: *Collected Papers of the 137th Joint Meeting of the Acoustical Society of America, European Acoustics Association and DAGA, Berlin.* Paper 2ASCA_4.

Beutnagel, M., Mohri, M. & Riley, M. (1999b). Rapid Unit Selection from a Large Speech Corpus for Concatenative Speech Synthesis. In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Budapest*, 2, 607-610.

Bigorgne, D., Boëffard, O., Cherbonnel, B., Emerard, F., Larreur, D., Le Saint-Milon, J. L., Métayer, I., Sorin, C. & White, S. (1993). Multi-Lingual PSOLA Text-to-Speech System. In: *Proceedings of the International Conference on Acoustics, Speech & Signal Processing '93*, 2, 187-90.

Black, A. W. (2002). Perfect Synthesis for all of the People all of the Time. In: *Keynote, IEEE Text-to-Speech Workshop, Santa Monica, CA.*

Black, A. W. & Campbell, W. N. (1995). Optimising Selection of Units from Speech Databases for Concatenative Synthesis. In: *Proceedings of the European Conference on Speech Communication & Technology (Eurospeech), Madrid*, 1, 581-584.

Black, A. & Lenzo, K. (2000a) *Building Voices in the Festival Speech Synthesis System*, DRAFT [http://www-2.cs.cmu.edu/~awb/]

Black, A. W. & Lenzo, K. A. (2000b). Limited Domain Synthesis. In: *Proceedings of the International Conference on Spoken Language Processing, Beijing*, 2, 411-414.

Black A. W. & Lenzo, K. (2001). Optimal Data Selection for Unit Synthesis. In: *Proceedings of the 4th ESCA Workshop on Speech Synthesis, Scotland.*

Black, A. W. & Taylor, P. (1994). CHATR: A Generic Speech Synthesis System. In: *Proceedings of the International Conference on Computational Linguistics, Japan*, 2, 983-986.

Black, A. W. & Taylor, P. (1997). Automatically Clustering Similar Units for Unit Selection in Speech Synthesis. In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Greece*, 2, 601-604.

Black, A. W., Taylor, P. & Caley, R. (1999). *The Festival Speech Synthesis System – System Documentation.* CSTR Edinburgh. Edition 1.4 for Festival 1.4.0.
[http://www.cstr.ed.ac.uk/projects/festival/manual/festival_toc.html]

Blouin C. J. & Bagshaw, P. C. (2000). Analysis of the Degradation of French Vowels induced by the TD-PSOLA Algorithm, in Text-to-Speech Context. In: *Proceedings of the International Conference on Spoken Language Processing*, 1, 709-712.

Boëffard, O., Miclet, I. & White, S. (1992). Automatic Generation of Optimized Unit Dictionaries for Text to Speech Synthesis. In: *Proceedings of the International Conference on Spoken Language Processing, Canada*, 1211 – 1214.

Boëffard, O., Cherbonnel, B., Emerard, F. & White, S (1993). Automatic Segmentation and Quality Evaluation of Speech Unit Inventories for Concatenation-Based, Multilingual PSOLA Text-to-Speech Systems. In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Berlin*, 1449-1452.

Boersma, P. (1993). Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound. In: *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, 17, 97-110.

Boersma, P. & Weenink, D. (1999). *Praat 3.8.38. A System for doing Phonetics by Computer.* [http://www.fon.hum.uva.nl/praat]

Breen, A. P. (1998). Speech Synthesis. In: *Proceedings of the Institute of Acoustics,* Windemere, UK, 20 (6), 239 – 244.

Breen, A., Bowers, E. & Welsh, W. (1996). An Investigation into the Generation of Mouth Shapes for a Talking Head. In: *Proceedings of the International Conference on Spoken Language Processing,* 4, 12 - 16.

Brieman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees.* Monterey, CA: Wadsworth & Brooks.

Campbell, N. W. (1999). A Call for Generic-Use Large-Scale Single-Speaker Speech Corpora and an Example of their Application in Concatenative Speech Synthesis. *Technical Publications ATR Interpreting Telecommunications Research Laboratories*, 42 – 47.

Campbell, N. & Black, A. (1996). Prosody and the Selection of Source Units for Concatenative Synthesis. In: J. van Santen *et al.* (Eds.) *Progress in Speech Synthesis*, London, UK: Springer Verlag.

Campbell, N. W., Higuchi, N. & Black, A. (1998*). CHATR: A Natural Speech Re-Sequencing Synthesis System*. Draft, April 8, 1998. [http://www.itl.atr.co.jp/chatr/papers.html]

Campos, G. & Gouvea, E. (1996). Speech Synthesis using the CELP Algorithm. In: *Proceedings of the International Conference on Spoken Language Processing '96.*

Carlson, R., Granstrom, B. & Klatt, D. H. (1979). Some Notes on the Perception of Temporal Patterns in Speech. In: B. Lindblom & S. Ohman (Eds.) *Frontiers of Speech Communication Research,* London: Academic Press, 233-243.

CCITT SG12 (1993). Subjective Performance Assessment of the Quality of Speech Output Devices. *Special Rapporteur for Questions 5/XII.*

Chappell, D. T. & Hansen, J. H. L. (1997). An Auditory-Based Measure for Improved Phone Segment Concatenation. In: *Proceedings of the IEEE International Conference on Acoustics, Speech & Signal Processing,* 3, 1639-1642.

Chappell, D. & Hansen, J. H. L. (1998). Spectral Smoothing for Concatenative Speech Synthesis. In: *Proceedings of the International Conference on Spoken Language Processing, Australia,* 5, 1935-1938.

Charpentier, F. & Moulines, E. (1989). Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech using Diphones. In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Paris,* 13-19.

Charpentier, F. & Stella, M. G. (1986). Diphone Synthesis using an Overlap-Add Technique for Speech Waveform Concatenation. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing,* 2015-2018.

Chou, F., Tseng, C. & Lee, L. (1999). Selection of Waveform Units for Corpus-Based Mandarin Speech Synthesis Based on Decision Trees and Prosodic Modification Costs. In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech),* 2295 – 2298.

Clark-Carter, D. (1999). *Doing Quantitative Psychological Research – from Design to Report.* UK: Psychology Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences* (2nd Edn.). NJ: Lawrence Erlbaum Associates.

Conkie, A. (1999). Robust Unit Selection System for Speech Synthesis. In: *Collected Papers of the 137th Meeting of the Acoustical Society of America & the 2nd Convention of the European Acoustics Association, Germany*, Paper 1PSCB.10.

Conkie, A. & Isard, S. (1996). Optimal Coupling of Diphones. In: *Progress in Speech Synthesis*, J. van Santen *et al.* (Eds), London, UK: Springer-Verlag, 293 - 305.

Cowley, C. K. & Jones, D. M. (1993). Assessing the Quality of Synthetic Speech. In: C. Baber & J. M. Noyes (Eds.) *Interactive Speech Technology*, 149 – 156.

Crespo, M., Velasco, P., Serrano, L. & Sardina, J. (1996). On the use of a Sinusoidal Model for Speech Synthesis in Text-to-Speech. In: J. van Santen *et al.* (Eds.) *Progress in Speech Synthesis*, London, UK: Springer Verlag, 57-70.

Crystal, D. (1987). *The Cambridge Encyclopaedia of Language*. Cambridge UK: Cambridge University Press.

Deketelaere, S., Derro, O. & Dutoit, T. (2001). Speech Processing for Communications: What's New? *Revue HF*, 5-24.

Dixon, N. R. & Maxey, H. D. (1968). Terminal Analog Synthesis of Continuous Speech using the Diphone Method of Segment Assembly. *IEEE Transactions*, AU-16, 40-50.

Donovan, R. (1996). *Trainable Speech Synthesis*. PhD thesis. Cambridge University. [ftp://svr-ftp.eng.cam.ac.uk/pub/reports/donovan_thesis.ps.Z]

Donovan, R. E. & Woodland, P. C. (1999). A Hidden Markov-Model-Based Trainable Speech Synthesizer. *Computer Speech & Language*, 13, 223-241.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis.* The Netherlands: Kluwer Academic Publishers.

Dutoit, T. & Leich, H. (1993). MBR-PSOLA: Text-to-Speech Synthesis based on an MBE Resynthesis of the Segments Database. *Speech Communication*, 13, 435-440.

Dutoit, T. & Leich, H. (1994). High Quality Text-to-Speech Synthesis: A Comparison of Four Candidate Algorithms. In: *Proceedings of the IEEE International Conference of Acoustics, Speech & Signal Processing*, 565-568.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F. & van der Vrecken, O. (1996). The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. In: *Proceedings of the International Conference on Spoken Language Processing, Philadelphia*, 3, 1393-1396.

Eagles (1996). Expert Advisory Group on Language Engineering Standards (EAGLES) Guidelines. DG XIII *Linguistic Research and Engineering.*
[http://www.ilc.cnr.it/EAGLES96/intro.html]

Edgington, M. & Lowry, A. (1996). Residual-Based Speech Modification Algorithms for Text-to-Speech Synthesis. In: *Proceedings of the International Conference on Spoken Language Processing*, 3, 1425-8.

Edgington, M., Lowry, A., Jackson, P., Breen, A. P. & Minnis, S. (1996a). Overview of Current Text-to-Speech Techniques: Part 1 – Text and Linguistic Analysis. *BT Technology Journal*, 14 (1), 68 – 83.

Edgington, M., Lowry, A., Jackson, P., Breen, A. P. & Minnis, S. (1996b). Overview of Current Text-to-Speech Techniques: Part 2 – Prosody and Speech Generation. *BT Technology Journal*, 14 (1), 84 – 99.

Fant, G. (1960). *Acoustic Theory of Speech Production.* The Hague: Mouton.

Flanagan, J. L. (1972). *Speech Analysis Synthesis & Perception.* New York: Springer–Verlag.

Fujimura, O. (1962). Analysis of Nasal Consonants. *Journal of the Acoustical Society of America (JASA),* 34, 1865 – 1875.

Fujimura, O. & Lovins, J. (1978). Syllables as Concatenative Phonetic Units. *In:* A. Bell, & J. Hooper (Eds.) *Syllables and Segments.* Amsterdam, North-Holland. 107-120.

Gelfand, S. A. (1998). *Hearing: An Introduction to Psychological and Physiological Acoustics.* NY: Marcel Dekker Inc.

Gibbon, D., Moore, R. & Winski, R. (Eds.), (1997). *Handbook of Standards and Resources for Spoken Language Systems.* Berlin & New York: Walter de Gruyter Publishers.

Gray, A. H. & Markel, J. D. (1975) Distance Measures for Speech Processing. *IEEE Transactions on Acoustic, Speech, & Signal Processing,* assp-24 (5), 380–391.

Griffin, D. W. & Lim, J. S. (1984). Signal Estimation from Modified Short-Time Fourier Transform. *IEEE Transactions on Acoustics, Speech & Signal Processing,* ASSP-32, (2), 236 – 243.

Halle, M., Hughes, G. W. & Radley, J. P. A. (1957). Acoustic Properties of Stop Consonants. *Journal of the Acoustical Society of America (JASA),* 29, 107 – 116.

Hamon, C., Moulines, E. & Charpentier, F. (1989) A Diphone Synthesis System based on Time-Domain Modifications of Speech. In: *Proceedings of the International Conference on Acoustics, Speech & Signal Processing, Glasgow,* 238 – 241.

Harris, J. (1978). On the use of Windows for Harmonic Analysis with the Discrete Fourier Transform. In: *Proceedings of the IEEE,* 66 (1).

't Hart, J., Collier, R. & Cohen, A. (1990). *A Perceptual Study of Intonation.* Cambridge: Cambridge University Press.

Hawkins, S., Heid, S., House, J. & Huckvale, M. (2000). Assessment of Naturalness in the ProSynth Speech Synthesis Project. In: *Proceedings of the IEE Colloquium on Speech Synthesis, London.*

Hess, W. (1983). *Pitch Determination of Speech Signals: Algorithms and Devices.* Berlin: Springer Verlag.

Hirokawa, T. & Hakoda, K. (1990) Segment Selection and Pitch Modification for High Quality Synthesis using Waveform Segment. In: *Proceedings of the International Conference on Spoken Language Processing,* 337-340.

Holmes, J. (1983). Formant Synthesisers – Cascade or Parallel? *Speech Communication,* 2, 251-273.

Holmes, J., Mattingly, I. G. & Shearme, J. N. (1964). Speech Synthesis by Rule. *Language and Speech,* 7, 127 – 143.

Holzapfel, M. & Campbell, N. (1998). A Nonlinear Unit Selection Strategy for Concatenative Speech Synthesis based on Syllable Level Features. In: *Proceedings of the International Conference on Spoken Language Processing, Australia,* 6, 2755-2758.

Hon, H., Acero, A., Huang, X., Liu, J. & Plumpe, M. (1998). Automatic Generation of Synthesis Units for Trainable Text-to-Speech Systems. In: *Proceedings of the IEEE International Conference on Acoustics & Speech Signal Processing,* 1, 293 – 296.

House, A., Williams, C. E., Hecker, M. H. & Kryter, K. D. (1965). Articulation Testing Methods: Consonantal Differentiation with a Closed Response Set. *Journal of the Acoustical Society of America (JASA),* 37, 158-166.

Howard-Jones, P. A. & the SAM Partnership (1992a). 'SOAP' – A Speech Output Assessment Package for Controlled Multilingual Evaluation of Synthetic Speech. In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech),* (1), 281-283.

Howard-Jones, P. A. & the SAM Partnership. (1992b). Specification of Listener Dimensions. *ESPRIT Project 2589 (SAM), Multilingual Speech Input/Output Assessment Methodology and Standardisation.* University College London, London. Stage Report 8, Part One.

Howell, P. (1993). Cue Trading in the Production and Perception of Vowel Stress. *Journal of the Acoustical Society of America (JASA)*, 94, 2063-2073.

Huang, X., Acero, A., Adcock, J., Hon, H., Goldsmith, J., Liu, J. & Plumpe, M. (1996). Whistler: A Trainable Text-to-Speech System. In: *Proceedings of the International Conference on Spoken Language Processing*, 4, 2387 – 2390.

Hughes, G. W. & Halle, M. (1956) Spectral Properties of Fricative Consonants. *Journal of the Acoustical Society of America (JASA)*, 28, 303 – 310.

Hunt, A. J. & Black, A. W. (1996). Unit Selection in a Concatenative Speech Synthesis System using a Large Speech Database. In: *Proceedings of the IEEE International Conference on Acoustics & Speech Signal Processing, Germany*, 1, 373-376.

IPA (1949). *The Principles of the International Phonetic Association: A Description of the International Phonetic Alphabet and the Manner of Using It, Illustrated by Texts in 51 Languages.* International Phonetic Association, Dept. of Phonetics, University College of London.

Itoh, K., Nakajima, S. & Hirokawa, T. (1994). A New Waveform Speech Synthesis Approach based on the COC Speech Spectrum. In: *Proceedings of the IEEE International Conference on Acoustics & Speech Signal Processing, Australia*, 1, 557 – 580.

ITU-T (1994). *Recommendation P.85 – A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices.* Study Group 12, ITU-T.
[http://www.itu.int]

ITU-T (1996). *Recommendation P.800 – Methods for Subjective Determination of Transmission Quality.* Study Group 12, ITU-T. [http://www.itu.int]

Iwahashi, N., Kaiki, N. & Sagisaka, Y. (1992). Concatenative Speech Synthesis by Minimum Distortion Criteria. In: *Proceedings of the IEEE International Conference on Acoustics & Speech Signal Processing, USA*, 2, 65-68.

Iwahashi, N. & Sagisaka, Y. (1995). Speech Segment Network Approach for an Optimal Synthesis Unit Set. *Computer Speech & Language*, 9, 335-352.

Jeida (1995). JEIDA Guideline for Speech Synthesiser Evaluation. *Research Report on Office Automation Equipment Standardization*, 221-241.
[http://www.slt.atr.co.jp/cocosda/output/jeida2.txt]

Jekosch, U. (1992). The Cluster-Identification Test. In: *Proceedings of the 1992 International Conference on Spoken Language, Banff, Canada*, 1, 205 – 209.

Jekosch, U. (1993). Speech Quality Assessment and Evaluation. In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Berlin*, 1387-1394.

Jekosch, U. & Pols, L. C. W. (1994). A Feature-Profile for Application-Specific Speech Synthesis Assessment and Evaluation. In: *Proceedings of the International Conference on Spoken Language Processing*, 3, 1319 – 1322.

Jilka, M., Möhler, G. & Dogil, G. (1999). Rules for the Generation of ToBI-based American English Intonation. *Speech Communication*, 28, 83 - 108.

Kawai, H., Higuchi, N., Simizu, T. & Yamamoto, S. (1994). Development of a Text-to-Speech System for Japanese based on Waveform Slicing. In: *Proceedings of the International Conference on Acoustics, Speech & Signal Processing, Australia*, 569-572.

Kent, R. D. & Read, C. (1992). *The Acoustic Analysis of Speech*. San Diego: Singular.

Klabbers, E. & Veldhuis, R. (2001). Reducing Audible Spectral Discontinuities. *IEEE Transactions on Speech and Audio Processing,* 9 (1), 39-51.

Klatt, D.H. (1979). Synthesis by Rule of Segmental Durations in English Sentences. *In:* Lindblom & Ohman (Eds.) *Frontiers of Speech Communication Research,* London: Academic Press, 287-299.

Klatt, D. H. (1982). The Klattalk Text-to-Speech System. In: *Proceedings of the International Conference on Acoustic & Speech Signal Processing,* 1589 – 1592.

Klatt, D. H. (1987). Review of Text-to-Speech Conversion for English. *Journal of the Acoustical Society of America (JASA),* 82 (3), 737-793.

Koenig, W., Dunn, H. K. & Lacy, L. Y. (1946). The Sound Spectrograph. *Journal of the Acoustical Society of America (JASA),* 17, 19 – 49.

Kortekaas, R. W. L. & Kohlrausch, A. (1997a). Psychoacoustical Evaluation of the Pitch-Synchronous Overlap-and-Add Speech-Waveform Manipulation Technique using Single-Formant Stimuli. *Journal of the Acoustical Society of America (JASA),* 101 (4), 2202-2213.

Kortekaas, R. W. L. & Kohlrausch, A. (1997b). Psychophysical Evaluation of PSOLA: Natural versus Synthetic Speech. In: *Proceedings of the 5th European Conference on Speech Communication & Technology (Eurospeech),* 5, 2497-2490.

Kortekaas, R. W. L. & Kohlrausch, A. (1999). Psychoacoustic Evaluation of PSOLA II. Double Formant Stimuli and the Role of Vocal Peturbation. *Journal of the Acoustical Society of America (JASA),* 105 (1), 552 – 535.

Kraft, V. & Portele, T. (1995). Quality Evaluation of five German Speech Synthesis Systems. *Acta Acustica,* 3, 351 – 365.

Kröger, B. (1992). Minimal Rules for Articulatory Speech Synthesis. In: *Proceedings of EUSIPCO (European Association for Signal Processing),* 1, 331-334.

Ladefoged, P. (1996). *Elements of Acoustic Phonetics.* 2nd Ed. Chicago: University Chicago Press.

Laroche, J. (1989). A New Analysis/Synthesis System of Musical Signals using Prony's Method. Application to Heavily Damped Percussive Sounds. In: *Proceedings of the International Conference on Acoustics, Speech & Signal Processing, Glasgow*, 2053 – 2056.

Laroche, J., Stylianou, Y. & Moulines, E. (1993). HNS: Speech Modification based on a Harmonic + Noise Model. In: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2, 550-553.

Larreur, D., Emerard, F. & Marty, F. (1989). Linguistic and Prosodic Processing for a Text-to-speech Synthesis System. In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Paris*, 510-513.

Lenzo, K. & Black, A. (2000). Diphone Collection and Synthesis. In: *Proceedings of the International Conference on Spoken Language Processing, Beijing, China.*

Linggard, R. (1985). *Electronic Synthesis of Speech.* Cambridge: Cambridge University Press.

Lowry, A. (1999). Personal Communication.

Macon, M. W. & Clements, M. A. (1996). Speech Concatenation and Synthesis using an Overlap-Add Sinusoidal Model. In: *Proceedings of the International Conference on Acoustics, Speech & Signal Processing*, 1, 361 – 364.

Macon, M. W., Cronk, A. E. & Wouter, J. (1998). Generalization and Discrimination in Tree-Structured Unit Selection. In: *Proceedings of the 3rd ESCA Workshop on Speech Synthesis, Australia*, 195-200.

Markel, J.D. & Gray, A. H. Jr. (1976). *Linear Prediction of Speech*, NY: Springer Verlag.

McAulay, R. & Quatieri, T. (1986). Speech Analysis/Synthesis based on a Sinusoidal Representation. *IEEE Transactions on Audio, Speech and Signal Processing*, 34, 744-754.

Meron, Y. & Hirose, K. (1999). Efficient Weight Training for Selection Based Synthesis. In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Budapest*, 5, 2319-2322.

Möbius, B. (2000). Corpus-Based Speech Synthesis: Methods and Challenges. *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stutgart)*, AIMS 6 (4), 87-116.

Morton, K. (1991). Expectations for Assessment Techniques Applied to Speech Synthesis. In: *Proceedings of the Institute of Acoustics*, 13 (2), 501-508.

Moulines, E. & Charpentier, F. (1988) Diphone Synthesis using a Multipulse LPC Technique. In: *Proceedings of SPEECH '98, 7th FASE Symposium, Edinburgh*, 47 – 53.

Moulines, E. & Charpentier, F. (1990). Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones. *Speech Communication*, 9, 453-467.

Moulines, E., Hamon, C. & Charpentier, F. (1989). High-Quality Prosodic Modifications of Speech using Time-Domain Overlap-add Synthesis. In: *Proceedings XII ieme colloque GRETSI.*

Moulines, E., Emerard, F., Larreur, D., Le Saint Milon, J. L., Le Faucher, L., Marty, F., Charpentier, F. & Sorin, C. (1990). A Real Time French Text-to-Speech System Generating High-Quality Synthetic Speech. In: *Proceedings of the International Conference on Acoustics, Speech & Signal Processing*, 1, 309-312.

Nakajima, S. (1994) Automatic Synthesis Unit Generation for English Speech Synthesis based on Multi-Layered Context Oriented Clustering. *Speech Communication*, 14, 313-324.

Nakajima, S. & Hamada, H. (1988). Automatic Generation of Synthesis Units based on Context Oriented Clustering. In: *Proceedings of the International Conference on Acoustics, Speech & Signal Processing, New York*, 659-662.

Neovius, L. & Raghavendra, P. (1993). Comprehension of KTH Text-to-Speech with Listening Speed Program. In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, 3, 1687-1690.

Niimi, Y., Kasamatu, M., Nishimoto, T. & Araki, M. (2001). Synthesis of Emotional Speech using Prosodically Balanced VCV Segments. In: *Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Scotland.*

O'Connor, J. D., Gertman, L. J., Liberman, A. M., Delattre, P. G. & Cooper, F. S. (1957). Acoustic Cues for the Perception of Initial /w, j, r, l/ in English. *Word*, 13, 24 – 43.

O'Shaughnessy, D. (1987). *Speech Communication – Human and Machine*, Addison Wesley.

Page, J. & Breen, A. (1996). The Laureate Text-to-Speech System, Architecture and Applications. *BT Technical Journal*, 14, 57-67.

Pols, L. C. W. & SAM-partners. (1992). Multi-Lingual Synthesis Evaluation Methods. In: *Proceedings of the International Conference on Spoken Language Processing*, 2, 181-184.

Portele, T. (1998) Just Concatenation – A Corpus-Based Approach and its Limits. In: *Proceedings of the 3rd ESCA/ COCOSDA Workshop on Speech Synthesis*, 61 – 71.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1992). *Numerical Recipes in C: The Art of Scientific Computing*, 2nd Edn. UK: Cambridge University Press.

Rahim, M. G., Goodyear, C. C., Kleijn, W. B., Schroeter, J. & Sondhi, M. M. (1993). On the use of Neural Networks in Articulatory Speech Synthesis. *Journal of the Acoustical Society of America (JASA)*, 2, 1109 – 1121.

Robinson, D. W. & Dadson, R. S. (1956). A Redetermination of the Equal Loudness Relations for Pure Tones. *British Journal of Applied Physics,* 7, 166-181.

Sagisaka, Y. (1988). Speech Synthesis by Rule using Optimal Selection of Non-Uniform Synthesis Units. In: *Proceedings of the International Conference on Acoustics, Speech & Signal Processing, New York,* 679-682.

Sagisaka, Y. (1990). On the Prediction of Global F0 Shape for Japanese Text-to-Speech. In: *Proceedings of the International Conference on Acoustics, Speech & Signal Processing,* 2, 49 – 52.

Sagisaka, Y. & Iwahashi, N. (1995). Objective Optimization in Algorithms for Text-to-Speech Synthesis. In W. B. Kleijn & K. K. Paliwal (Eds), *Speech Coding and Synthesis,* Amsterdam: Elsevier, 685-706.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirchberg, J. (1992) ToBI: A Standard for Labeling English Prosody. In: *Proceedings of the International Conference on Spoken Language Processing,* 867 – 870.

Sommers, M. S. & Kewley-Port, D. (1996). Modeling Formant Frequency Discrimination of Female Vowels. *Journal of the Acoustical Society of America (JASA),* 99, 3770 – 3781.

Sonntag, G. P. & Portele, T. (1996). A Framework to Evaluate and Verify the Presence of Linguistic Concepts in the Prosody of Spoken Utterances. In: *Proceedings of the 3rd SPEAK!-Workshop, Budapest.*

Spiegel, M. F., Altom, M. J., Macchi, M. J. & Wallace, K. L. (1990). Comprehensive Assessment of the Telephone Intelligibility of Synthesised Natural Speech. *Speech Communication,* 9, 279-291.

Sproat, R. & Olive, J. (1995). An Approach to Text-to-Speech Synthesis. In W. B. Kleijn & K. K. Paliwal (Eds), *Speech Coding and Synthesis,* Amsterdam: Elsevier, 611-633.

Sproat, R., Ostendorf, M. & Hunt, A. (1999). *The Need for Increased Speech Synthesis Research.* Report of the 1998 NSF Workshop for Discussing Research Priorities & Evaluation Strategies in Speech Synthesis. [http://www.research.att.com/~rws/newindex/report10.pdf]

Stella, M. G. & Charpentier, F. J. (1985). Diphone Synthesis using Multipulse Coding and a Phase Vocoder. In: *Proceedings of the International Conference on Acoustics, Speech & Signal Processing,* 740 – 743.

Stevens, K. N. (1990). Control Parameters for Synthesis by Rule. In: *Proceedings of the ESCA Tutorial Day on Speech Synthesis, Autrans,* 27-37.

Stevens, S. S. & Volkmann, J. (1940). The Relation of Pitch to Frequency: A Revised Scale. *American Journal of Psychology,* 53, 329 – 353.

Stylianou, Y. (1998). Concatenative Speech Synthesis using a Harmonic + Noise Model. In: *Proceedings of the 3rd ESCA Speech Synthesis Workshop.*

Stylianou, Y. (2001). Applying the Harmonic plus Noise Model in Concatenative Speech Synthesis. In: *IEEE Transactions on Speech & Audio Processing.*

Stylianou, Y., Laroche, J. & Moulines, E. (1995). High Quality Speech Modification based on a Harmonic + Noise Model. In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech).*

Syrdal, A. K., Conkie, A. & Stylianou, Y. (1998a). Exploration of Acoustic Correlates in Speaker Selection for Concatenative Synthesis. In: *Proceedings of the 5th International Conference on Spoken Language Processing, Australia.*

Syrdal, A., K., Stylianou, Y., Garrison, L., Conkie, A. & Schroeter, J. (1998b). TD-PSOLA versus Harmonic plus Noise Model in Diphone Based Speech Synthesis. In: *Proceedings of the International Conference on Acoustics, Speech & Signal Processing, Seattle,* 273 – 276.

Takeda, K., Abe, K. & Sagisaka, Y. (1990). On Unit Selection Algorithms and their Evaluation in Non-Uniform Speech Synthesis. In: *Proceedings of the ESCA Workshop on Speech Synthesis, France*, 35-38.

Turner, C. W., Zwislocki, J. J. & Filion, P. R. (1989). Intensity Discrimination Determined with Two Paradigms in Normal and Hearing Impaired Subjects. *Journal of the Acoustical Society of America (JASA)*, 86, 109 – 115.

Valbret, H., Moulines, E. & Tubach, J. P. (1992). Voice Transformation using PSOLA Technique. *Speech Communication*, 11, 175-187.

van Bezooijen, R. & van Heuven, V. (1997). Assessment of Synthesis Systems. In: Gibbon *et al.* (Eds). *Handbook of Standards and Resources for Spoken Language Systems*, Berlin: Walter Gruyler & Co., 481-562.

van den Heuval, H., Cranen, B. & Rietveld, T. (1996). Speaker Variability in the Coarticulation of /a, i, u/. *Speech Communication*, 18, 113-130.

van Santen, J. P. H. (1993). Perceptual Experiments for Diagnostic Testing of Text-to-Speech Systems. *Computer Speech and Language*, 7, 49-100.

van Santen, J. P. H. (1997) Prosodic Modeling in Text-to-Speech Synthesis. In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Rhodes, Greece.*

van Santen, J. P. H. & Buschbaum, A. L. (1997). Methods for Optimal Text Selection. In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Greece*, 2, 553-556.

Veldhuis, R. (1998). A Computationally Efficient Alternative for the Liljencrants-Fant Model and its Perceptual Evaluation. *Journal of the Acoustical Society of America (JASA)*, 103 (1), 566 – 571.

Vine, D. S. G., Sahandi, R. & Longster, J. (1999). Recording Concatenative Units for Speech Synthesis using a Reference Pitch Prompt. In: *Proceedings of the International Workshop - Speech & Computer (SPECOM '99)*, 174-177.

Violaro & Boëffard (1998). A Hybrid Model for Text-to-Speech Synthesis. In: *IEEE Transactions on Acoustics, Speech & Signal Processing*, 6 (5), 426 – 434.

Voiers, W. D. (1983). Evaluating Processed Speech using the Diagnostic Rhyme Test. *Speech Technology*, Jan-Feb, 30-39.

Wang, W. J., Campbell, N. W., Iwahashi, N. & Sagisaka, Y. (1993). Tree-Based Unit Selection for English Speech Synthesis. In: *Proceedings of the International Conference on Acoustics, Speech & Signal Processing, USA*, 2, 191-194.

Wells, J., Barry, W., Grice, M., Fourcin, A. & Gibbon, D. (1992). *Standard Computer-Compatible Transcription*. ESPRIT Project 2589, SAM-UCL-037.

Wightman, C. W. & Talkin, D. T. (1996). The Aligner: Text-to-Speech Alignment using Markov Models. In: J. van Santen *et al.* (Eds.) *Progress in Speech Synthesis*, London, UK: Springer Verlag, 313 - 323.

Witten, I.H. (1982). *Principles of Computer Speech*. London: Academic Press.

Wouters, J. & Macon, M. W. (1998). A Perceptual Evaluation of Distance Measures for Concatenative Speech Synthesis. To Appear in: *Proceedings of the International Conference on Spoken Language Processing, November 98*.

Wouters, J. & Macon, M. W. (2000). Spectral Modification for Concatenative Speech Synthesis. In: *Proceedings of the International Conference on Acoustics, Speech & Signal Processing*, 941-944.

Wright, H. N. (1960). Audibility of Switching Transients. *Journal of the Acoustical Society of America (JASA)*, 32, 138.

Zera, J., Onsan, Z. A., Nguyen, Q. T. & Green, D. M. (1993). Auditory Profile Analysis of Harmonic Signals. *Journal of the Acoustical Society of America (JASA)*, 93, 3431 – 3441.