Spectral Discontinuity in Concatenative Speech Synthesis - Perception, Join Costs and Feature Transformations

Barry Kirkpatrick

A dissertation submitted in fulfilment of the requirements for the award of

Doctor of Philosophy

to the

School of Computing

Dublin City University

Supervisors: Dr. Darragh O'Brien & Dr. Ronán Scaife

2010

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: _____ ID No.: 98665197 Date: _____

Acknowledgements

During the completion of this thesis I gained help and support from many people and I would like to thank them. Firstly, many thanks to my supervisors Darragh O'Brien and Ronán Scaife for constant guidance, support, insightful discussions and constructive suggestions throughout the course of my PhD.

I would like to thank Andrew Errity for many useful discussions that proved useful in identifying constructive research and helped in the avoidance of infeasible and potentially unproductive studies. I would also like to thank everyone that contributed to the listening tests in the perceptual experiment for their time and patience.

Contents

Abstract

This thesis explores the problem of determining an objective measure to represent human perception of spectral discontinuity in concatenative speech synthesis. Such measures are used as join costs to quantify the compatibility of speech units for concatenation in unit selection synthesis. No previous study has reported a spectral measure that satisfactorily correlates with human perception of discontinuity. An analysis of the limitations of existing measures and our understanding of the human auditory system were used to guide the strategies adopted to advance a solution to this problem.

A listening experiment was conducted using a database of concatenated speech with results indicating the perceived continuity of each concatenation. The results of this experiment were used to correlate proposed measures of spectral continuity with the perceptual results. A number of standard speech parametrisations and distance measures were tested as measures of spectral continuity and analysed to identify their limitations. Time-frequency resolution was found to limit the performance of standard speech parametrisations. As a solution to this problem, measures of continuity based on the wavelet transform were proposed and tested, as wavelets offer superior time-frequency resolution to standard spectral measures. A further limitation of standard speech parametrisations is that they are typically computed from the magnitude spectrum. However, the auditory system combines information relating to the magnitude spectrum, phase spectrum and spectral dynamics. The potential of phase and spectral dynamics as measures of spectral continuity were investigated. One widely adopted approach to detecting discontinuities is to compute the Euclidean distance between feature vectors about the join in concatenated speech. The detection of an auditory event, such as the detection of a discontinuity, involves processing high up the auditory pathway in the central auditory system. The basic Euclidean distance cannot model such behaviour. A study was conducted to investigate feature transformations with sufficient processing complexity to mimic high level auditory processing. Neural networks and principal component analysis were investigated as feature transformations.

Wavelet based measures were found to outperform all measures of continuity based on standard speech parametrisations. Phase and spectral dynamics based measures were

found to correlate with human perception of discontinuity in the test database, although neither measure was found to contribute a significant increase in performance when combined with standard measures of continuity. Neural network feature transformations were found to significantly outperform all other measures tested in this study, producing correlations with perceptual results in excess of 90%.

Keywords: Speech synthesis, unit selection, join cost, concatenative synthesis, spectral discontinuity, perceptual experiment, feature extraction, wavelets, speech perception, phase spectrum, spectral dynamics, feature transformation, neural networks.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Thesis overview

Speech is the primary form of human communication and as such a demand exists for real world computer applications that can communicate with human users through speech. Most users of speech technology are highly critical of unnatural sounding synthesised speech and this has proved to be a limiting factor in the widespread adoption of speech synthesis technology. An ideal speech synthesis system should be capable of producing high quality, natural-sounding, expressive speech that is indistinguishable from speech produced by a human speaker. Current speech synthesis technology is quite far from this goal and many challenges remain.

1.1.1 Speech synthesis

A number of methods exist to produce a speech waveform in speech synthesis. Currently the prevailing methods are statistical parametric synthesis and concatenative-based synthesis (King and Karaiskos, 2009; Karaiskos et al., 2008). Statistical parametric synthesis produces the speech waveform by generating speech parameters from statistics acquired from a speech database. These parameters are employed to synthesise the speech waveform to match a desired target output utterance. In concatenative synthesis a sequence of pre-recorded speech units are connected together in sequence to produce the overall target

utterance. Statistical parametric synthesis produces speech that is smooth, consistent in quality and flexible for manipulating voice characteristics. The disadvantage is that it can sound muffled and lacks the crisp sound quality of real speech. Concatenative synthesis can vary significantly in quality depending on the compatibility of successive speech units and it is difficult to make modifications to the voice characteristics. As concatenative synthesis uses natural speech, high quality, natural-sounding speech can be produced provided the joins between concatenated units are not perceptually noticeable. The objective of producing a concatenative speech synthesis system that produces speech with a minimum - ideally zero - number of audible joins is what motivates the work carried out and described in this thesis.

1.1.2 Unit selection speech synthesis

A particular type of concatenative synthesis is unit selection speech synthesis. In unit selection synthesis each target unit of speech is selected from a set of possible candidate units. The sequence of units is selected with the aim of producing the best possible quality speech output with respect to the desired target utterance. Unit selection based concatenative synthesis is currently considered state-of-the-art in text-to-speech (TTS) synthesis (King and Karaiskos, 2009; Lu et al., 2009; Karaiskos et al., 2008; Ling et al., 2008) and has been the dominant technology since its development (Hunt and Black, 1996). Unit selection synthesis is capable of producing highly natural-sounding synthetic speech, although, as with all forms of concatenative synthesis it is also capable of producing low quality speech. When the joins in concatenated speech are inaudible the synthesised speech can sound very natural. When the joins are audible and perceived as discontinuities, the quality degrades significantly. Human listeners are very quick to reject the resulting low quality speech which requires significantly more listener effort to interpret. It is this lack of consistent quality that is currently limiting the widespread use of unit selection based TTS in many commercial applications.

The development of unit selection synthesis is due mainly to the availability of low-cost memory and high computational power rather than new developments in speech science.

This is the case for most of the developments in concatenative speech synthesis since its initial emergence as viable synthesis technique (Breen, 1994). The availability of low-cost memory enables large databases of recorded speech to be employed for unit selection with many instances of each possible unit, each with varied prosodic and spectral characteristics. High computational power enables the timely selection of units at run-time to synthesise the target utterance. A significant knowledge gap exists in that it is not fully understood what distinguishes the natural-sounding speech from the low quality speech produced from unit selection. The heart of the problem lies in understanding the conditions under which the join between two concatenated segments of speech becomes an audible discontinuity. Without a complete understanding of this problem it is not clear what is the best manner to select consecutive speech units, from the units available in a given database, to produce the optimum quality speech output.

1.1.3 Problem statement

In this thesis the problem of spectral discontinuity in concatenative speech synthesis is investigated. When two units of speech are concatenated, sometimes natural-sounding speech is produced and on other occasions the join is clearly audible and sounds discontinuous. The goal of this thesis is to investigate under what conditions the join is audible and to quantify its audibility. This will result in a better understanding of the processing of abrupt spectral transitions in the human auditory system and will also provide a metric to select the optimum units of speech in unit selection synthesis leading to more reliable high quality speech synthesis.

1.1.4 Goals of the thesis

The ultimate goal of this thesis is: ***To develop a deeper understanding of human perception of discontinuity for concatenative TTS and to apply this knowledge to the development of new measures of spectral continuity that correlate with human auditory perception.***

This is a broad ranging, abstract objective and needs to be broken down into a set of

more refined objectives that provide a more definite path towards the ultimate solution. Defining these objectives reflects the nature of the approach adopted. These objectives are as follows:

1. To analyse the limits of current measures of spectral continuity and to gain an understanding of why these measures fail to consistently correlate with human perception of discontinuity.

2. To identify the underlying auditory processes that are active in the identification of spectral discontinuities. Such knowledge could be exploited in the development of new measures of continuity.

3. To explore and identify new independent sources of discontinuity and to quantify the level of correlation between measures computed from these new independent sources with human perception of discontinuity.

4. To establish a more sophisticated framework for measuring discontinuity that enables the application of advanced pattern recognition algorithms.

1.2 Thesis organisation

The thesis is structured as follows:

- **Chapter 2** contains a broad ranging literature review. This review covers: the state-of-the-art in speech synthesis, a review of the literature investigating perceptually salient distance measures for unit selection and an overview of relevant topics in speech perception.

- **Chapter 3** describes the perceptual experiment conducted to label a database of concatenated speech identified as perceptually continuous or discontinuous by human listeners. This chapter contains the results of this experiment and details of the design of the speech database and its analysis. The procedure applied to correlate the perceptual results with the proposed distance measures is also presented.

- **Chapter 4** presents a systematic comparison of standard spectral measures. Each measure is tested for its ability to detect discontinuities in the test database. The results are presented and the measures are analysed to identify the limitations of such measures for the task of detecting spectral discontinuities in concatenated speech.

- **Chapter 5** introduces new spectral measures derived from the phase spectrum and the wavelet transform. Each of the measures are tested with respect to their ability to detect discontinuities in the test database. Each measure is analysed and compared with the standard spectral measures.

- **Chapter 6** presents an investigation of spectral dynamics as a source of discontinuity in concatenated speech. A feature extraction method for spectral dynamics is presented and applied to a number of standard feature sets. The results for detecting discontinuities with spectral dynamic measures are presented. The results are further analysed to determine if spectral dynamic mismatch represents an independent source of discontinuity.

- **Chapter 7** introduces a feature space framework for the task of detecting discontinuities. Feature space transformations are investigated to determine if they can enhance the performance of a given feature set in detecting discontinuities.

- **Chapter 8** summarises the work presented in the thesis and presents a set of conclusions. Further research is also discussed.

1.3 Contributions of the thesis

The work presented in this thesis makes a number of original contributions.

- Establishment of a database of concatenated speech with perceptual results to label each test sample in the database as continuous or discontinuous for experimentation.

- Identification of the limitations of standard feature extraction methods and conditions in which they are unsuitable for the task of detecting discontinuities.

- The development of new spectral measures that help in overcoming the limitations of standard feature extraction techniques.

- Identification of spectral dynamic mismatch as a potential source of discontinuity, developing a spectral dynamic feature extraction method and determining the degree of correlation with human perception of discontinuity.

- Identification of phase spectrum mismatch as a potential source of discontinuity, investigating suitable feature sets to represent the phase spectrum and determining the degree of correlation with human perception of discontinuity.

- The proposal of a novel feature space framework and feature transformations to detect spectral discontinuities and evaluation of the proposed system.

1.3.1 Publications

The contributions of this thesis have resulted in the following publications to date.

1. B. Kirkpatrick, D. O'Brien and R. Scaife, 'A comparison of spectral continuity measures as a join cost in concatenative speech synthesis', in Proc. of the Irish Signal and Systems Conference (ISSC), Dublin, 2006.

2. B. Kirkpatrick, D. O'Brien and R. Scaife, 'Feature extraction for spectral continuity measures in concatenative speech synthesis', in Proc. International Conference on Spoken Language Processing (ICSLP), Pittsburgh PA, USA, September, 2006.

3. B. Kirkpatrick, D. O'Brien, R. Scaife, and A. Errity, 'Spectral dynamics as a source of discontinuity in concatenative speech synthesis', in Proc. of the 15th Int. Conf. on Digital Signal Processing (DSP), Cardiff, Wales, July 2007.

4. B. Kirkpatrick, D. O'Brien, R. Scaife, and A. Errity, 'On the role of spectral dynamics in unit selection speech synthesis',in Proc. of Interspeech 2007 - Eurospeech, Antwerp, Belgium, August 2007.

5. B. Kirkpatrick, D. O'Brien and R. Scaife, 'Feature transformation applied to the detection of discontinuities in concatenated speech', in Proc. of the 6th ISCA Workshop on Speech Synthesis (SSW6), Bonn, Germany, August 2007.

Chapter 2

Literature Review

An ideal TTS system should be able to produce speech of equal quality to that of a human speaker reading out the same input text. Currently no such system exists and a human listener can easily identify artificially generated speech. To improve current TTS systems it is essential to understand the key factors that influence the quality of the speech generated and to develop strategies that address these issues. Many factors are known to influence the perceived quality of synthetic speech. In concatenative speech synthesis the audibility of joins between successive speech units is the primary limiting factor in producing consistently high quality speech.

In this chapter a literature review is presented that consists of three main components: (1) text-to-speech synthesis with emphasis on unit selection, (2) join costs and (3) the auditory system. Firstly text-to-speech synthesis technologies and techniques are reviewed including waveform generation techniques and a review of benchmark TTS systems. A literature review is presented based on previous studies investigating join costs for unit selection speech synthesis. A review is presented on selected topics in human auditory perception that is required to establish the framework and foundations for later chapters.

2.1 Speech synthesis

This section contains a review of speech synthesis systems, waveform generation techniques, a more detailed presentation of unit selection synthesis and a review of selected speech synthesis systems.

2.1.1 TTS system architecture

All TTS systems share a common architecture with two distinct components; one for natural language processing (NLP) and the other for speech generation (Huang et al., 2001; Dutoit, 2001). This is depicted in Fig. 2.1. The NLP module receives the input text and analyses it to determine the structure of the text and to determine the phonetic composition of each word. This involves a number of tasks which can be further categorised into text analysis, phonetic analysis and prosody generation.

The text analysis stage involves document structure detection, text normalisation and linguistic analysis. Phonetic analysis converts abstract lexical symbols to phonemic representation. This involves homograph disambiguation, morphological analysis and letter-to-sound conversion (Huang et al., 2001; Dutoit, 2001). The prosodic generator module typically produces the intonation pattern, durations and relative amplitudes to match the underlying message, based on the linguistic analysis produced earlier.

The output from the NLP module forms the input to the speech generation module. This module is responsible for generating a corresponding speech waveform from the data supplied by the NLP module. This produces the final output utterance corresponding with the original text input. Many techniques exist to generate the speech waveform and are discussed in the next section.

2.1.2 Waveform generation

The role of the waveform generation module is to synthesise the speech signal from the parameters provided by the NLP module. A number of different techniques exist to produce the speech waveform including formant synthesis, concatenative synthesis, statistical parametric synthesis and articulatory synthesis. All of these approaches can be categorised

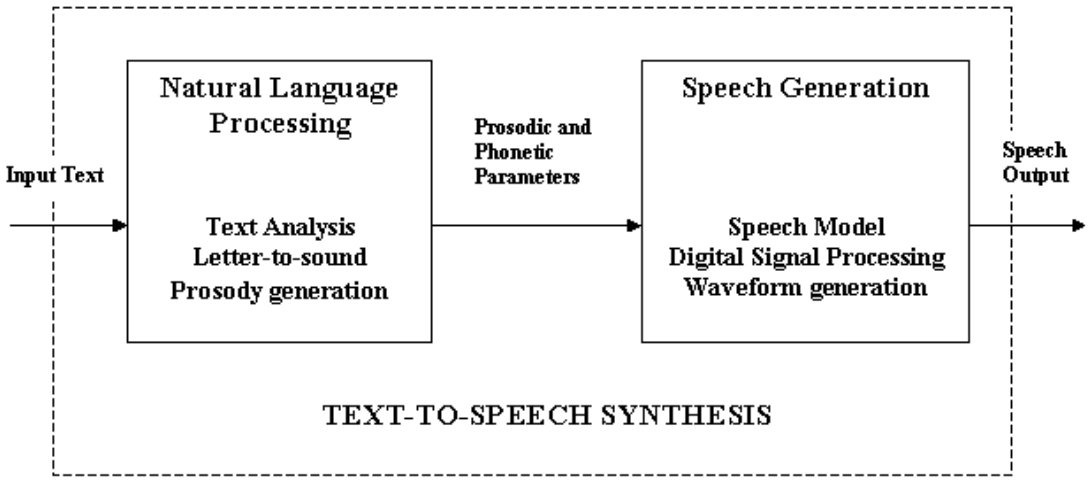

Figure 2.1: Block diagram of a generic speech synthesis system indicating the primary components; NLP and waveform generation.

as being either data-driven (corpus-based) or rule-based synthesis. In rule-based synthesis a broad set of rules are applied to produce the desired output sounds for a target utterance. In data-driven synthesis the output is built from recorded speech data.

Formant synthesis

Formant synthesis was the first widely adopted method of speech synthesis. Formant synthesis uses the source-filter model of speech production. In the source-filter model the voice source is modelled as noise for unvoiced speech and as the glottal waveform for voiced speech. Speech is produced by passing an excitation signal, representing the glottal waveform or noise, through a filter that represents the vocal tract. The parameters of the filter are 'hand derived' to correspond with known formant values for target phones. This is an example of generation by rule. With current technology the formant values and transitions could be learned from a database of speech similar to the approach of statistical parametric speech synthesis (Huang et al., 2001).

Formant synthesis by rule produces intelligible speech although it does not sound like

natural speech and it does not retain the characteristics of individual speakers. This limits its use in practice. Advantages are that it does not require a recorded database of speech and is suitable for embedded applications. Formant synthesisers are often employed in screen readers for the visually impaired, as they can produce produce intelligible speech at very high speaking rates; something that is not currently possible with concatenative synthesis.

Statistical parametric synthesis

In statistical parametric synthesis a database of speech is analysed to determine the statistical parameters of its contents (Zen et al., 2009). These statistics are used in a parameter generation algorithm that produces speech features corresponding with a target sound. These features are subsequently employed to construct the desired speech waveform. Producing the speech waveform from the parameters requires a speech model. Typically a vocoder-type model is employed. As it uses signal processing techniques to produce the speech waveform it does not sound as natural as recorded speech. The resulting speech can sound muffled or buzzy. Most current statistical parametric synthesis systems are Hidden Markov Model (HMM) synthesisers.

The key advantage of statistical parametric synthesisers over concatenative synthesisers is the consistent nature of the speech output. The quality does not vary like in concatenative synthesis, although it is generally accepted that the best quality concatenated speech is superior in quality to speech generated by a statistical parametric synthesiser. Statistical parametric synthesis offers a flexible framework for research and can be extended to incorporate ideas from other synthesis technologies; for example in Ling et al. (2009) concepts from articulatory synthesis are integrated with a HMM based synthesiser to achieve better quality speech synthesis. Statistical parametric synthesis requires a smaller database than unit selection synthesis. The memory required for the speech generation engine is small and it is suitable for embedded applications.

Articulatory synthesis

Articulatory models can be used to produce speech using a rule-based system. In human speech production the articulators in conjunction with the input airflow determine the sound produced. It has been demonstrated that articulatory models can be employed with only 15 parameters (Huang et al., 2001). Common articulatory parameters are; opening area of the lips, location and size of constriction by the tongue, opening to the nasal cavities, average glottal area and the rate of expansion/contraction of the vocal tract behind a constriction. Current articulatory synthesis techniques are not of the quality of formant, concatenative or statistical parametric synthesisers.

Concatenative synthesis

In concatenative synthesis the speech waveform is generated by concatenating pre-recorded speech units, see Fig. 2.2. Concatenative synthesis is a data-driven technique and quality depends critically on coverage in the speech database. In order for a sound to be produced it must be represented in the database. A concatenative synthesiser can produce natural sounding speech as it uses recordings of natural speech. The quality of concatenative synthesis is inconsistent and the likelihood of consistently producing output speech at the maximum possible quality is low. When two percptually incompatible units of speech are concatenated the join is clearly audible and is referred to as a discontinuity. Discontinuities are distracting for a listener and even a single discontinuity in a sentence can break the flow of the utterance and significantly reduce the overall perceived quality (Klabbers et al., 2007).

A key issue in concatenative synthesis is the choice of speech unit. For example, phonemes would appear to be a natural choice as they are the fundamental unit of all speech sounds. However concatenation of phonemes has proved to be difficult and does not result in intelligible speech due to coarticulatory effects that are difficult to model (Huang et al., 2001). A diphone is a unit of speech that contains the transition between two phones. Diphones have proved to be more successful than phones at producing quality speech output. Diphone concatenation places joins in the stationary vowel centre and

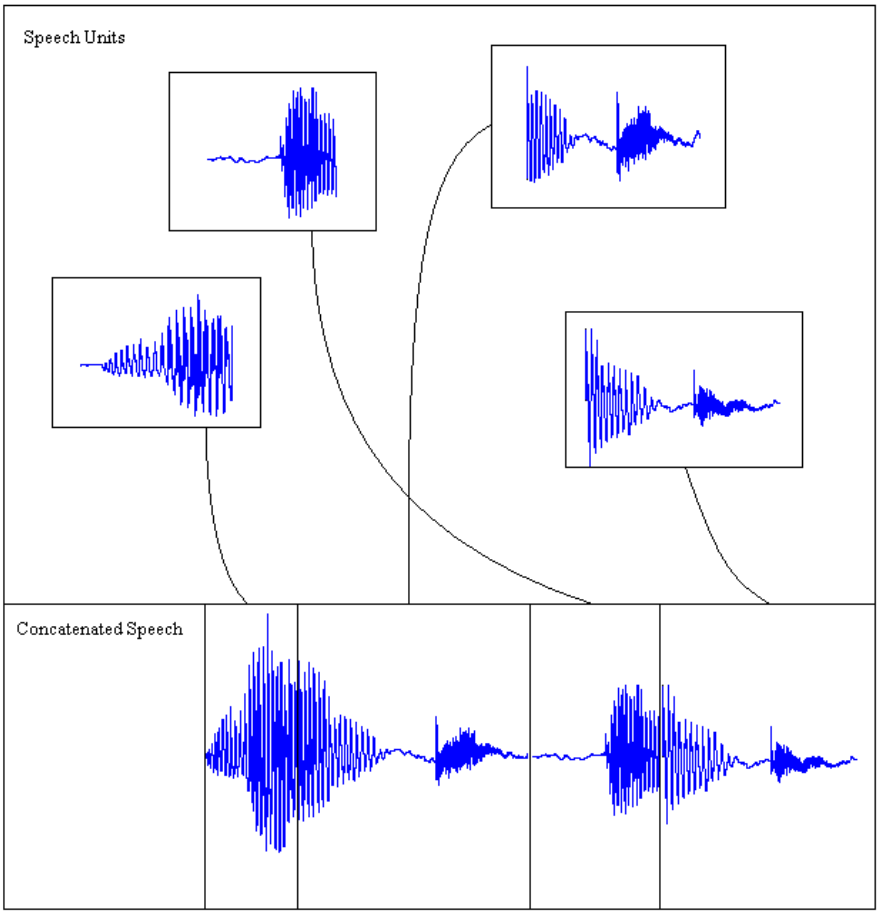

Figure 2.2: Illustrating the concatenation of speech to produce a desired output waveform.

captures rapid spectral transitions in the heart of the unit. This is advantageous as it is more likely that discontinuities will occur in a join contained in a transient. The use of diphones means that much more units need to be represented in the database. English consists of approximately 42 phonemes giving rise to approximately 1700 (42×42) possible diphones. In reality many of these phone transitions never occur in English and only 1300 diphones are required for complete coverage by a diphone database. A general trade off exists between the size of a unit and the number of units required to provide sufficient coverage to produce a waveform to represent an arbitrary text input. The general principle is that as units get shorter a more flexible approach can be adopted and smaller databases are required. However this results in many more concatenations, which inevitably leads to

more audible discontinuities and a subsequent compromise in output quality. Using longer units requires larger databases but results in fewer concatenations and consequently can be capable of producing high quality speech. Limited domain applications are suited to using longer units in which the lexicon is closed and relatively small.

2.1.3 Unit selection synthesis

Unit selection speech synthesis is a specific type of concatenative speech synthesis. In unit selection the speech database contains many instances of each possible unit, each of which has varied prosodic and spectral characteristics. This allows the selection of a unit from a set of potential candidate units, as illustrated in Fig. 2.3. Units are selected to match the target characteristics and to minimise the potential of concatenating consecutive units that will result in a discontinuity.

Two dominant strategies have emerged as methods for selecting the optimum sequence of units. Hunt and Black presented a model for unit selection in Hunt and Black (1996) and the other dominant method was presented in Donovan and Eide (1998). Hunt and Black's model is based on defining a target cost and a join cost. The target represents the degree of match between candidate units and the desired target unit and the join cost represents the degree of match between two consecutive units. The units are selected to minimise the overall cost. Donovan and Eide's method adopts a unit clustering strategy in which the units populate a decision tree. The location of the units within the decision tree depends on phonetic and prosodic characteristics with similar units being clustered together. Both methods are relatively similar and each depends on the ability to define suitable criteria to quantify the degree of compatibility between consecutive units.

Target cost

The target cost is used to represent the difference between the ideal desired unit and the units in the database. This is computed as a weighted sum of the differences from features such as; *F0, phonetic, stress, syllable position, utterance position* and *position within a word.*

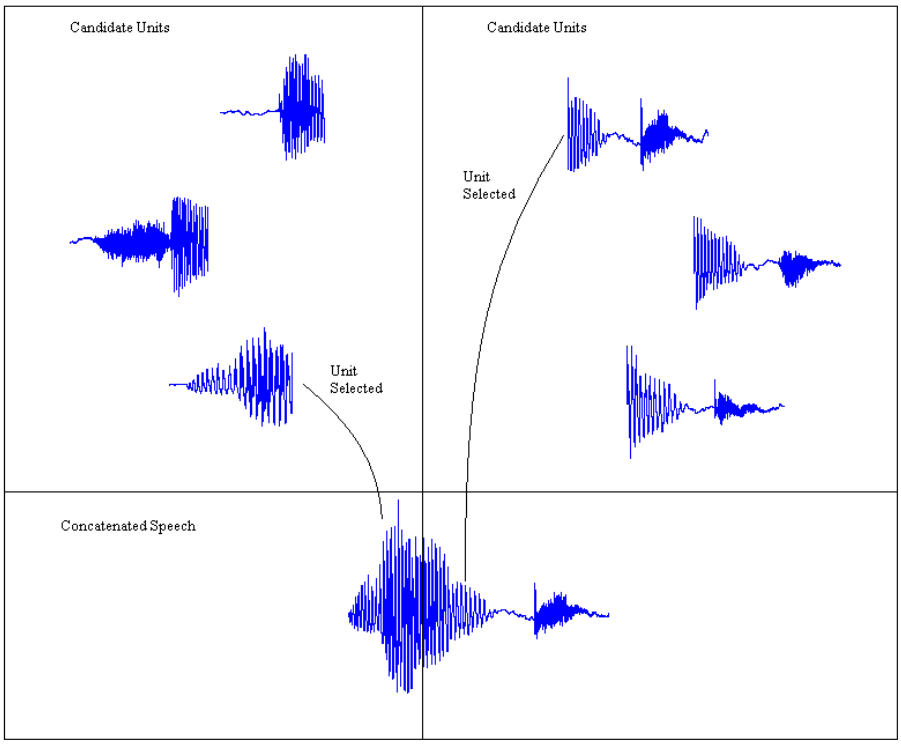

Figure 2.3: Illustrating the concatenation of two speech units were each unit is selected from a set of candidate units.

Join cost

The join cost in unit selection is used to represent the compatibility of two consecutive units. An ideal join cost should correspond with human perception of discontinuity; producing a large measure when an audible discontinuity is present and a measure of approximately zero when the join is inaudible. An ideal join cost should also be able to rank joins that are relatively more discontinuous.

The join cost is typically computed as a number of sub-costs including *F0, energy* and *spectral mismatch*. The spectral sub-cost is usually computed as the Euclidean distance between a set of Mel-Frequency Cepstral Coefficients (MFCC) from each unit. A detailed description of spectral feature extraction is presented in Chapter 4. Spectral join costs do not correlate sufficiently with human perception and have been identified as a key factor limiting the quality and consistency of speech produced with unit selection synthesisers.

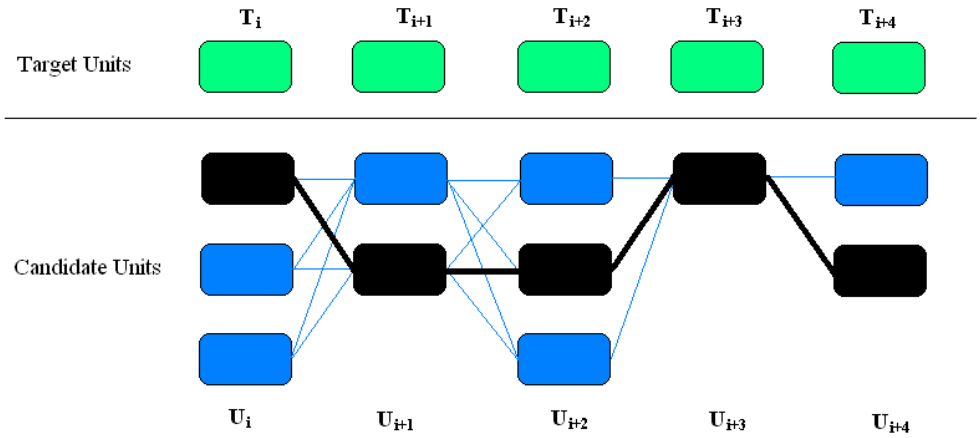

Figure 2.4: Illustrating the selection of the optimum sequence of units from the available candidate units using a Viterbi search through a state transition network using transition and target costs.

Viterbi search

In most unit selection systems the optimum sequence of units are selected by applying the Viterbi algorithm (Forney, 1973) as described in Hunt and Black (1996). A state transition network is constructed to represent the available candidate units for a target utterance. The cost of state occupancy is represented by the target cost and the cost of a transition between two states is represented by the join cost. The Viterbi algorithm computes the optimum pathway through the state transition network and as such determines the optimum sequence of units based on minimising the target and join costs. This scheme is illustrated in Fig. 2.4 in which the target units are placed in the top row and the candidate units are alligned below. A sample optimum path is illustrated.

The quality of the speech output from unit selection depends greatly on the coverage of the recordings. In general more recordings typically lead to better coverage and in turn to better synthesised speech. Practical considerations limit the size of a speech synthesis database and the amount of memory required may preclude many applications (Breen and Jackson, 1999; Breen, 2000). Recording a large database is difficult, costly and labour

intensive, with great difficulty in sustaining consistent voice quality in all recordings over significant durations of time (Stylianou, 1999; Breen, 2000).

2.1.4 Synthesis technologies

In recent years a number of benchmarks have been set in speech synthesis. Each corresponding with a particular speech synthesis system. The following section discusses a number of these synthesis systems starting with the first effective demonstration of a unit selection system, ATR v-talk (Sagisaka et al., 1992), and concludes with current HMM based statistical parametric synthesis and unit selection systems. This is not an exhaustive account of all speech synthesis systems but is intended to give an overview of the current state-of-the-art and an understanding of how the current systems have evolved.

ATR v-talk

The v-talk system was produced in ATR Japan and was the first synthesis system to provide an effective demonstration of unit selection (Sagisaka et al., 1992). Previous systems had only one instance of each possible unit available in the database whereas the v-talk system employs a database with multiple instances of each possible unit. The v-talk system automatically selects units based on minimising the spectral distance between the desired target units and the units available in the database. This system worked for Japanese and produced high quality speech. Japanese has a simpler phonetic structure than English. To produce the equivalent quality synthesis in English with a system of this type remained a significant challenge at the time.

CHATR

The CHATR speech synthesis system was also developed at ATR Japan (Black and Taylor, 1994; Campbell, 1996). It built upon the success of the v-talk system and generalised the techniques to multiple languages. CHATR, like the v-talk system, also employs unit selection synthesis although the implementation had been further developed. The units are of non-uniform size and the selection procedure uses both acoustic and prosodic features.

The system employs a state transition network to select the optimum sequence of units as proposed in (Hunt and Black, 1996) by defining appropriate target and join costs.

Festival

The Festival system was developed by Black and Taylor at CSTR in Edinburgh for synthesising speech from an inventory of diphone units (Black and Taylor, 1997). Festival was designed to be modular and flexible in order to enable research and experimentation in speech synthesis. Festival provides a means to develop and test new independent modules that can be run within the Festival system without having to develop all the other components required to achieve a functional speech synthesis system. Festival has become the standard benchmark speech synthesis system and due to its flexible nature is used in many systems that followed its release. Many further releases of Festival exist (Clark et al., 2004).

BT Laureate

The Laureate speech synthesis system was developed by British Telecom (Breen and Jackson, 1998; Page and Breen, 1998). This system generates the output speech by unit selection. It differs from other unit selection systems in that it performs unit selection based upon phonological features and does not use acoustic information. Each unit is represented by time aligned annotations, units with the same annotations are effectively equivalent in this system. Speech is generated by selecting units that optimise a global cost. A similarity metric is used for unit selection that is defined on segmental identity and supra-segmantal environment. Dynamic programming is used to determine the sequence of units.

IBM Trainable

The IBM trainable speech synthesis system was originally developed at IBM (Donovan and Eide, 1998). It uses a speaker-dependent decision-tree state-clustered hidden Markov model to automatically segment the database. To synthesise speech the system used dynamic programming for selecting from the units aligned to each leaf in the decision tree during

training. The dynamic programming used cost functions based on continuity, duration, pitch and energy. The system used TD-PSOLA (Moulines and Charpentier, 1990) to match the target prosody. The system aimed to select units such that the need for prosodic modification would be minimised.

Next-Gen

The Next-Gen synthesis system was developed at AT&T Labs (Beutnagel et al., 1999). The system built upon many of the systems that had gone before it, such as Festival, CHATR and AT&T's own Flextalk system. The Next-Gen system employed what was judged to be the best components of these preceding speech synthesis systems. In addition the Next-Gen system employed the harmonic plus noise model to generate the speech waveform (Stylianou, 2001). This system was ranked first in a study by Alvarez and Huckvale (2002) in a study of six commercially available systems and was found to produce natural sounding speech.

HTS

The HTS system was developed at Nitech Japan. It is a HMM based speech synthesis system and has been under development for many years. It has recently emerged as a viable alternative to unit selection based synthesis. In the 2005 Blizzard Challenge (Black and Tokuda, 2005) a number of the test sentences produced by the Nitech HTS system (Zen and Toda, 2005) achieved mean opinion scores above that of unit selection synthesisers (Bennett, 2005). The HTS system has continued to rank highly in subsequent Blizzard challenges. Developments employing the global variance technique in the parameter generation algorithm (Toda et al., 2005) have contributed to the improvement in quality.

2.2 Join costs for unit selection

In unit selection systems following the framework proposed by Hunt and Black (1996) a target cost and a join cost are both defined and are subsequently used to select the optimum sequence of units. The target cost is easily defined and the natural default parameters, $F0$,

phonetic, etc. employed for this task have been found to perform satisfactorily. Defining an effective join cost has proved to be much more challenging, in particular the spectral sub-cost has proven to be a limiting factor to date. The energy and $F0$ sub-costs are relatively well understood and represent perceptually important information regarding continuity at a join and each can be succinctly represented in a single parameter. Spectral continuity is not well understood and studies to date have demonstrated that current measures do not exhibit a significant correlation with human perception.

2.2.1 Previous studies investigating join costs

This section contains a literature review of studies comparing join costs for unit selection speech synthesis. A summary of these studies is contained in Table 2.1.

Wouters and Macon

Wouters and Macon (Wouters and Macon, 1998) conducted a study to evaluate the performance of a set of candidate join costs. In this study they employed a database of 166 test words where each word contained a join in a vowel centre. Four different vowels were considered. Listeners were asked to rate the level of discontinuity on a 5 point scale. The perceptual results were correlated with the join costs to quantify the agreement between the human results and each of the join cost measures. In general the correlation was found to be low. The best ranking measures were Mel-cepstral based measures with a correlation of 0.64. The performance of these measures marginally improved with the addition of delta features to a correlation of 0.66. Wouters and Macon also reported that using only delta features as a join cost resulted in a correlation with human perception similar to that of standard features.

Klabbers and Veldhuis

In the studies conducted by Klabbers and Veldhuis (1998, 2001) a number of spectral measures were compared for the task of detecting discontinuities. An experiment was conducted in which human listeners were asked if they could detect a discontinuity in a short

section of synthesised speech. The test database contained 2415 samples consisting of five Dutch vowels, each of which contained a join in the vowel centre. Klabbers and Veldhuis (Klabbers and Veldhuis, 2001) found the Kullback-Leibler distance between LPC power spectra to be the best predictor of discontinuities. They also reported that MFCC-based measures had a very low level of correlation with human perceptual results and for the task of predicting audible discontinuities, were only marginally better than random guessing. Overall the degree of correlation between human perceptual results and the join costs was not high. In a further study the contribution of various sources of discontinuity were investigated (Klabbers et al., 2007). In Klabbers et al. (2007) the novel concept of employing formant re-synthesis to manipulate individual sources of discontinuity is presented. The results of this study indicate that formant frequency mismatch is the most significant source of discontinuity of the sources tested, followed by mismatch in energy. Formant bandwidths were found to be the least significant source of discontinuity.

Vepa and King

In a series of studies conducted by Vepa and King (Vepa et al., 2002; Vepa and King, 2003, 2004a, 2006), a number of join costs were investigated and a new spectral measure tailored as a join cost for spectral continuity was proposed (Vepa and King, 2003). They also proposed the use of Mulitple Centroid Analysis (MCA) (Crowe and Jack, 1987) as an alternative to formant estimates. The new measure was based on the Kalman filter and is discussed further in the section 2.2.2. In each of these studies the test database of speech stimuli consisted of sentences of concatenated speech. The joins under examination were contained in American English diphthongs in a word that was stressed in the sentence. Listeners were asked to rate the degree of perceived discontinuity on a scale from 1 to 5. The listener results were correlated with the join costs. In Vepa et al. (2002) it is reported that the inclusion of delta features did not significantly improve performance and in some cases caused the performance to decrease. In Vepa and King (2006) it is reported that the LSF-based measure is most consistent with human perceptual results, although the preference of LSFs over other measures is not consistently statistically significant.

Stylianou and Syrdal

Stylianou and Syrdal conducted a study in which a broad range of standard speech parametrisations were tested for the task of detecting discontinuities in concatenated speech (Stylianou and Syrdal, 2001). The test database consisted of 2016 concatenated words and each word contained a single join in the vowel centre. Listeners were asked to state if they could detect a discontinuity in the test words. The perceptual results were correlated with the candidate join costs. No measure was found to correlate satisfactorily with human perception of discontinuity. The Kullback-Leibler distance between FFT-based power spectra was found to be the most successful measure, closely followed by an MFCC-based measure. In general non-parametric measures were found to outperform parametric methods such as LPC and PLP. LSFs were found to be the worst performing measure.

Chen and Campbell

Chen and Campbell conducted a study in which synthesised speech samples were evaluated by listeners (Chen and Campbell, 1999). Each speech sample was scored by the listeners on a scale of 1 to 5 and compared with the join costs. A join cost based on the Bispectrum (Mendel, 1991) was found to have the best correlation with human perceptual results. It is conjectured by Chen and Campbell that phase continuity is important and that the Bispectrum is more successful as a join cost as it encapsulates certain phase information.

Donovan

A study conducted by Donovan (2001) found that existing measures did not correlate satisfactorily with human perception of join quality. Donovan proposed a novel join cost that modified standard MFCC features. The MFCCs were modified to remove any contribution to the join cost computed from inaudible components of the signal. During the computation of the MFCCs a threshold was introduced to set to zero coefficients judged to be below the threshold of hearing. The new measure was found to correlate better with human perception of discontinuity than any of the standard measures tested in this study. The Kullback-Leibler was the second best measure reported. This study also reported an

increase in performance with the inclusion of delta features and double delta features.

Bjørkan et al.

Bjørkan et al. (2005) compared a number of distance measures for the of task detecting discontinuities in two Norwegian long vowels. Most of the measures tested produced very similar results. A weighted linear combination of features was found to produce the best results. $F0$ was found to be the best detector of discontinuities and Bjørkan et al. argue that results being corrupted by $F0$ discontinuities could be a source of inconsistency in other studies.

Additional studies

A number of additional studies have been conducted that have proposed a new measure as opposed to systematic comparison of standard measures. In Tsuzaki and Kawai (2002) the Auditory Image Model (AIM) is investigated as a join cost for concatenative speech synthesis. AIM is a computational model of the peripheral auditory system (Patterson et al., 1995). The results indicated an advantage over LPC-based measures, but no significant advantage over MFCCs. In Pantazis et al. (2005) a distance measure is proposed based on a non-linear harmonic model and amplitude modulation (AM) and frequency modulation (FM) features. Pantazis et al. report an improvement of 90% in the detection of discontinuities over the Kullback-Leibler distance between power spectra, which was found to be the best standard measure for this databse. It should be noted that a pattern recognition procedure was employed that used the same database for testing and training. In Pantazis and Stylianou (2007) the same feature sets are presented and the database is split with 80% for training and 20% for testing. The reported gains in performance are significantly less. This study does suggest that potentially significant gains can be achieved by employing advanced pattern recognition techniques to detect discontinuities.

Summary

Many studies have presented conflicting results with measures that ranked highly in one study performing poorly in another. For example, Stylianou and Syrdal found the Euclidean distance between LSFs to be the worst predictor of audible discontinuity, a finding that is inconsistent with that of Vepa and King. Similarly, Klabbers and Veldhuis reported that the Euclidean distance between MFCCs ranked poorly as a predictor of audible discontinuities, although this measure ranked highly in many other studies.

It is difficult to make direct comparisons between studies as many used different strategies to test the distance measures. Furthermore, each of the studies used different databases and different criteria to rank candidate join costs. A summary of the features tested and the best and worst performing features for each study are presented in Table 2.1.

2.2.2 Purpose-designed join costs

To date most join costs investigated have used standard speech parametrisations. Donovan (2001) modified standard MFCCs to better suit the task of a spectral join cost although little work has been done to develop measures specifically tailored for the problem of representing spectral continuity. Two join costs that have been proposed and designed specifically for the function of representing spectral continuity in unit selection synthesis are that of Vepa and King (2003) and Bellegarda (2004).

Kalman filter based join cost

Vepa and King (2003) proposed a measure in which the Kalman filter is used to model LSF trajectories. The Kalman filter is trained on natural speech to model the dynamic behaviour of LSFs for each phone. Once the Kalman filter parameters are determined for a phone it can be applied across joins contained within that phone. The prediction model inherent in the Kalman filter is thought to reflect the underlying articulatory movement. The trajectory created by the prediction model is used to compare against the actual trajectories across the join. The proposed join cost is defined from the error between the trajectories produced by the model and the actual trajectories observed across the join.

Study	Features Tested	Best	Worst
Klabbers and Veldhuis (1998)	Formants, Power spectra, MFCC, Likelihood ratio, log-spectrum, Partial Loudness	LPC power spectra (with KL)	MFCC
Wouters and Macon (1998)	Cepstra,LSF, Log Area Ratios, Itakura distance, PLP and Delta features	Mel-cepstra + deltas	Log Area Ratios
Chen and Campbell (1999)	LPC, LSF, MFCC, Bispectrum, Mellin Transform, Wigner Ville Distribution, Residual MFCC	Bispectrum	Wigner-Ville Distribution
Stylianou and Syrdal (2001)	FFT, Log FFT, PLP, LSF, Cepstra	FFT (with KL)	LSF
Donovan (2001)	Power Spectra, Cepstra, Perceptually Modified Cepstra, Log FFT, Itakura-Saito, Deltas, Double Deltas	Perceptually Modified Cepstra	Log FFT
Vepa et al. (2002)	LSF, Kalman Join Cost, MCA	LSF	Kalman Join Cost
Tsuzaki and Kawai (2002)	AIM, LPC, MFCC	AIM	LPC
Bjørkan et al. (2005)	Power spectra, LPC cepstra, MFCC, Pitch-synchronous cross correlation	Weighted linear combination of features	Pitch-synchronous cross correlation

Table 2.1: Perceptual studies to evaluate join costs, summarising the features tested and the best and worst features for each study.

When the Kalman filter is applied across a phone containing a join, the model trajectories closely match the measured trajectories until a join is encountered. The magnitude of the mismatch between the model trajectory and the observed trajectory reflect the lack of continuity of the underlying articulator dynamics in the speech production system. This measure was found to correlate with human perceptual results although it did not provide a better correlation then standard spectral representations such as LSFs (Vepa and King, 2006).

Bellegarda's modal decomposition

Bellegarda (2004) developed a join cost inspired by latent semantic analysis. The analysis procedure operates directly on the speech database to produce a transform that is tailored to represent continuity at the join. The central concept is to develop a transform matrix from the speech data, using Singular Value Decomposition (SVD) analysis (Golub and Van Loan, 1996), that can be used as an alternative to Fourier analysis. Bellegarda refers to this as a 'Boundary Centric' approach. The following steps outline the key components of this method:

- A matrix is constructed consisting of single pitch periods of speech. Each of these frames straddle a unit edge.
- SVD analysis is performed on the matrix of speech frames.
- The left modal matrix resulting from SVD analysis is adopted as the transform matrix.
- Each speech frame representing a unit is transformed by the left modal matrix resulting from SVD analysis.
- A distance measure is computed from the transformed features using the *Cos* distance (Moon and Stirling, 2000).

A diagram illustrating the modal decomposition is illustrated in Fig. 2.5. In this approach the basis functions of the transform derived from SVD analysis are directly de-

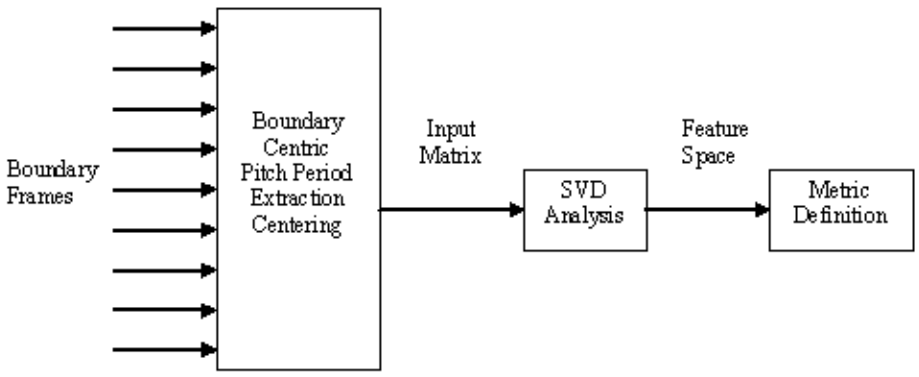

Figure 2.5: Block diagram illustrating Bellegarda's modal decomposition-based join cost.

pendent on the speech units at unit boundaries. Bellegarda argues that such a tailored transform is inherently more suited to the task then general purpose Fourier analysis. It is also argued that as the transform is a real transform, it does not separate into phase and magnitude components, like Fourier analysis. As a result the phase information is not discarded; as with conventional spectral analysis of speech. Bellegarda has reported success with this method in unit selection synthesis (Bellegarda, 2006, 2004) and unit boundary training (Bellegarda, 2007).

2.2.3 Spectral smoothing

A number of studies have been conducted to develop algorithms that smooth the spectra across the join between concatenated units (Dutoit, 2001; Chappell and Hansen, 2002; Vepa and King, 2006). These algorithms are referred to as spectral smoothing techniques. The objective is to smooth the parameters across the join to remove discontinuities in the concatenated speech. The problem of defining a perceptually salient join cost is closely related to the problem of developing an effective spectral smoothing technique. To define an effective join cost requires knowing what features of the signal cause discontinuities. To perform spectral smoothing the discontinuous features are interpolated across the join

to smooth the transition with the aim of removing the discontinuity. Without a complete understanding of what gives rise to the discontinuity it is difficult to determine what features of the signal should be modified.

LSFs have good interpolation properties and Dutoit (2001) demonstrated that they are a suitable choice for producing smoother spectral transitions between units. In Chappell and Hansen (2002) a number of different spectral smoothing algorithms are compared. Optimal coupling (Conkie and Isard, 1997) was found to be the most successful of the techniques considered. In optimal coupling the boundaries of the units are varied such that the unit edges can be selected to minimise any mismatch between the units. This technique is critically dependent on the choice of join cost used to determine the optimum location for the join. It was concluded that no algorithm tested was consistently successful as a spectral smoothing technique to remove discontinuities. In some cases the smoothing algorithm decreased the audibility of a discontinuity and in other cases caused the quality to degrade further. It was also reported that a suitable measure of discontinuity is required to direct the smoothing algorithm. In Vepa and King (2006) a spectral smoothing technique based on the Kalman filter is reported. This technique is compared with linear interpolation of LSFs. Interpolation of LSFs produced the best results and in a number of instances no smoothing was preferred to either smoothing technique.

No study to date has reported a spectral smoothing algorithm that is satisfactory. This problem depends critically on understanding spectral discontinuity and defining a suitable join cost to represent spectral discontinuity.

2.3 Speech perception

A mutually informing relationship has existed between developments in speech perception and speech synthesis since the success of early formant synthesisers. Speech synthesis provides a framework to create speech stimuli to identify specific perceptual phenomena. For example Gobl and NíChasaide (2003) used a formant synthesiser to investigate the acoustic correlates of emotionally affective speech. Klabbers et al. (2007) used a similar

framework to determine the relative importance of independent sources of discontinuity. Understanding the underlying perceptual processes at work when humans detect discontinuities in concatenative speech is another example of the relationship between synthesis and perception. In this section speech perception and the human auditory system are reviewed and discussed in relation to human perception of discontinuities in concatenated speech.

To date, strategies for investigating join costs have predominantly focused on standard speech representations. These parameterisations have largely been developed for applications like automatic speech recognition (ASR) and speech coding. The results of these studies have been disappointing suggesting that these techniques are fundamentally limited. The objective here is to identify auditory mechanisms that may be potentially relevant to the detection of discontinuities. This knowledge can provide a deeper understanding of the problem and guide the development of new spectral join costs. In the following sections key features of the peripheral and central auditory systems are presented and processes specifically related to dynamic processing and event detection are highlighted.

2.3.1 The peripheral auditory system

When a human listener detects a sound it is first processed by the peripheral auditory system. The peripheral auditory system consists of the outer, middle and inner ear and is connected to the central auditory system by the auditory nerve. The primary components of the peripheral auditory system are depicted in Fig. 2.6. The objective of the peripheral auditory system is to transform the incoming acoustic signal into a form suitable for processing in the central auditory system. When speech is sensed by a human, the peripheral auditory system encodes the signal from an acoustic signal into a neural firing pattern in the auditory nerve. The encoding process facilitates subsequent higher level decision making processes in the central auditory system and retains critical information for speech perception. A number of key stages in the peripheral auditory system are of interest with respect to gaining deeper insight into human perception of discontinuities.

- Pre-cochlear processing: The stages in the peripheral auditory system that precede

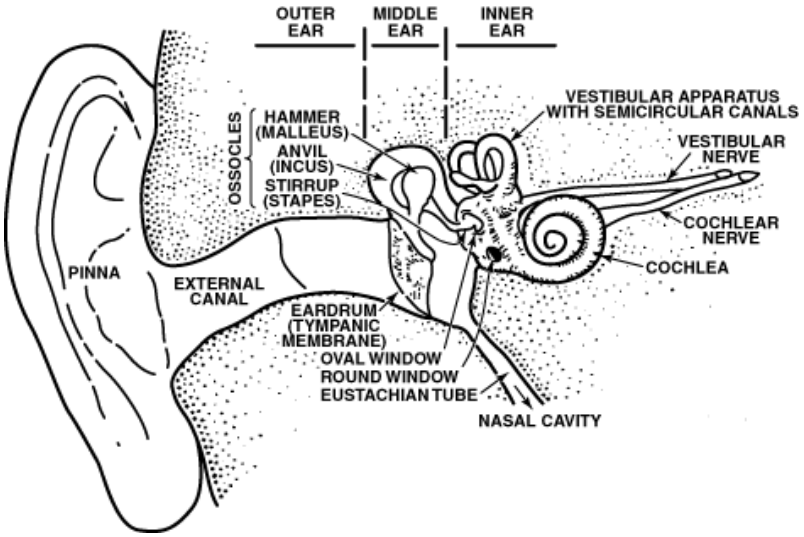

Figure 2.6: Diagram of the human ear, after Flanagan (1972).

the cochlea are typically modelled as a single functional unit. This system of physiological elements is effectively modelled from a signal processing perspective by a single filter (Patterson et al., 1995).

- Frequency selectivity: The basilar membrane is responsible for separating an incoming signal into a set of distinct overlapping frequency bands (Flanagan, 1960, 1972). Different locations along the basilar membrane respond to input stimuli of distinct frequencies (von Békésy, 1960) creating a frequency-place transformation. Each location behaves similar to a bandpass filter. The frequency decomposition of the peripheral auditory system is often modelled in signal processing applications as a filterbank of bandpass filters (Meddis, 1999; Lyon and Mead, 1988). Each filter in the filterbank represents an auditory channel. The bandwidth of each filter increases with frequency, modelling the increased frequency discrimination at low frequencies that decreases at higher frequencies. Most filterbank models employ linear filters, but this is a simplifying approximation as the non-linear nature of the basilar membrane transformation is not completely understood and is difficult to model. The output of each frequency band is characterised by its instantaneous frequency, phase and amplitude envelope, each of which excite particular neural firing patterns for higher level auditory processing.

- Neural encoding: After the incoming acoustic signal is decomposed into its constituent frequency components, hair cells located on the basilar membrane are responsible for transforming mechanical vibrations into electrical signals that are transmitted to the auditory nerve (Delgutte, 2002). Hair cells are categorised as inner and outer hair cells depending on their location, see Fig. 2.7. Inner and outer hair cells perform different tasks. Inner hair cells act as sound transducers and generate the neural signal causing neural spikes to be generated in the auditory nerve fibre. Outer hair cells play an accompanying role to facilitate the inner hair cell in reliably detecting the incoming signal and can be thought of as an amplifier (Ashmore and Gale, 2004). Movement of the basilar membrane causes movement in the inner and outer hair cells that in turn triggers neurotransmitters which create an electrical impulse in the auditory nerve. A number of significant alterations occur in this transformation. The inner hair cell has characteristics that are critical in understanding auditory processing of non-stationary acoustic signals. The hair cell is known to exhibit adaptation behaviour 'which correlates to the way attention is directed in an auditory scene' (Purwins et al., 2000). Computational models of the hair cell have been developed by Meddis (Meddis et al., 1990). The auditory image model (AIM) (Patterson et al., 1995) employs the Meddis hair cell model.

- Adaptation: The auditory system is known to display specific behaviour to deal with transient signals. When a new stimulus is detected an auditory nerve will fire at a high rate for approximately 5 to 15 ms (Smith, 1977) and this rate decreases steadily over the next 100 ms. The decay in the discharge rate is referred to as adaptation. Adaptation can occur on different time scales ranging from a few ms to several seconds and is often categorised into short-term and long-term adaptation. Short-term adaptation is assumed to be more relevant in the detection of spectral discontinuity. Adaptation enhances spectral contrast in a signal (Delgutte, 2002) that contains spectral dynamics. This has been demonstrated in numerous psychoacoustic experiments (Summerfield et al., 1986). Hair cells exhibit adaptation.

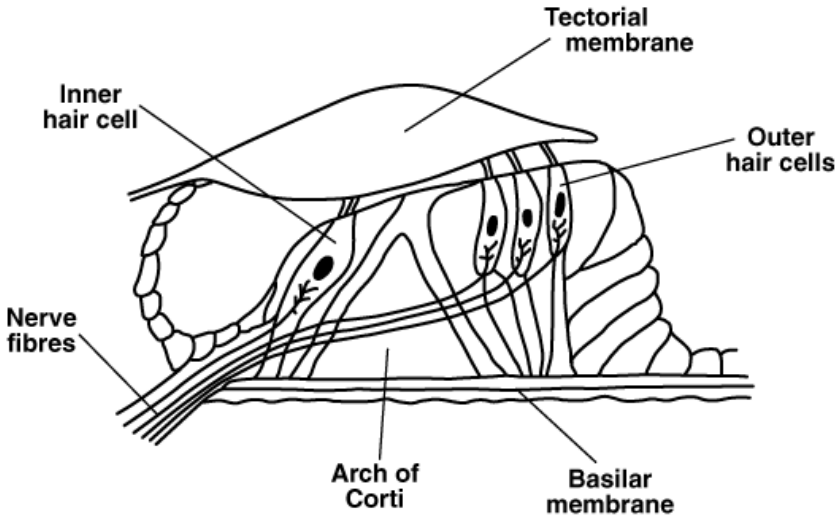

Figure 2.7: Diagram of the basialr membrane, hair cells and auditory nerve, after Flanagan (1972).

The average discharge rate of impulses in the auditory nerve and the temporal discharge patterns have been demonstrated to convey information relating to formant structure (Young, 2007). For a given auditory channel the discharge pattern contains information relating to the magnitude of the stimulus, the temporal envelope, the frequency and transient behaviour in the channel. The auditory nerve transmits this information to the central auditory system for further high level processing.

2.3.2 The central auditory system

The central auditory system is supplied with information from the peripheral auditory system by the auditory nerve. The peripheral auditory system has transformed the signal and decomposed it in terms of its frequency content, amplified features of interest and discarded certain information. This information is presented to the central auditory system in which a large number of high level tasks critical to sound perception are performed (Bregman, 1990; Moore, 2004), for example source separation, sound interpretation etc. The central auditory system has many processing units including; the cochlear nucleus, inferior colliculus and the primary auditory cortex, each of which perform different tasks. The constituent blocks of the peripheral auditory system are connected in series. In contrast,

the central auditory system has a much more complicated topology with many parallel interconnections and significant coupling between different locations.

When firing patterns produced from acoustic transients in the peripheral auditory system pass up the auditory nerve to the central auditory system many complex processes act on the signal. A number of these processes are specifically related to the detection and enhancement of spectral changes in the signal (Quatieri, 2002; Yost, 2007). Such processes are likely to play a significant role in the detection of discontinuities. For example, onset cells occur in the cochlear nucleus and in most higher level locations in the auditory system. These cells process transients and fire when an onset is detected (Quatieri, 2002). It has been demonstrated that approximately one third of the neurons in the auditory cortex will only fire when the auditory system is stimulated by a transient acoustic signal (Yost, 2007).

Lateral inhibition occurs in the primary auditory cortex and is associated with processing dynamic features in the auditory signal (Pickles, 1988; Yost, 2007). Lateral inhibition enhances changes in a signal along the frequency axis when neighbouring neurons are coupled such that one neuron can inhibit the firing rate of the other when a significant difference occurs in their inputs. Lateral inhibition occurs in the antero-ventral cochlear nucleus (Quatieri, 2002). Results have been reported that suggest lateral inhibition is active in the detection of sudden changes in frequency (May et al., 1999).

Studies into the how the brain processes sudden changes in auditory stimulus have gained momentum over the last decade as a result of non-invasive electroencephalogram (EEG) and magnetoencephalogram (MEG) measurements. The phenomenon of mismatch negativity (MMN) has received significant attention and is used as a means to investigate how the auditory system responds to changes in the auditory environment (Naatanen, 1995). MMN is the term used to describe the electrical response in the auditory system to abrupt changes in a stimulus and is characterised by negative electrical response in the brain due to any discriminable change detected in a repetitive sound. In Kraus et al. (1992) it is stated that 'The MMN appears to be an extremely sensitive electrophysiologic index of minimal acoustic differences in speech stimuli'. In May et al. (1999) it is argued that

MMN is due to both adaptation and lateral inhibition. MMN gives an indication of the nature and complexity of the processes involved in processing sudden spectral changes in the auditory system.

Much is unknown about the processing in the central auditory system. The level of processing is complex and deviates significantly from the strategies adopted in conventional signal processing feature extraction algorithms. Much of the processing is by neural circuits with a complex network of coupled interconnections and exhibits highly non-linear behaviour. Conventional speech feature extraction paradigms have evolved from linear system theory and do not relate well to processing in the central auditory system, although they do relate to certain aspects of processing in the peripheral auditory system.

2.3.3 Auditory processing and discontinuity detection

When a spectral discontinuity is detected in concatenated speech it is due to the incompatibility of the spectra on either side of the join. Signal components can change in a number of ways at the join; an abrupt termination of signal components, an abrupt onset of signal components and more subtle changes in signal components sustained across the join. In this thesis the detection of a discontinuity is interpreted as the human auditory system identifying a new auditory event. This could occur from an abrupt change in the amplitude, phase or instantaneous frequency in an auditory channel. The central auditory system will then process the change and identify if it is compatible with the existing auditory stream or if it signifies an independent event (Bregman, 1990).

Other approaches could be adopted to relate existing perceptual models to the detection of discontinuities. For example, auditory masking models have been found to be highly successful in speech coding and could be developed to relate to the detection of discontinuities. In auditory masking a dominant signal component can mask another weaker component if the components are close in either time or frequency. The auditory processes that give rise to masking, and in particular temporal masking, may provide further understanding of discontinuity detection in the auditory system. No existing models of auditory masking (Yost, 2007; Strope and Alwan, 1997; Jesteadt et al., 1982) where found to be

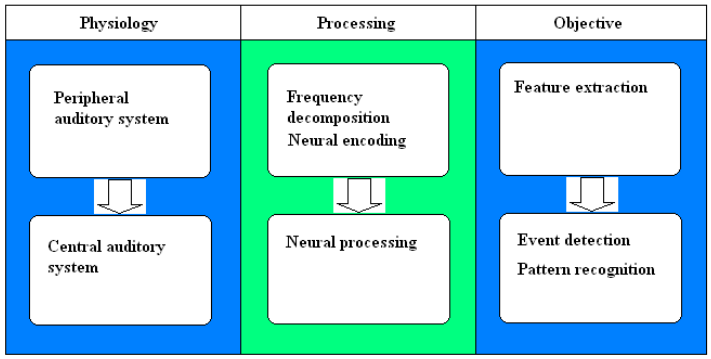

Figure 2.8: Diagram illustrating the peripheral and central auditory system, the nature of the processing and the role fulfilled by each.

readily applicable as a means to relate to the detection of discontinuities and it was decided to pursue event detection as a model to guide the development of new measures.

The detection of an auditory event invoked by an abrupt change in spectra involves high level neural processing. Processing at this level is not completely understood and is considerably more involved than conventional speech feature extraction techniques. The model adopted in this thesis is to interpret the peripheral auditory system as a feature extraction phase and the central auditory system as an advanced pattern recognition system for detecting abrupt spectral changes. This is illustrated in Fig. 2.8. This model provides a systematic way to break down the task of developing a system to detect discontinuities into two distinct phases, each of which has a parallel in the auditory system: a feature extraction module that corresponds with the peripheral auditory system and a pattern recognition module that corresponds with the central auditory system. Knowledge of the neural encoding process in the peripheral auditory system can serve as a guideline in deciding what information should be retained by the feature extraction algorithm. The complexity of the processes in the central auditory system suggests that a sophisticated algorithm is required in order to successfully detect spectral discontinuities using the information provided by the feature extraction component.

2.4 Summary and conclusion

In this chapter the literature in three key areas is reviewed: speech synthesis technology, perception of join costs and human auditory perception. A number of speech synthesis technologies were reviewed, with a particular emphasis on unit selection concatenative synthesis. The problem of defining a suitable join cost for unit selection was discussed. Previous studies investigating perceptually salient join costs for unit selection synthesis were reviewed. No obvious conclusion can be drawn from these studies as the results are largely inconsistent. The only consistent aspect of these studies is that no join cost reported has a satisfactory degree of correlation with human perception of discontinuity in concatenated speech. An overview of the human auditory system was presented. Features of the auditory system relevant to the detection of discontinuities were highlighted, including; frequency decomposition, adaptation, lateral inhibition and other complex and less well-understood processes in the central auditory system.

To develop a system that agrees with human perception of discontinuity will require an approach that extracts the appropriate features from the speech signal and performs pattern recognition at a sufficiently sophisticated level to detect discontinuities. Feature extraction should mimic key aspects of the peripheral auditory system to ensure that perceptually meaningful information is employed in the join cost. A framework for further processing of the features to discriminate between continuous and discontinuous joins should have sophistication and complexity similar to the processes that are activated in the central auditory system for the detection of sudden spectral transitions, in order to have scope to potentially model this behaviour.

Current join costs fail to produce a satisfactory level of correlation with human perception, this is well documented in the studies to date. A mismatch exists between conventional speech feature extraction algorithms and the processes in the auditory system to process transients. Considering this a number of possibilities exist to explain why current join costs under perform.

- The information extracted is insufficient to represent spectral continuity at the join.

This would indicate that alternative features need to be extracted. Almost all features tested to date are derived from the magnitude spectrum. Perceptual cues encoded in the neural firings however also relate to spectral dynamics and the phase spectrum (Patterson, 1987; Furui, 1986a; Yost, 2007).

- The computation of the Euclidean distance between features is inconsistent with the processes that detect sudden spectral change in the auditory system. The Euclidean distance or a similar metric is employed in most existing systems. Complex neural processes are used to identify new events in the central auditory system. The Euclidean distance in no way models this behaviour and could not be expected to produce similar results.

Considerable scope exists to improve join costs both in terms of feature extraction and pattern recognition. The feature extraction should ideally represent the information encoded by the peripheral auditory system that is passed on to the auditory nerve. The pattern recognition component should ideally have sufficient scope to model the complex neural circuitry involved in processing transients and detecting events in the central auditory system.

Chapter 3

Perceptual Experiment

In order to investigate human perception of discontinuity a database of concatenated speech was created. A perceptual experiment was conducted using the database. The objective was for human listeners to label each concatenation continuous or discontinuous. This enabled the creation of a framework to correlate human perception of discontinuity with objective measures and to thereby quantify the suitability of each candidate join cost. This chapter contains an overview of the issues involved in designing a perceptual experiment to evaluate join costs. The details regarding the construction of the test corpus and perceptual listening test are detailed and the subsequent results are presented and analysed. The analysis procedure to relate human perceptual results with the candidate join costs is presented. This procedure is also employed in this chapter to analyse the test database for known sources of discontinuity such as $F0$ and energy mismatch.

3.1 Overview and literature review

A number of issues arise in designing a perceptual experiment to evaluate spectral join costs.

- How to construct a database of test stimuli that the subjects are to judge based on the perceived level of discontinuity.
- The structure of the test utterances, e.g. words, phrases or sentences.

- Isolating spectral mismatch as the only source of discontinuity.
- The challenge of obtaining consistent, reliable perceptual results.
- Relating human perceptual results with the proposed candidate join costs.

3.1.1 Stimuli construction

Two dominant approaches exist for constructing a database of test stimuli for evaluating join costs. In the first approach a particular join cost is employed in a synthesis system to generate a target test utterance. This test utterance is then evaluated in terms of quality by human listeners and compared with a baseline system (Chen and Campbell, 1999). The primary limitation of this method is that separate perceptual listening tests are required to evaluate the performance of each candidate join cost. Other factors that influence this approach are the size of the database and the variability of both prosodic and spectral parameters. For example, the spectral match may improve but it may result in a degradation in the prosodic match and the perceptual results may not reflect the improvement in the spectral match.

The second approach is to define a closed set of natural units and to construct the test stimuli exclusively from these units (Klabbers and Veldhuis, 1998; Wouters and Macon, 1998; Stylianou and Syrdal, 2001; Bjørkan et al., 2005). Since the units for concatenation are predetermined, the cost function is not required for unit selection. The perceptual listening test needs to be completed only once and the results can be reused for testing any number of potential join costs. A join cost can be computed between each pair of concatenated units and in turn can be evaluated by correlating the join cost with the perceptual results.

3.1.2 Utterance length

In creating the test stimuli there are a number of factors that need to be considered regarding the structure of the utterance. For example test sentences were used by Vepa et al. (2002) and Chen and Campbell (1999), while Founda et al. (2001) employed phrases,

Stylianou and Syrdal (2001) and Wouters and Macon (1998) employed words and both Klabbers and Veldhuis (2001) and Donovan (2001) employed even shorter duration stimuli of 130 and 120 ms respectively. Employing longer utterances allows the impact of the wider context to be accounted for but has the disadvantage that it becomes more difficult for the listener to focus on the impact of an individual join. Choosing shorter utterances reduces the impact of neighbouring contexts and makes it easier for the listener to zero in on the join. It is unclear what role neighbouring contexts contribute to the perception of discontinuities. Depending on how the utterance is created, a longer utterance may contain many joins. Discontinuities from a number of joins in the same utterance make it difficult for listeners to judge the continuity of an individual join.

3.1.3 Isolating the source of discontinuity

A major difficulty in testing the perception of discontinuity is isolating a single source of discontinuity in the test database. Signal processing techniques have been employed in some studies to remove other sources of discontinuity (Klabbers and Veldhuis, 1998; Bellegarda, 2004), such as $F0$ and energy mismatch in order to isolate spectral mismatch as the only potential source of discontinuity. However signal processing may degrade the quality of the speech stimuli causing it to sound less natural and may introduce artefacts. This in turn may influence the perceptual results of the human listener. An alternative approach is to analyse the test database to identify discontinuous joins that have a high probability of arising from a known alternative source of discontinuity. For example, in Bjørkan et al. (2005) test stimuli with large $F0$ mismatch are removed from the database.

3.1.4 Perceptual results

In order to generate perceptual results, human listeners must provide a response to the level of discontinuity they perceive for each test stimulus. Typically in the case where a closed set of units were employed to construct the test stimuli each listener is asked to judge a stimulus as continuous or discontinuous or alternatively to rate the perceived degree of discontinuity on a scale. Gathering accurate results is a difficult and laborious task, as

listeners may respond inconsistently particularly when rating degrees of discontinuity on a scale. Donovan (2001) for example, reports a particularly low correlation (a mean listener to listener correlation of 0.39) in a study where listeners rated the degree of discontinuity on a scale. Providing suitable references for continuous and discontinuous speech at each level of the scale may reduce this problem. Deciding if a join is continuous or discontinuous is easier than rating the degree of discontinuity.

3.1.5 Relating perceptual results to join costs

Once perceptual results have been successfully collected a strategy is required to relate the perceptual results with distance measures computed from the proposed join costs. For the case of perceptual results that rate the degree of discontinuity on a scale a correlation coefficient is computed to evaluate the correlation between the proposed join cost and the perceptual results. When the perceptual results are a binary continuous/discontinuous classification, receiver operator characteristic (ROC) curves are typically used to evaluate the performance of the candidate join costs (Klabbers and Veldhuis, 2001; Stylianou and Syrdal, 2001).

Klabbers et al. (2007) provides a number of guidelines to ensure more reliable perceptual results drawing on experience from previous studies. These can be summarised as follows:

- Listeners should judge one join at a time. When presented with an utterance that contains more than one concatenation, the perception of one discontinuity can have an effect on judging adjacent joins.

- Providing a reference stimulus produces more reliable results. Without a reference listeners may rely on preceding stimuli as a reference point for both continuous and discontinuous speech.

- Try to isolate a single source of discontinuity. A number of known sources of discontinuities exist; focusing on one source of mismatch at a time will simplify the experiment and analysis.

A summary of the approaches adopted in previous studies is presented in Table 3.1.

Study	Test stimuli	Perceptual results	Evaluation
Klabbers and Veldhuis (1998)	2248 vowels: /a:/, /A/, /i/, /I/, /u/, SAMPA notation	Continuous or discontinuous binary choice	ROC curves
Wouters and Macon (1998)	166 English words vowels: /aa/, /ae/, /iy/, /uw/, OGI-bet notation	Rate degree of discontinuity on 5 point scale	Correlation coefficient
Chen and Campbell (1999)	6 Japanese sentences	Rate degree of discontinuity on 5 point scale	Correlation coefficient
Stylianou and Syrdal (2001)	2016 English words MRT word list	Continuous or discontinuous binary choice	ROC curve, Bhattacharyya distance
Donovan (2001)	112 English CV pairs voiced speech phone classes	Rate degree of discontinuity on 5 point scale with MOS	Correlation coefficient
Founda et al. (2001)	Greek phrases	Choose preference	Percentage of preferences
Vepa et al. (2002)	5 English sentences diphthongs: /ey/, /ow/, /ay/, /aw/, /oy/	Rate degree of discontinuity on 5 point scale	Correlation coefficient
Tsuzaki and Kawai (2002)	438 Japanese stimuli	Distinguish natural and synthetic stimuli	Percentage of preferences
Bellegarda (2004)	Concatenated words; vowels and consonants	Compare two utterances and rank based on discontinuity	Number of preferences
Bjørkan et al. (2005)	Norweigian long vowels: /A:/ and /e:/, SAMPA notation	Continuous or discontinuous binary choice	ROC curves

Table 3.1: Previous perceptual studies to evaluate join costs.

3.2 Approach adopted

Based on the the points noted in section 3.1, the following approach was adopted in order to construct a perceptual experiment with the aim of achieving consistent and reliable results, enabling the establishment of a framework to correlate human perceptual results with the proposed join costs.

- The database of test stimuli should be constructed from a closed set of natural speech units such that any number of join costs can be tested with the same perceptual results.
- The test database should exhibit a broad phonetic coverage containing joins in many vowel types within varying contexts.
- Each test utterance should be a single word with a single join contained in the vowel centre. This allows listeners to focus on one join at at time and reduces the distraction due to the pitch contour and linguistic context associated with longer utterances.
- Listeners should be required to listen to each test word and judge it continuous or discontinuous. Minimising the complexity of the task for the listener ensures more consistent and reliable results between listeners. Listeners will be allowed to listen to the test words as many times as they require.
- Reference words should be provided in order to promote consistency.
- Minimal signal processing should be applied in order to retain the natural quality of the speech.
- Incidences of discontinuities that have a high probability of being due to $F0$ or energy discontinuity should be identified and removed from the test database in order to isolate a set of stimuli that contain spectral discontinuities.

3.2.1 Test corpus

A database of test stimuli was constructed adopting the approach of Stylianou and Syrdal (2001). The inventory of units consisted of 300 words recorded from an adult male. The words recorded are presented in Table 3.2 and are taken from the Modified Rhyme Test (MRT) (House et al., 1963). The inventory of words was recorded in a hemi-anechoic recording studio at a sampling frequency of 16 kHz. The 300 word inventory consisted of 50 sets of 6 words. Almost all of the words have the structure of an initial consonant, vowel and final consonant. Two sets are vowel final and a small number of words end in a consonant cluster. Within each set the words share the same vowel nucleus and differ in the initial or final consonant (the onset and coda). For example, set 1: *'went, sent, bent, dent, tent, rent'*, varies in the initial consonant and set 3: *'pat, pad, pan, path, pack, pass'*, varies in the final consonant.

The database of 300 words was used to construct a test database of concatenated words. Each of the 300 words was divided into two half-words. For each of the 50 sets this results in 12 half words, 6 left half words and 6 right half words, which made up the left and right units respectively. Every possible combination within a set of left and right units were concatenated, yielding 36 test words for each set and 1800 test words in total. Within each set of 36 concatenated words, 30 result from left and right units not originating in the same word. The remaining 6 concatenated words were created from units originating in the same word and are resynthesised versions of the original recordings. Only test words containing left and right units originating from different recorded words have the potential to contain audible discontinuities, 1500 in total. The 300 resynthesised words were used as control words to validate each listeners perceptual results. If a listener judged a control word as discontinuous that listeners results for that set were rejected.

The test words were concatenated using pitch synchronous overlap and add that exploited pitch marking to maintain $F0$ continuity across the join. Pitch continuity must be maintained across the join otherwise a discontinuity will occur that is the result of a bad concatenation procedure as opposed to incompatible speech units. The pitch marks were obtained using Praat software (Boersma and Weenink, 2005). This simple concatenation

Set 1	went	sent	bent	dent	tent	rent
Set 2	hold	cold	told	fold	sold	gold
Set 3	pat	pad	pan	path	pack	pass
Set 4	lane	lay	late	lake	lace	lame
Set 5	kit	bit	fit	hit	wit	sit
Set 6	must	bust	gust	rust	dust	just
Set 7	teak	team	teal	teach	tear	tease
Set 8	din	dill	dim	dig	dip	did
Set 9	bed	led	fed	red	wed	shed
Set 10	pin	sin	tin	fin	din	win
Set 11	dug	dung	duck	dud	dub	dun
Set 12	sum	sun	sung	sup	sub	sud
Set 13	seep	seen	seethe	seek	seem	seed
Set 14	not	tot	got	pot	hot	lot
Set 15	vest	test	rest	best	west	nest
Set 16	pig	pill	pin	pip	pit	pick
Set 17	back	bath	bad	bass	bat	ban
Set 18	way	may	say	pay	day	gay
Set 19	pig	big	dig	wig	rig	fig
Set 20	pale	pace	page	pane	pay	pave
Set 21	cane	case	cape	cake	came	cave
Set 22	shop	mop	cop	top	hop	pop
Set 23	coil	oil	soil	toil	boil	foil
Set 24	tan	tang	tap	tack	tam	tab
Set 25	fit	fib	fizz	fill	fig	fin
Set 26	same	name	game	tame	came	fame
Set 27	peel	reel	feel	eel	keel	heel
Set 28	hark	dark	mark	bark	park	lark
Set 29	heave	hear	heat	heal	heap	heath
Set 30	cup	cut	cud	cuff	cuss	cub
Set 31	thaw	law	raw	paw	jaw	saw
Set 32	pen	hen	men	then	den	ten
Set 33	puff	puck	pub	pus	pup	pun
Set 34	bean	beach	beat	beak	bead	beam
Set 35	heat	neat	feat	seat	meat	beat
Set 36	dip	sip	hip	tip	lip	rip
Set 37	kill	kin	kit	kick	king	kid
Set 38	hang	sang	bang	rang	fang	gang
Set 39	took	cook	look	hook	shook	book
Set 40	mass	math	map	mat	man	mad
Set 41	ray	raze	rate	rave	rake	race
Set 42	save	same	sale	sane	sake	safe
Set 43	fill	kill	will	hill	till	bill
Set 44	sill	sick	sip	sing	sit	sin
Set 45	bale	gale	sale	tale	pale	male
Set 46	wick	sick	kick	lick	pick	tick
Set 47	peace	peas	peak	peach	peat	peal
Set 48	bun	bus	but	bug	buck	buff
Set 49	sag	sat	sass	sack	sad	sap
Set 50	fun	sun	bun	gun	run	nun

Table 3.2: MRT word list recorded for unit inventory.

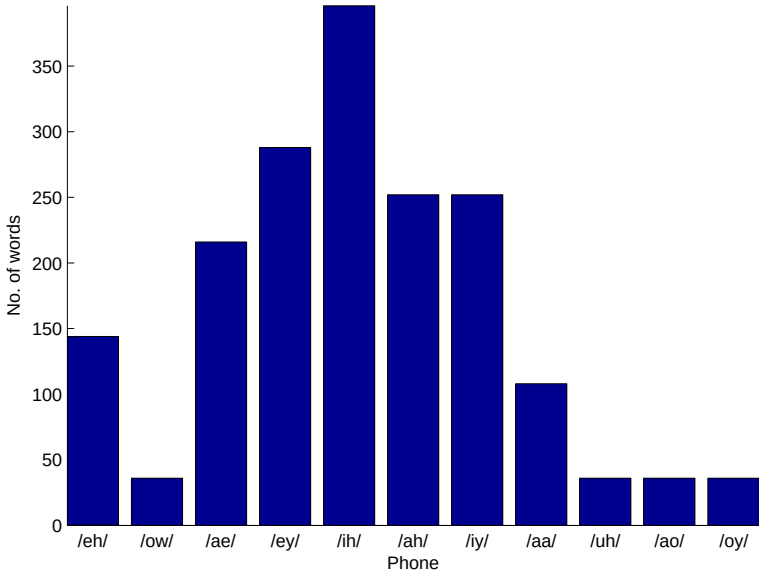

Figure 3.1: Bar chart indicating the number of occurrences of each phone type containing a concatenation in the 1800 word database of test stimuli.

procedure was adopted to minimise the complexity of the signal processing such that any acoustic artefacts that arose were due to the incompatibility of the units and not due to signal processing artefacts or bad concatenation procedure. The overlap and add method will smooth the spectrum between the units in the region of overlap. This may remove some potential discontinuities from the test database.

3.2.2 Phonetic coverage

In the test database all the joins are contained in vowel centres. The number of incidences of joins by each vowel type is indicated in Fig. 3.1. The vowels are labelled with ARPABET notation (Huang et al., 2001). A phonetic description of each vowel type is given in Table 3.3. A number of previous studies have reported the rate of occurrence of discontinuities based on the phone class or phonetic description. For example Klabbers and Veldhuis (2001) found significantly more discontinuities occurred in /u/ (73%) than in /a:/ (17.1%). Syrdal (2001) found that discontinuities were more likely to occur in diphthongs, an idea that was further explored in the work of Vepa et al., their study (Vepa and King, 2006) deals exclusively with joins contained in diphthongs.

The number of occurrences of each context, as defined by the the onset and coda, of

ARPABET	Vowel class	Description	No. words	Example
/iy/	Long monophthong	Front close	252	feel
/aa/		Back open	108	dark
/ao/		Open mid back round	36	law
/ae/		Front open	216	pat
/eh/	Short monophthong	front open-mid	144	ten
/ih/		Front close	396	hit
/ah/		Open mid-back	252	cut
/uh/		Back close mid round	36	book
/oy/	Diphthong	/ao/ to /ih/	36	oil
/ow/		/eh/ to /ih/	36	hold
/ey/		Front close-mid	288	day

Table 3.3: Phonetic description and the number of occurrences of each phone in the database containing a join.

ARPABET	Voicing	Phone class	No. in onset	No. in coda
/m/	Voiced	Nasal	36	36
/n/			0	144
/ng/			0	36
/l/	Voiced	Liquid	36	144
/r/			36	36
/ey/	Voiced	Vowel	0	36
/ao/			0	36
/d/	Voiced	Plosive	72	72
/b/			108	0
/g/			0	36
/p/	Unvoiced	Plosive	180	72
/t/			72	108
/k/			108	72
/s/	Unvoiced	Fricative	180	72
/f/			36	0
/hh/	Unvoiced	Glottal fricative	36	0

Table 3.4: Description of each phonetic context in the database that surrounds the vowel centre containing the join on the left, onset, and right hand side, coda.

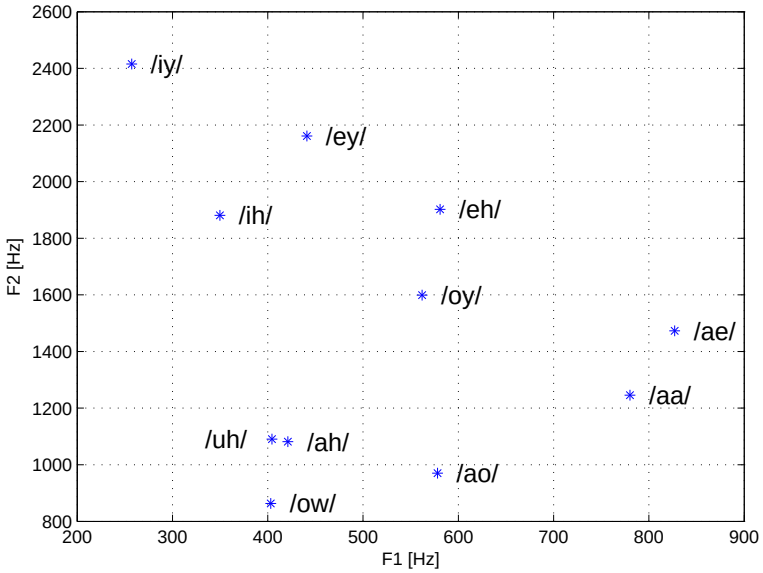

Figure 3.2: Formant chart computed from the unit inventory employing formant estimates from the centre of the voiced region for each of the 300 recorded words.

each test word is presented in Table 3.4. The context surrounding the vowel containing the join is important as co-articulation will influence the realisation of the vowel nucleus. Syrdal (2001) reported that discontinuities were more likely to occur in joins surrounded by the sonorants /l, r, m, n, ng/.

3.2.3 Speaker and speaker characteristics

The recorded speaker was an adult male with an Irish accent and was a non-professional speaker. The speaker was requested to deliver the words for recording in a neutral voice. No carrier sentence was employed for the recording. The speaker was also requested to, within reason, maintain consistent pitch and loudness for each of the six words within any given set. This was somewhat facilitated by the rhyming nature of the words within each set.

Natural variations occur in speech during the production of a given phoneme and the physiology of a given speaker will impose the characteristics of that speaker on the utterance produced. For example formant values vary depending on the length and shape of the vocal tract of the speaker. A formant chart computed from the the inventory of 300 words recorded is illustrated in Fig. 3.2. This illustrates the location of the formants for the

Vowel	$\sigma(F1)$	$\sigma(F2)$
/iy/	38.995	36.883
/aa/	74.189	108.635
/ao/	14.628	25.712
/ae/	121.351	103.709
/eh/	63.021	335.653
/ih/	62.12	540.644
/ah/	56.165	45.9817
/uh/	29.383	227.404
/oy/	57.705	79.876
/ow/	28.577	56.822
/ey/	41.385	102.617

Table 3.5: Indicating the standard deviation of $F1$ and $F2$ for each vowel type in the unit inventory.

speaker with respect to vowel coverage in the database. Formants were computed at the unit boundaries (the vowel centre of each recorded word) over a 40 ms Hanning window. The formant estimates were computed from the poles of a 16^{th} order all-pole LPC model. The LPC analysis was carried out using the covariance method (Hayes, 1996). The raw speech was pre-emphasised with the filter $H(z) = 1 - 0.95z^{-1}$. The formant values in Fig. 3.2 represent the mean formant value computed for each vowel type in the database. For a given speaker variations will also occur. The standard deviation, σ, computed for $F1$ and $F2$ for each vowel type is indicated in Table 3.5. Notable variance exists for a number of vowels in both $F1$ and $F2$. If a speaker exhibits significant variance in the formant structure of a given vowel, joins made within that vowel set have a greater potential to exhibit spectral mismatch.

A pitch histogram for the speaker is given in Fig. 3.3. $F0$ analysis was performed using the YIN algorithm (de Cheveigné and Kawahara, 2002) with a window length of 40 ms. Pitch was estimated throughout the voiced region of each of the 300 word inventory. These estimates were used to construct the pitch histogram in Fig. 3.3. The pitch of the speaker is important from a perceptual point of view. Syrdal (2001) found the rate of occurrence of discontinuities for male and female speakers to be significantly different, with a higher occurrence for female speech. The spacing of harmonics in the frequency domain depends

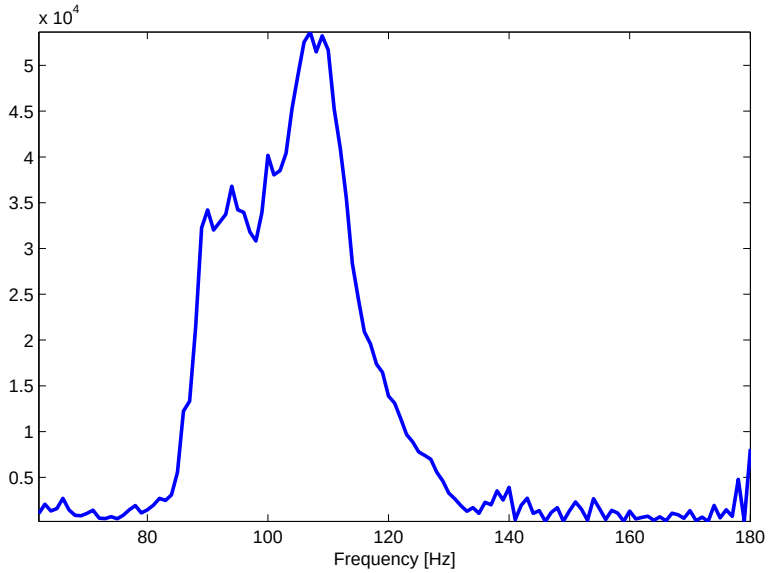

Figure 3.3: Pitch histogram computed from all pitch estimates throughout the voiced region of the 300 recorded words.

directly on the pitch and dictates the number of harmonics within each auditory channel of a human listener. Harmonic spacing has proven to be important in other perceptually driven speech applications. For example in the speech coding literature the perceptual role of the phase spectrum has been documented to be considerably different for both male and female speech due to differences in pitch (Skoglund and Kleijn, 2000). Kim (2001) proposes a model based on the number of harmonics in a specific critical band of the auditory spectrum to explain the effect of pitch on phase perception between high and low pitch speakers.

3.3 Subjective test

The perceptual test was divided into subtests, each containing 36 concatenated words. At the start of a subtest the listener was presented with examples of audible discontinuities for reference. The test required the listener to make a forced decision for each test word, continuous or discontinuous. The listener was provided with the original recorded words that were used to create each of the concatenated words for comparison, which served as a reference for continuous natural speech. Each test word could be repeated as often as

the listener requested. Each subtest contained six control words to validate each listener's results. Each listener undertook the test in a quiet environment using headphones. Twelve listeners in total contributed perceptual results with coverage of three listeners per subtest. A majority scoring system was employed to decide if a test word was continuous or discontinuous.

Listeners were asked not to perform more than two subtests in succession. Despite this listeners still reported that detecting discontinuities in the test words provided was a difficult and laborious task.

3.3.1 Results of the perceptual experiment

From the perceptual experiment a total of 434 discontinuities were detected by the human listeners after employing the majority scoring system on the individual perceptual results. Of the words judged to be discontinuous by the majority voting system, 66.13% of the words were unanimously judged disontinuous and the remaining 33.87% were judged discontinous on the basis of having a majority of listeners judging the word discontinuous. This leaves a total of 1366 continuous words deemed to be continuous. Of the continuous words 82.87% of them were unanimously judged continuous by the listeners. The remaining 17.13% were deemed continuous on the basis of having a majority of listeners judging the word continuous.

3.3.2 Phonetic analysis of results

The number of discontinuities detected by vowel type is listed in Table 3.6. Although no distinct and obvious pattern emerges a number of points can be noted.

- The largest percentage of discontinuities across all vowel is for the diphthong /oy/. Previous studies have reported that discontinuities are more likely to occur in diphthongs than in monophthongs due to the spectral dynamics in diphthongs.
- The diphthong /ow/ is the only vowel in the database to contain zero discontinuities.
- The two vowels, /ao, ow/, contain the lowest percentages of discontinuities. This

ARPABET	No. Discon	% Discon	No. of words
/iy/	69	27.381	252
/aa/	31	28.703	108
/ao/	3	8.333	36
/ae/	40	18.5185	216
/eh/	31	21.5278	144
/ih/	114	28.7879	396
/ah/	63	25	252
/uh/	7	19.444	36
/oy/	13	36.111	36
/ow/	0	0	36
/ey/	63	21.875	288

Table 3.6: The number and percentage of discontinuities with respect to the vowel type containing the join.

difference is significantly lower than for all other vowel types. These are the only two incidences of vowels with lip rounding in the database.

- Vowels with large standard deviations in formant values across the database coincide with vowels with higher rates of discontinuity. The three vowels with the highest rates of discontinuity, /oy, ih, aa/ are a subset of the vowels with the four highest standard deviations in $F1$, /aa, eh, ih, oy/.

- Vowels with low standard deviations in formant values across the database coincide with fewer occurrences of discontinuities. The two vowels with the least number of discontinuities, /ow, ao/ are a subset of the vowels with the three smallest variances in $F1$, /ao, ah, ow/.

A number of vowels in the database only contain 36 test words. The number of discontinuities for these vowels in the database is low. This could be the reason why more definitive patterns do not emerge in the results presented.

The number and percentage of discontinuities for each vowel category is illustrated in Table 3.7. Counter to previous studies the lowest rate of discontinuity is found for the diphthong /ow/. This may be explained by the low variance in $F1$ in /ow/. In contrast, the highest rate of discontinuities was found in the diphthong /oy/. With the exception of

Vowel type	No. Discon	% Discon	No. of words
Long vowel	143	23.3	612
Short vowel	215	25.966	828
Diphthong	76	21.11	360

Table 3.7: The number and percentage of discontinuities with respect to each vowel category in the database.

rounded vowels, /ao, ow/, and the diphthong /oy/ all vowels had relatively similar rates of discontinuity. More insightful results were obtained by considering vowels individually and noting the variance in formants than by categorising them into diphthongs and long and short monophthongs.

A breakdown of the occurrence of discontinuities based on phonetic context for varying onset and coda is illustrated in Table 3.8 and Table 3.9 respectively. The largest rate of discontinuity was found for /m/ when it occurred in the coda, with 52.78% of test words ending in /m/ containing a discontinuity. High rates of discontinuity were also obtained for codas containing /t/ and /r/. The lowest rates of detection occured for test words with vowel endings. The largest incidence of discontinuities for contexts with respect to the onset ocurred for /d/, with 45.83% of test words begining in /d/ containing discontinuities. High rates of discontinuity were also obtained for onsets containing /f/ and /t/. This suggests that both the context and the vowel centre influence the likelihood of a discontinuity ocurring.

3.4 Relating subjective and objective measures

In order to quantify the performance of each candidate measure it is necessary to use a performance metric that measures the degree of correlation between the candidate measures of spectral continuity and the human perceptual results. The area under the ROC curve (AUC) is taken as the performance measure throughout the remainder of this thesis.

ARPABET	No. in onset	No. Discon	% Discon
/m/	36	6	16.667
/n/	0	0	0
/ng/	0	0	0
/l/	36	4	11.111
/r/	36	10	27.778
/ey/	0	0	0
/ao/	0	0	0
/d/	72	33	45.8333
/b/	108	21	19.444
/g/	0	0	0
/p/	180	47	26.111
/t/	72	28	38.889
/k/	108	20	18.519
/s/	180	48	26.667
/f/	36	15	41.667
/hh/	36	5	13.889

Table 3.8: The number and percentage of discontinuities in terms of the consonant context in the onset.

ARPABET	No. in coda	No. Discon	% Discon
/m/	36	19	52.778
/n/	144	23	15.972
/ng/	36	10	27.778
/l/	144	25	17.361
/r/	36	10	33.333
/ey/	36	2	5.556
/ao/	36	3	8.333
/d/	72	25	34.722
/b/	0	0	0
/g/	36	4	11.111
/p/	72	17	23.611
/t/	108	38	35.185
/k/	72	15	20.833
/s/	72	20	27.778
/f/	0	0	0
/hh/	0	0	0

Table 3.9: The number and percentage of discontinuities in terms of the consonant context in the coda.

3.4.1 Receiver operating characteristic curves

Each measure was evaluated by generating an ROC curve (Duda and Hart, 2001). Two probability density functions, $p(\kappa|0)$ and $p(\kappa|1)$, were estimated for each distance measure based on the perceptual results for continuous and discontinuous joins respectively. ROC curves were calculated from the probability density functions and provided information regarding the separability of $p(\kappa|0)$ and $p(\kappa|1)$, for each distance measure. The ROC curves were computed by calculating the hit rate, P_H (3.1), the probability of correctly detecting a discontinuity and the false alarm rate, P_{FA} (3.2), the probability of classifying a continuous join as discontinuous, for varying distance thresholds κ.

$$P_H(\kappa_0) = \int_{\kappa_0}^{\infty} p(\kappa|1)d\kappa \tag{3.1}$$

$$P_{FA}(\kappa_0) = \int_{\kappa_0}^{\infty} p(\kappa|0)d\kappa \tag{3.2}$$

The ROC curve was constructed by plotting the pairs P_H and P_{FA} for each threshold value, κ_0, from 0 to ∞.

Area under the ROC curve

After computing the ROC curve the AUC (3.3) is calculated. The area calculation assumed a piecewise linear approximation between consecutive points on the ROC curve. The area under each piecewise linear segment was computed, and summed to give an estimate of the AUC value.

$$AUC = \int_0^1 P_H dP_{FA} \tag{3.3}$$

The AUC can be interpreted as the probability of correctly classifying a join. The AUC value is bound between 0 and 1. An ideal measure that perfectly discriminates between continuous and discontinuous joins would get an AUC value of 1 and an arbitrary or random measure with no discriminating information would get an AUC value of approximately 0.5, corresponding with the probability of pure chance in a binary classification task.

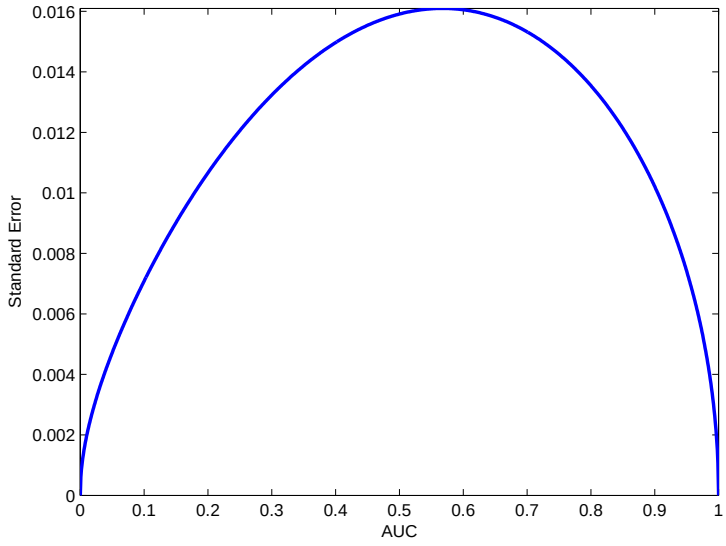

Figure 3.4: Standard error versus the AUC value.

Standard error for ROC curves

The standard error, SE, of the AUC value computed from the ROC curve can be estimated by equation 3.4 (Hanley and McNeil, 1982).

$$SE = \sqrt{\frac{A(1-A) + (N_d - 1)(Q_1 - A^2) + (N_c - 1)(Q_2 - A^2)}{N_d N_c}} \tag{3.4}$$

Q_1 and Q_2 are defined in equations 3.5 and 3.6 respectively.

$$Q_1 = \frac{A}{2 - A} \tag{3.5}$$

$$Q_2 = \frac{2A^2}{1 + A} \tag{3.6}$$

The AUC value is represented by A in equation 3.4. N_d and N_c represent the number of discontinuous and continuous samples in the dataset respectively. Equation 3.4 depends only on the the AUC value and the number of continuous and discontinuous samples. In the database the number of continuous and discontinuous samples is constant. A number of discontinuities may be omitted from the computation of certain results, causing N_d to decrease. This occurs if they are identified as having a high probability of containing a discontinuity that is due to either $F0$ or energy mismatch. The corresponding standard

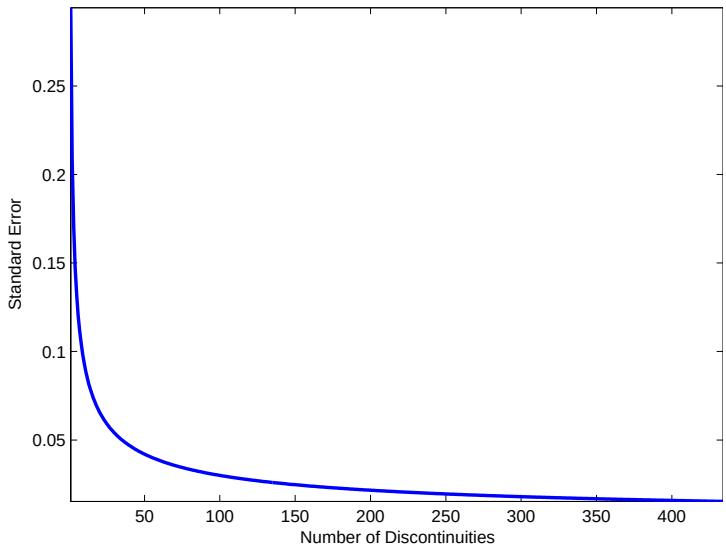

Figure 3.5: Standard error plotted against the number of discontinuities.

error for all possible AUC values from 0 to 1 is illustrated in Fig. 3.4 for 434 discontinuous joins and 1366 continuous joins. The maximum SE estimated is just above 0.016. Note that the curve is not symmetric as the number of continuous and discontinuous joins are not equal. The standard error increases as the AUC value approaches approximately 0.5. This reflects that a more significant error is likely to occur when the PDFs are significantly overlapped. Fig. 3.4 indicates the degree of confidence with which we can interpret the AUC values computed for each candidate measure of spectral continuity. A measure that produces an AUC value that is within the standard error margins of another measure is not statistically, significantly better or worse than that measure.

Fig. 3.5 illustrates the impact of the number of discontinuities, N_d, on the standard error when the AUC value is held constant at 0.7. The standard error decreases monotonically as the number of discontinuities, N_d, increases. When the AUC value is increased the curve is shifted downwards decreasing the standard error and vice-versa when the AUC value decreases towards 0.5. This is important as the number of discontinuities may be reduced for the computation of the AUC value if a number of test words are judged to contain discontinuities that are due to $F0$ or energy mismatch. This curve gives an indication of the number of discontinuities required for a statistically valid evaluation of the proposed join costs.

3.5 Analysis of results

Test words judged to contain an audible discontinuity in the perceptual test may have contained a discontinuity that was not due to spectral mismatch. Based on the perceptual results the test database was analysed with respect to known sources of discontinuity. The objective was to identify words from the test stimuli that were judged as discontinuous due to a discontinuity from a known source of discontinuity. This enables the performance metrics to be computed with and without the inclusion of such words. In this study we were primarily concerned with spectral mismatch so audible discontinuities due to other sources could skew results.

3.5.1 Known sources of discontinuity

Other, known sources, of discontinuity are due to $F0$ mismatch and energy mismatch across a join. To address this issue the database of test words was analysed to identify test words containing an audible discontinuity that was potentially due to $F0$ or energy mismatch.

Fundamental frequency - $F0$

$F0$ trajectories were computed from the original 300 words and $F0$ values were extracted at the unit boundaries to represent each of the units in the inventory. $F0$ analysis was performed using the YIN algorithm (de Cheveigné and Kawahara, 2002) with a window length of 40 ms. The $F0$ distance measure, D_{F0}, was taken as the absolute difference between $F0$ values from the left and right units, equation 3.7. Probability density functions were estimated from the set of $F0$ distance measures for both continuous and discontinuous joins and are illustrated in Fig. 3.6. The PDFs were estimated by constructing a histogram of 40 bins in the range from 0 to the maximum $F0$ distance and normalising (the sum of all probabilities equals one) the histograms to behave like a PDF. Fig. 3.6 indicates the degree of overlap between the distributions of continuous and discontinuous joins with respect to the D_{F0}. The distribution for continuous joins with respect to D_{F0} reduces monotonically from a maximum value corresponding to a distance of zero. The probability decreases approximately exponentially as D_{F0} increases. The distribution for discontinuous joins

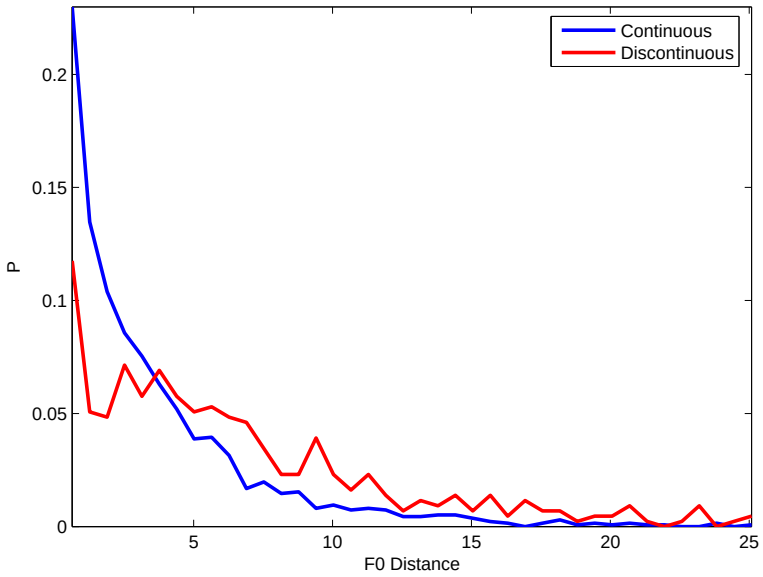

Figure 3.6: Estimated PDFs for continuous and discontinuous joins based on the absolute $F0$ distance.

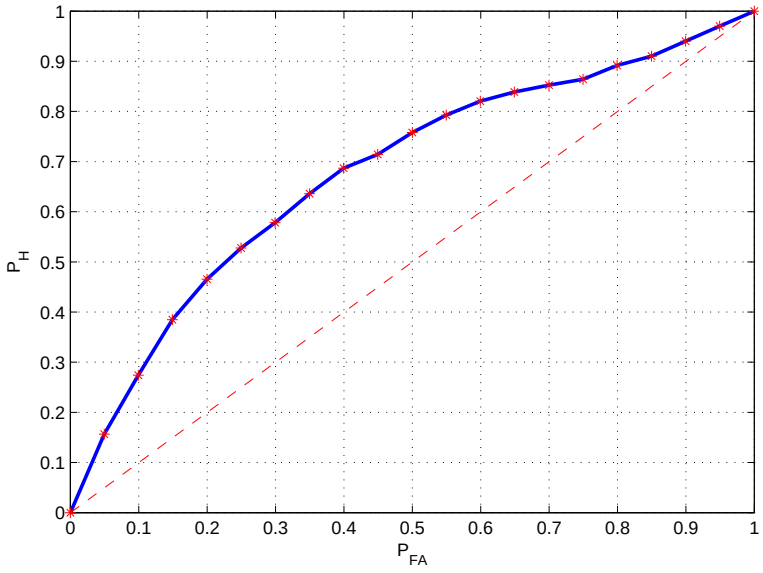

Figure 3.7: ROC curve based on the absolute $F0$ distance (AUC = 0.6784).

was found to be more evenly spread than the distribution for continuous joins and the probability of a join being discontinuous remains higher than the probability of join being continuous above a threshold of approximately 4 Hz.

$$D_{F0} = |F0_{left} - F0_{right}| \tag{3.7}$$

The ROC curve relating the human perceptual results and $F0$ distance measure is illustrated in Fig. 3.7. The AUC value computed from the ROC curve is 0.6784, indicating that $F0$ correlates with the human perceptual results of discontinuity for the test database. This suggests that a number of discontinuities are due to $F0$ mismatch. In order to identify joins that may contain a discontinuity due to $F0$ mismatch, all discontinuous joins were ranked with respect to D_{F0}. The joins with largest D_{F0} were iteratively removed until an AUC value of below 0.55 was achieved. The AUC value of 0.55 was selected as a having a sufficiently low correlation with the perceptual results to ensure that the joins remaining had a significantly low probability of containing a discontinuity due to $F0$. The threshold at which this was achieved corresponded with an absolute $F0$ distance of 7.2 Hz. From the original 434 discontinuities 290 are below the threshold and the remaining 144 are above the threshold. The 144 joins above the threshold have a significant probability of containing a discontinuity due to $F0$ mismatch.

Energy

Energy coefficients were extracted at the unit boundaries for each left and right unit in the inventory. A window length of 10 ms was employed to compute the energy coefficient. The energy distance measure, D_E, was taken as the absolute difference between energy values from the left and right units, equation 3.8. Probability density functions were estimated from the set of energy distances for both continuous and discontinuous joins and are illustrated in Fig. 3.8. The PDFs were estimated as with the $F0$ based analysis. The distribution for continuous joins with respect to energy reduces monotonically similar to the case for D_{F0}. The probability decreases approximately exponentially as the degree of

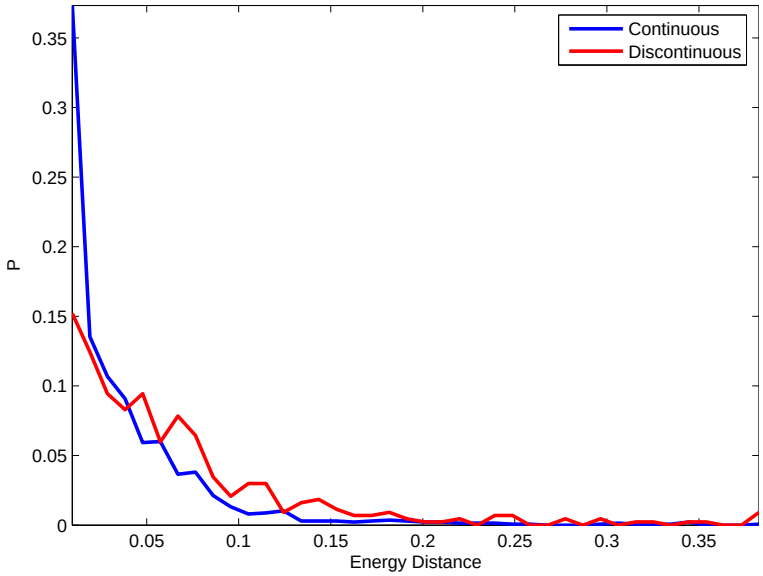

Figure 3.8: Estimated PDFs for continuous and discontinuous joins based on the absolute energy distance.

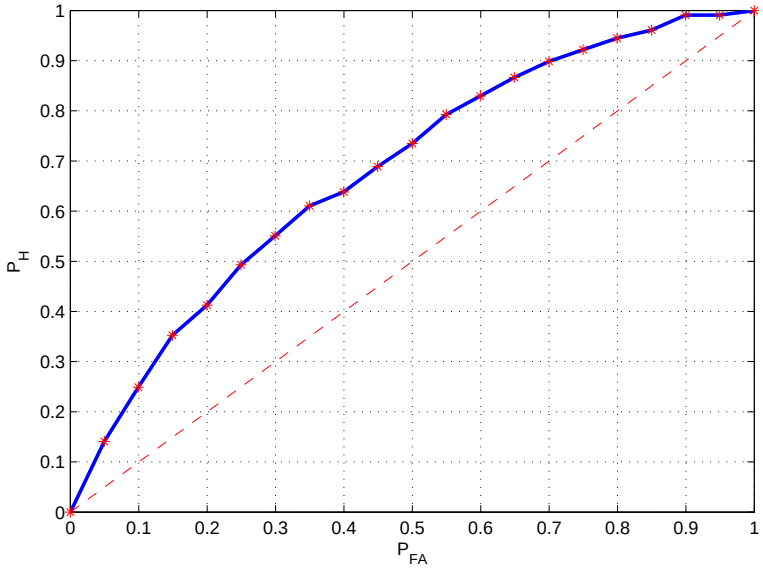

Figure 3.9: ROC curve based on the absolute energy distance (AUC = 0.678).

energy mismatch increases. The distribution for discontinuous joins also decreases consistently as the D_E increases. The probability of a join being discontinuous remains higher after a specific threshold is reached as illustrated in Fig.3.8 for D_{F0}.

$$D_E = |E_{left} - E_{right}| \tag{3.8}$$

The choice of window length was found to have a significant impact on the AUC value. The AUC value corresponding with Fig. 3.9 is 0.678. A threshold was computed as in the $F0$ analysis to remove joins with large energy mismatch, such that the resulting AUC value is reduced to below 0.55. The number of joins containing discontinuities below the threshold was 288, these joins have a significantly low probability of containing a discontinuity due to energy mismatch across the point of concatenation. The remaining 146 joins have a significant probability of containing discontinuities due to energy mismatch.

After removing stimuli with large $F0$ and energy mismatch the database contains 231 discontinuous samples and 1366 continuous samples. This corresponds to a standard error of approximately 0.02 for an AUC value of 0.7. The maximum standard error for this number of discontinuities is 0.021 and occurs for an AUC value of 0.59.

3.6 Summary and conclusion

In this chapter the issues involved in designing a perceptual experiment were considered. A perceptual experiment was described in which human perceptual results were obtained for the perception of discontinuity in concatenated speech. From the experiment a total of 434 discontinuities were detected by the human listeners from a database of 1800 test stimuli. A detailed description of the phonetic coverage of the unit inventory was presented along with an analysis of the rate of occurrence of discontinuity. It was found that the standard deviation of formant estimates for a given vowel could provide a general indication if that vowel was likely to exhibit significant rates of discontinuity or not. Vowels that involve lip rounding were found to contain particularly low incidences of discontinuities. Consonant contexts about a join were also found to influence the occurrence of discontinuities in the

test stimuli. The evaluation procedure to relate perceptual results with distance measures computed from proposed join costs was presented. The same evaluation procedure was employed to identify discontinuities in the database that had a high probability of being due to $F0$ or energy mismatch. For some results it was difficult to draw a definitive conclusion even when certain trends could be observed. This may be due to the limited size of the test database. This is a particularly limiting factor when the test database is considered on a phone by phone basis as certain phones only contain 36 test words and only a small percentage of those words contain a discontinuity.

As well as containing results in its own right this chapter has established a framework to evaluate candidate join costs based on relating the perceptual results by employing the AUC as a performance metric. This framework is used extensively in subsequent chapters to correlate candidate measures of spectral continuity with perceptual results.

Chapter 4

Investigation of Standard Spectral Measures

In speech processing applications, the raw speech signal is typically transformed into a representation that is more suitable for the task at hand. Important signal characteristics should be easily accessible in the transformed domain in order to extract the salient features of the speech signal. Many spectral representations are used in speech processing: the Fourier transform is ubiquitous, MFCCs have become dominant in ASR while LPC representations have become prevalent in speech coding and speech modification. It is not clear what spectral representation is most suitable to represent human perception of spectral continuity in a join cost for unit selection synthesis.

A number of standard spectral representations of speech have been tested as features for spectral join costs for unit selection synthesis. The results of these studies are largely inconclusive and often the results for a specific measure varies considerably between studies, with the most consistent outcome being that no measure correlates satisfactorily with human perception of discontinuity. In speech synthesis systems, MFCCs in conjunction with the Euclidean distance measure are typically used as the spectral join cost. This is largely due to the popularity of MFCCs in other speech applications such as ASR. It is not yet clear however if such parametrisations make for suitable spectral join costs and what limitations they might have for this task. In this chapter a number of standard spectral

representations common in speech processing applications are investigated for the task of detecting discontinuities in the test database described in Chapter 3.

4.0.1 Chapter objectives and overview

The objectives of this chapter are as follows;

- To systematically compare the performance of conventional speech parametrisations as measures of spectral continuity.
- To identify feature extraction subtleties that cause variations in the results.
- To determine what characteristics of each feature set gives rise to an increase or decrease in performance.
- To investigate the limitations of standard spectral representations as measures of spectral continuity.
- To compare the results obtained against previous studies.

This chapter goes beyond a basic comparison of feature sets and seeks to provide the foundation necessary to develop measures, specifically designed to represent human perception of spectral continuity, presented in subsequent chapters. To achieve this requires a deeper understanding of the limitations of conventional feature sets and an appreciation of why these measures do not sufficiently represent human perception of discontinuity. This knowledge can be applied to guide the development of new feature sets to represent spectral continuity.

In this chapter a set of standard speech representations are tested for the task of detecting discontinuities in concatenated speech and their limitations are investigated. The candidate distance measures and feature sets are presented. A detailed presentation of the techniques required for computing each of the candidate feature sets is contained in this chapter. The results for the task of detecting discontinuities in the test database with each of the candidate measures is presented. An analysis of the results is performed to explore

possible sources of variation and the limitations of the candidate features. Conclusions are presented, providing a link between the results, analysis and the objectives of the chapter.

4.1 Join costs

The computation of a join cost to represent a measure of spectral continuity between two units of speech has two distinct stages. The first stage involves the extraction of features from the speech signal for each unit and the second stage requires quantifying the degree of mismatch between the features representing each unit. This section outlines the features and distance measures tested in this chapter as spectral join costs.

4.1.1 Spectral features

The following feature sets are considered in this chapter:

- DFT power spectra ($PS(DFT)$) and log power spectra ($LogPS(DFT)$).
- Harmonic power spectra (PS_h) and log harmonic power spectra ($LogPS_h$).
- LPC power spectra ($PS(LPC)$), cepstral coefficients ($LPCC$) and LSF's (LSF) on both Mel and linear frequency scales.
- MFCCs computed from both DFT ($MFCC(DFT)$)and LPC ($MFCC(LPC)$) magnitude spectra.
- PLP power spectra ($PLPPS$), LSFs ($PLPLSF$) and cepstral coefficients ($PLPCC$).

These feature sets can be broadly categorised into three groupings. Firstly those based on Fourier-like expansions: DFT and Harmonic representations. Secondly, those based on LPC models that relate to physical models of speech production and lastly those based on perceptual modelling: MFCCs and PLP.

4.1.2 Distance measures

In order to quantify the degree of similarity between two feature vectors the distance between the vectors is calculated. A number of distance measures are considered: the

absolute distance (l_1), the Euclidean distance (l_2) the Cos distance and the symmetric Kullback-Leibler distance (D_{skl}) (Klabbers and Veldhuis, 2001).

The l_p distances, equation 4.1, represent a measure of length between the feature vectors $\mathbf{x}$ and $\mathbf{y}$. The choice of p establishes the geometric properties of the distance, in this study $p = \{1, 2\}$, corresponding to absolute and Euclidean distance respectively.

$$l_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{N} |x(i) - y(i)|^p \right)^{1/p} \tag{4.1}$$

The Cos distance, equation 4.2, measures the cosine of the angle θ between the two feature vectors $\mathbf{x}$ and $\mathbf{y}$.

$$Cos(\theta) = \frac{\mathbf{x}^T \mathbf{y}}{|\mathbf{x}| \, . \, |\mathbf{y}|} \tag{4.2}$$

This measure is bound between 1 and -1. A value of 1 indicates that the vectors correlate and are linearly dependent although they may have different scale. A value of 0 indicates that the vectors are orthogonal.

The Kullback-Leibler divergence was proposed in Kullback and Leibler (1951) and is a measure of the separability of distributions. It is a non-symmetric measure and in many applications a symmetric version of the measure has become popular, referred to as the symmetric Kullback-Leibler. The symmetric Kullback-Leibler distance measures the distance between two probability distributions, $x(\omega)$ and $y(\omega)$.

$$D_{skl}(x(\omega), y(\omega)) = \frac{1}{4\pi} \int_0^{2\pi} (x(\omega) - y(\omega)) log \frac{x(\omega)}{y(\omega)} d\omega \tag{4.3}$$

Power spectra and LSFs can be treated as probability distributions, as they are always positive and can be normalised to behave like probability distributions and as such the symmetric Kullback-Leibler distance can be applied to these feature sets. The Kullback-Leibler distance for power spectra is given in equation 4.3. Power spectra as opposed to log power spectra are used in conjunction with the Kullback-Leibler distance as log power spectra will give negative values for power spectra values less than one. This measure is not appropriate for cepstral representations. The discrete sum approximation was used

to implement the symmetric Kullback-Leibler distance using discrete features (Vepa and King, 2004b).

$$D_{skl}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} (x(i) - y(i)) log \frac{x(i)}{y(i)} \tag{4.4}$$

In-depth discussion on approximate and exact computation of D_{skl} is presented in Klabbers and Veldhuis (2003).

4.2 Feature extraction

In this section the details of computing each of the standard feature sets is presented. Feature extraction was found to be important and is presented in detail. This section contains information on preprocessing of the speech signal followed by the key characteristics and implementation details for each of the candidate feature sets. The features sets are categorised as Fourier-like methods, LPC based features and perceptual features.

4.2.1 Preprocessing

Feature extraction was performed over the last window of speech in each left hand unit and over the first window of speech in the right hand unit. A number of common steps precede the computation of each of these feature sets. All features were extracted using a pitch synchronous window one pitch period in duration. This was found to be the optimum windowing strategy and is discussed further in section 4.4.3. The raw speech was pre-emphasised with the filter $H(z) = 1 - 0.95z^{-1}$ and Hanning windowed.

4.2.2 Fourier methods

The Fourier transform is central to many signal processing applications and in speech processing signals are often represented in the Fourier domain. In this section details of computing the Fourier transform spectrum and the harmonic spectrum are presented.

Fourier power spectra

The Fourier transform is an operation that transforms a time domain signal into the frequency domain as a weighted sum of complex exponentials (Oppenheim and Schafer, 1975; Quatieri, 2002). The discrete time Fourier transform as defined for discrete time signals, $x[n]$, is given in equation 4.5 and its inverse is defined in equation 4.6. This definition results in a representation, $X(\omega)$, that is a continuous function of frequency, ω.

$$X(\omega) = \sum_{-\infty}^{\infty} x\,[n]\, e^{-j\omega n} \tag{4.5}$$

$$x\,[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\omega) e^{jwn} d\omega \tag{4.6}$$

In order to exploit the frequency domain representation in practical applications it is necessary to sample the frequency domain signal, equation 4.7, and further to consider finite durations of the signal as defined in the discrete short-time Fourier transform (STFT), equation 4.8, where $w[n]$ is the window function (Allen and Rabiner, 1977; Quatieri, 2002). The choice of window length and the subsequent spacing of samples on the frequency axis is critical to ensure neighbouring harmonics are effectively represented in the Fourier domain.

$$X(n,k) = X(n,\omega)|_{\omega=\frac{2\pi}{N}k} \tag{4.7}$$

$$X(n,k) = \sum_{-\infty}^{\infty} x\,[m]\, w\,[n-m]\, e^{-j\frac{2\pi}{N}km} \tag{4.8}$$

Uniform sampling in the frequency domain constrains the computation of the Fourier transform to a specified set of frequency sample points. The length N of the finite duration window function, $w[n]$, determines the time-frequency resolution of the spectral estimate. In speech applications N is typically selected to give a window of about 20-30 ms. The frequency sampling interval is $\frac{2\pi}{N}$. For voiced speech this means that the shortest possible window length that can resolve individual harmonics is a window length equal to one pitch period. The spacing between individual harmonics in the Fourier domain becomes shorter than the frequency sampling interval for window lengths shorter than a pitch period and

as such individual harmonics can not be effectively represented with a window length of less then one pitch period.

With time-varying signals, longer windows can cause spectral smearing effects on the spectral estimate at a given frequency as the corresponding signal component may not be stationary throughout the duration of the window. This limits the size of the window. Larger windows yield better frequency resolution and shorter windows yield better time resolution. When employing a short-time window for spectral estimation it inherently assumes that the windowed signal is stationary. This is generally a reasonable approximation for voiced speech provided an appropriate window length is chosen. This is typically the case for vowel centres where the spectrum is relatively stationary from one pitch period to the next provided the window of speech contains a small number of pitch periods with relatively constant spectra.

The information contained within each Fourier coefficient, equation 4.9, becomes useful when it is separated into magnitude, $|X\omega)|$, and phase,$\angle X(\omega)$, components as represented in equations 4.10, 4.11 and 4.12.

$$X(\omega) = X_r(\omega) + jX_i(\omega) \tag{4.9}$$

$$X(\omega) = |X\omega)|\, e^{j\angle X(\omega)} \tag{4.10}$$

$$|X\omega)| = (X_r(\omega)^2 + X_i(\omega)^2)^{1/2} \tag{4.11}$$

$$\angle X(\omega) = tan^{-1}\left(\frac{X_r}{X_i}\right) \tag{4.12}$$

Traditionally it has been the magnitude spectrum that has been employed in speech applications due to its perceptual significance, although recent studies have demonstrated the perceptual importance of the phase spectrum (Alsteris and Paliwal, 2006; Paliwal and Alsteris, 2005).

The Fourier power spectrum and log power spectrum were tested for their ability to detect discontinuities in the test database. A 512 point fast Fourier transform (FFT) was employed to compute the magnitude spectrum and squared to compute the power

spectrum. In this study the FFT magnitude spectrum and log power spectrum are tested for the task of detecting discontinuities in the test database.

Harmonic power spectra

Speech can be viewed as having a harmonic component and an aperiodic or noise component. The harmonic component is of particular interest in processing voiced speech. Discontinuities have been found to be particularly problematic (Klabbers and Veldhuis, 2001; Syrdal, 2001) in voiced speech. Such findings motivate investigating the harmonic component as a candidate distance measure to represent spectral continuity. Removing potentially redundant information from the signal, in this case the unvoiced or aperiodic speech component, could improve the performance of the distance measure. A number of techniques exist for separating the harmonic and aperiodic components of speech (Yegnanarayana et al., 1998; Bailly, 1999; Stylianou, 2001). The approach adopted here is that of Stylianou (2001) and is outlined below.

The harmonic estimate, $\mathbf{x}_h$, is computed that optimally fits the original time domain signal in a mean squared error sense. This procedure can be interpreted as a sub-sampled Fourier transform such that only the complex exponentials that coincide with harmonic components are retained in the expansion. Unlike the Fourier transform the frequency of the harmonics are not restricted to the sample points of the frequency grid. The computation of the harmonic components depends on a pitch estimate to determine the frequencies at which the harmonic components occur. Inaccurate pitch estimates lead to inaccurate estimates of the harmonic components.

In this study the harmonic power spectrum and log harmonic power spectrum are tested for the task of detecting discontinuities in the test database. Pitch estimation was carried out with the YIN algorithm (de Cheveigné and Kawahara, 2002). The harmonics were estimated up to 6000 Hz, spectral content above this frequency was assumed to be unvoiced.

4.2.3 Linear predictive techniques

The speech production apparatus is often modelled by the source-filter model (Markel and Gray, 1976). The filter component of the model represents the vocal tract and the source represents the airflow entering the vocal system from the lungs. In the source-filter model of speech the filter is typically modelled as an all-pole filter, $H(z)$ equation 4.13, as the vocal tract only gives rise to spectral resonances which can be effectively modelled by poles. Nasal sounds which give rise to anti-resonances can be modelled by zeros. In this thesis only all-pole modelling is considered. The characteristic polynomial, $A(z)$ in equation 4.14, contains information about the vocal tract configuration and can be represented in many different parametric forms.

$$H(z) = \frac{B}{A(z)} \tag{4.13}$$

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{4.14}$$

A number of algorithms exist to compute the all-pole model (Hayes, 1996); the autocorrelation, covariance and Burg methods. Each of these methods has its specific advantages and disadvantages. A preliminary investigation indicated that each of the algorithms were found to yield similar results with no distinct preference emerging. The autocorrelation method was employed to compute a 16^{th} order all-pole model. The autocorrelation-based LPC coefficients can be calculated from equation 4.15.

$$\mathbf{a} = \mathbf{R}^{-1}\mathbf{r} \tag{4.15}$$

In equation 4.15 $\mathbf{a}$ is the vector of LPC coefficients from the characteristic polynomial, $\mathbf{r}$ is the truncated autocorrelation vector and R is the autocorrelation matrix.

A number of LPC-based feature sets exist with different features preferred for different applications. LSFs are often preferred for signal modification algorithms due to their good interpolation properties and the guarantee of a stable filter subsequent to modification. Cepstral-based parameters are often preferred for pattern recognition tasks. The LPC-based representations considered include: LPC power spectra, LPC cepstra and LSFs,

each of which is discussed below.

LPC power spectrum

The power spectrum, $P(e^{jw})$, can be computed from the all-pole model by equation 4.16. This represents an estimate of the spectral envelope of the vocal tract.

$$P(e^{jw}) = \frac{|B|^2}{\left|1 + \sum_{k=1}^{p} a_k e^{-jkw}\right|^2} \tag{4.16}$$

The spectral envelope was computed at 512 sample points along the frequency axis. The sampling was chosen to be consistent with the FFT length for FFT-based power spectrum.

LPC cepstrum

A relationship exists that relates the LPC coefficients, a_k, to cepstrum coefficients (Schroeder, 1981), c_n. The recursive relation (4.17) was used to compute the LPC cepstrum.

$$c_n = -a_n - \frac{1}{n}\sum_{k=1}^{n-1} k c_k a_{n-k} \tag{4.17}$$

The first 20 cepstral coefficients were computed and the first was discarded yielding 19 coefficients as the LPC cepstral feature vector.

Line spectral frequencies

The roots of the characteristic polynomial contain information regarding the resonant peaks in the spectrum of the all-pole model. The frequency of the peaks can be computed as the phase angle of the complex roots.

The LSF representation (Itakura, 1975) of LPC models has become popular in speech processing due to the their useful properties (Soong and Juang, 1984). LSFs are computed by constructing two new polynomials, $P(z)$and $Q(z)$ equations 4.18 and 4.19 respectively, from the characteristic polynomial $A(z)$. The characteristic polynomial contains all the information in the LPC model relating to the shape of the spectral envelope. The LSFs are computed as the phase angles of the roots of $P(z)$ and $Q(z)$. The roots of the polynomials

$P(z)$ and $Q(z)$ always lie on the unit circle and as such are fully described by the phase angle of their respective roots.

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1}) \tag{4.18}$$

$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1}) \tag{4.19}$$

4.2.4 Perceptual representations

Perceptually-based features have become popular in many speech applications. Humans can generally outperform human-engineered solutions for most speech applications, for example sound source separation and ASR. This has resulted in the development of feature sets that try to mimic certain processes within the human auditory system. The most popular perceptually-based features are MFCCs and PLP.

Mel-frequency cepstral coefficients

MFCCs attempt to model a number of known characteristics of the human auditory system. Simplified models of auditory processes are employed to produce spectral estimates more consistent with the auditory response. MFCCs have proven to be highly beneficial in ASR (Davis and Mermelstein, 1980).

The following auditory processes are modelled in the computation of MFCCs;

- Mel-frequency scale, modelling the non-linear frequency scale of the auditory system.
- Critical band frequency resolution mimicking that of the basilar membrane.
- Instantaneous amplitude compression.

In the computation of MFCCs the power spectrum is estimated, this is often implemented using either LPC or DFT power spectrum. Subsequent to this the Mel-scale filterbank is applied to the spectrum by applying a set of triangular filters designed to

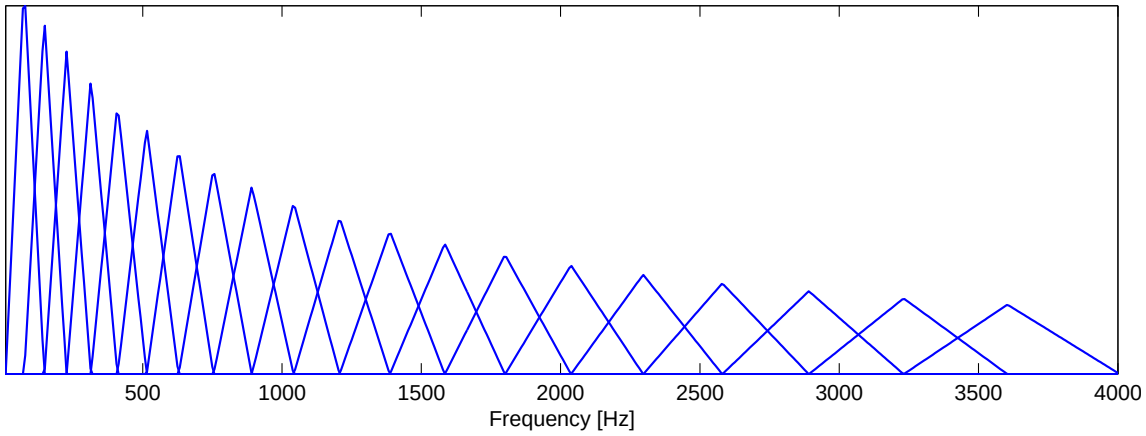

Figure 4.1: MFCC filters, indicating both relative weighting and frequency resolution.

mimic the critical bandwidths of the peripheral auditory system on the non-linear Mel-frequency scale, equation 4.21. The centre frequencies and band edges of the filters are linear below 1000 Hz and increase logarithmically with frequency above this point. The filters are normalised with respect to their bandwidths.

The Mel-scale filterbank employed in the MFCC implementation used in this thesis is illustrated in Fig. 4.1 and employs 20 filters spanning the frequency range from 0 Hz up to 4000 Hz. Frequencies above 4000 Hz were discarded. The log operation is performed on the filterbank outputs and the DCT is applied yielding the MFCCs. The implementation employed resulted in 20 MFCC coefficients. The first coefficient was discarded resulting in feature vectors with 19 coefficients. Although the filterbank attempts to model the frequency resolution of the auditory system it is ultimately limited by the time frequency resolution of the spectral estimation stage. In this thesis MFCCs were computed using power spectra computed from both the DFT and LPC analysis.

Perceptual linear prediction

The approach to auditory modelling in PLP (Hermansky, 1990) is quite similar in principle to that employed in the calculation of MFCCs. The human auditory system is modelled by a non-linear frequency scale, the application of a filterbank and an amplitude compression stage. The models used to implement the non-linear frequency scale, filterbank and amplitude compression differ in implementation to MFCCs, but aim to model the same

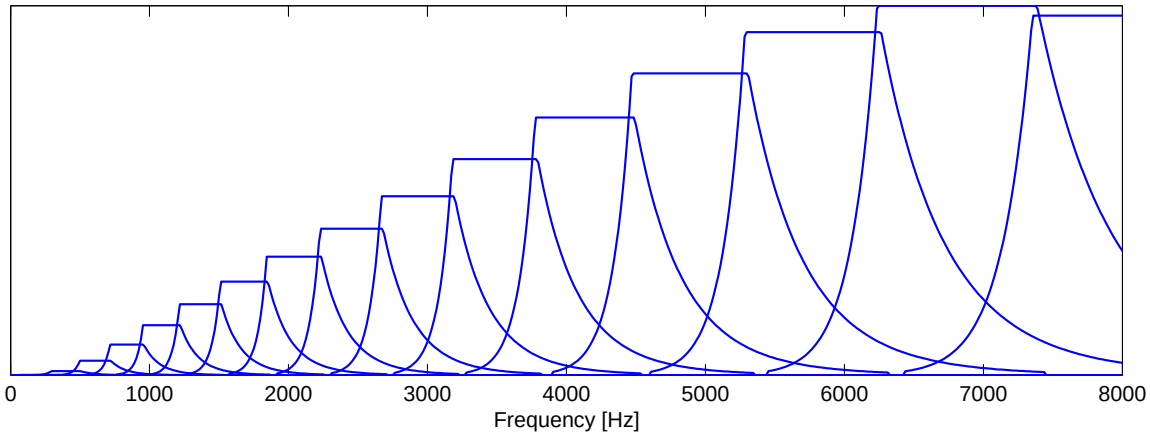

Figure 4.2: PLP filters, indicating both relative weighting and frequency resolution.

phenomena. The critical difference with PLP over MFCCs is that the final approximation of the auditory spectrum is used to produce an LPC model yielding a significantly different parametric form. In the computation of the PLP model the power spectrum is estimated and input to the filterbank. The filterbank implementation is illustrated in Figure 4.2. The Bark scale, equation 4.20, is employed as the non-linear frequency scale (Gold and Morgan, 2000).

$$f_{\text{Bark}} = 6 \log \left(\frac{\omega}{1200\pi} + \left(\left(\frac{\omega}{1200\pi} \right)^2 + 1 \right)^{1/2} \right) \tag{4.20}$$

The filters are one Bark apart and are approximately rectangular. An equal-loudness transformation is also included to model the unequal sensitivity of human hearing at different frequencies. The amplitude compression stage is implemented by computing the cubic root of the filterbank outputs. The resulting approximation of the auditory power spectrum is transformed into an autocorrelation sequence by taking its inverse Fourier transform. The autocorrelation sequence is used to apply the standard autocorrelation technique to compute the coefficients of an all-pole model.

As with standard LPC models a number of parametric forms exist to transform the all-pole model parameters. In this thesis the PLP parameters were transformed into power spectra, LSFs and cepstral coefficients. The PLP implementation used a 5^{th} order PLP model.

Mel LSF

Perceptual based frequency warping is often used on LSFs. In this thesis Mel LSFs are considered. Warping the LSFs onto a Mel-scale has the overall effect that a mismatch between two LSFs in a higher frequency region contributes relatively less to the overall distance than a similar mismatch in low frequency regions. The Mel-scale approximates the frequency scale of the human auditory system (Flanagan, 1972).

$$f_{\text{Mel}} = (1127)\log\left(1 + \frac{f_{\text{Hz}}}{700}\right) \tag{4.21}$$

4.3 Results

The results for each of the candidate feature sets and distance measures are presented in this section. Each measure is tested for its ability to detect discontinuities in the test database. The human perceptual results are related to the distance measure by ROC curves. The AUC value is used to quantify the effectiveness of each candidate measure. A number of feature sets have results omitted for D_{skl} as this measure is not applicable to those feature sets. The results are presented in three subsections: Fourier-based measures, LPC measures, perceptual features.

4.3.1 Fourier-based measures

The results for detecting discontinuities with the power spectra and log power spectra from both the Fourier spectra and harmonic spectra are presented in Table 4.1. Most of the DFT-based measures have similar AUC values regardless of the choice of distance measure. Log spectra were found to outperform standard power spectra and the best measure from this section is the absolute distance between DFT log power spectra. The harmonic-based measures did not score as well as the DFT-based measures. The same trend emerges for the harmonic measures regarding an improvement in performance with log power spectra compared to standard power spectra.

Features	l_1	l_2	Cos	D_{skl}
$LogPS(DFT)$	0.758	0.754	0.756	-
$PS(DFT)$	0.752	0.71	0.732	0.751
$LogPS_h$	0.699	0.697	0.674	-
PS_h	0.642	0.643	0.678	0.665

Table 4.1: Results for each of the spectral transformation feature sets and distance measure combination, the table entries indicate the AUC.

4.3.2 Linear predictive features

The results for detecting discontinuities with LPC-based distance measures are presented in Table 4.2. LSF-based measures are found to be the most successful form of LPC features and scored the highest AUC values. The choice of distance was found to have little impact on the results. Similar to Fourier-based measures, log spectra were found to outperform standard power spectra.

Features	l_1	l_2	Cos	D_{skl}
LSF Linear	0.751	0.742	0.74	0.767
$LPCC$	0.753	0.75	0.74	-
$LogPS(LPC)$	0.745	0.745	0.752	-
$PS(LPC)$	0.748	0.683	0.722	0.744

Table 4.2: Results for each of the LPC-based feature sets and distance measure combination, the table entries indicate the AUC.

4.3.3 Perceptual features

The results for the MFCC and PLP measures are presented in Table 4.3. Results for MFCCs outperform the measures for PLP. MFCCs computed from DFT power spectra were found to outperform LPC-based MFCCs. Similar to LPC, LSF measures were found to produce the best results of all possible parametric forms of the PLP model. Mel-scale warping was found to improve the performance of LSF measures when compared to standard linear LSFs. MFCCs scored an AUC value of 0.776 when used with the l_2 distance. This is the highest AUC value for perceptual features in this section and is the highest AUC value for all candidate measures tested in this chapter.

Features	l_1	l_2	Cos	D_{skl}
$MFCC(DFT)$	0.77	0.776	0.742	-
LSF Mel	0.767	0.769	0.766	0.773
$MFCC(LPC)$	0.757	0.766	0.734	-
$PLPLSF$	0.745	0.749	0.745	0.766
$PLPCC$	0.738	0.743	0.722	-
$PLPPS$	0.713	0.695	0.694	0.719

Table 4.3: Results for each of the perceptually-based feature sets and distance measure combination, the table entries indicate the AUC.

4.4 Results analysis

In this section the results are analysed to gain a deeper insight into the advantages and disadvantages of each measure, to identify sources of variation in the results and the limitations of each measure. This section is organised into three subsections covering: comparison of spectral measures, phonetic breakdown of results and time-frequency resolution limitations of the spectral measures.

4.4.1 Comparison of spectral measures

A summary of all standard spectral measures tested in this chapter are contained in Table 4.4. The measures are arranged according to the highest AUC value recorded for each measure. The highest AUC value was found to be 0.776 for MFCCs computed from DFT spectra and the second highest was LPC-based LSFs on a Mel-scale that scored an AUC value of 0.773. The lowest AUC value was computed for the harmonic power spectrum and had an AUC value of 0.642. A plot indicating the standard error for the range of AUC values in Table 4.4 is given in Fig 4.3. The standard error for the best performing measure is 0.0188. Many of the measures tested are statistically equivalent in that they have AUC values that differ by less then the standard error. For example, a typical measure from the results presented with an AUC value of 0.75 has a standard error of 0.0195. Adding and subtracting the standard error from 0.75 gives an AUC range of 0.7305 to 0.7695. Measures within this range do not offer any significant difference in performance. Almost all measures tested occupy this band of AUC values with the exception of some of the

best and worst measures. Many of the measures presented are statistically equivalent. A comparison of the ROC curves for MFCCs, LSFs on a Mel-scale, DFT log power spectra and the LPC cepstrum are illustrated in Fig. 4.4, the Euclidean distance was the measure used to produce these ROC curves.

Features	l_1	l_2	Cos	D_{skl}
$MFCC(DFT)$	0.77	0.776	0.742	-
LSF Mel	0.767	0.769	0.766	0.773
LSF Linear	0.751	0.742	0.74	0.767
$MFCC(LPC)$	0.757	0.766	0.734	-
$PLPLSF$	0.745	0.749	0.745	0.766
$LogPS(DFT)$	0.758	0.754	0.756	-
$LPCC$	0.753	0.75	0.74	-
$PS(DFT)$	0.752	0.71	0.732	0.751
$LPCPS$	0.748	0.683	0.722	0.744
$LogPS(LPC)$	0.745	0.745	0.752	-
$PLPCC$	0.738	0.743	0.722	-
$PLPPS$	0.713	0.695	0.694	0.719
PS_h	0.642	0.643	0.678	0.665
$LogPS_h$	0.699	0.697	0.674	-

Table 4.4: Summary of results for all spectral distance measures tested, the table entries indicate the AUC.

The ROC curves are all quite similar; they score perfectly in the upper right corner which corresponds with short distances with the performance dropping off as the spectral distance increases. High detection rates at short distances result in the plateau that can be observed in the upper right hand corner of each of the ROC curves. This is due to the successful detection of natural speech in the database. of MFCCs outperform the other feature sets in the region approximately halfway through the curve, from false alarm rates of 0.2 to 0.5, which corresponds with the location of the optimum threshold for separating continuous and discontinuous joins. Although MFCCs are the best performing measure with the highest AUC value they do not significantly outperform many of the other measures. Measures based on DFT, LPC and MFCCs all perform similarly. PLP measures and the harmonic spectrum are less successful.

Despite the fact that no feature set emerges as an obvious and outright measure to

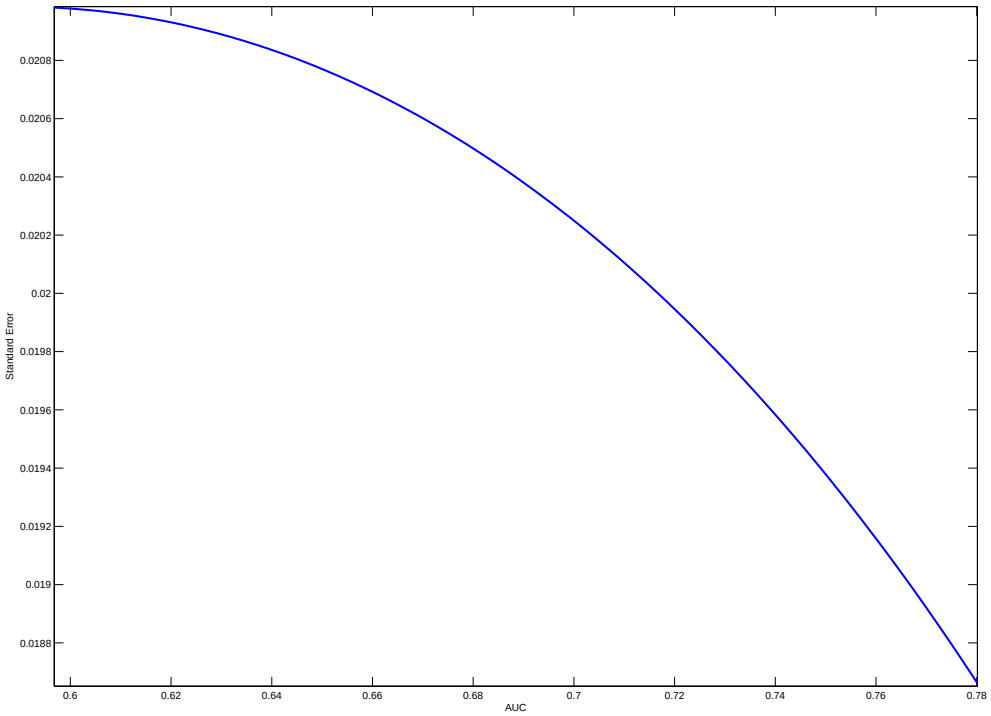

Figure 4.3: Plot indicating the standard error for the range of AUC values computed for standard spectral measures.

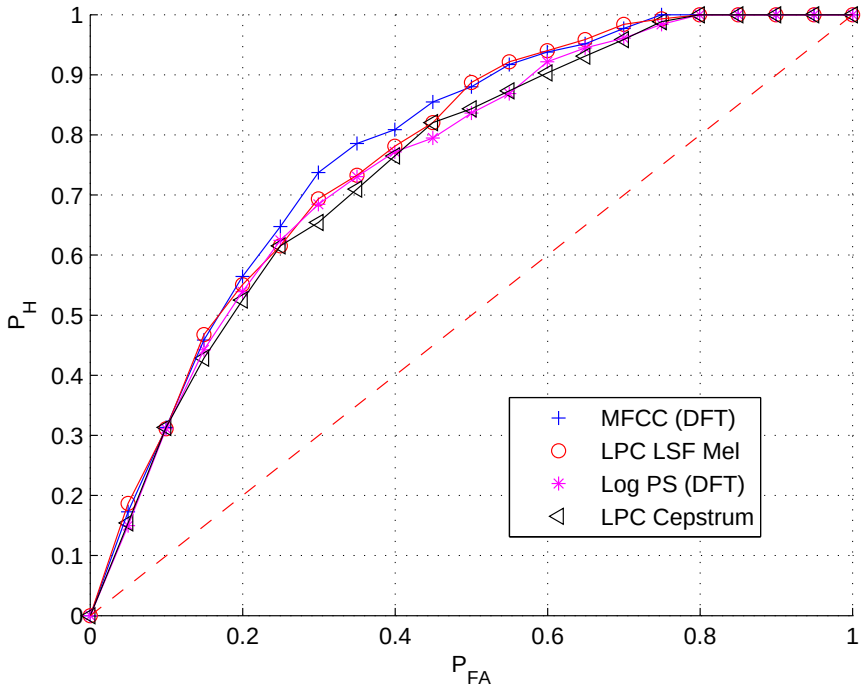

Figure 4.4: Comparing the ROC curves for selected feature sets employing the l_2 distance.

detect spectral discontinuities a number of useful observations can be made based on the results presented.

- Perceptual modelling improves performance through the application of amplitude compression and non-linear frequency scales. For example, DFT power spectra scores an AUC of 0.71 (with Euclidean distance) and after applying perceptual modelling to get MFCCs from the DFT spectra the AUC value increases to 0.776 (with Euclidean distance). The subsequent perceptual modelling significantly improves performance with respect to the standard error value. Mel-scale LSFs also outperform LSFs on a linear frequency scale.

- Certain parametric forms are more suitable then others. Regardless of the information contained within a feature set it needs to be represented in a form suitable for pattern recognition tasks. For example, LSFs outperform all other parametric forms for linear predictive based models. This is the case for LPC and PLP representations. MFCCs are specifically designed to be suitable for pattern recognition tasks as the final features output are decorrelated by the DCT which is an advantage of this representation.

- The choice of distance measure does not significantly influence the AUC value for a particular feature set.

Perceptual modelling has proved useful on two fronts: non-linear amplitude compression of the spectral estimates and frequency scale warping. Log amplitude power spectra outperform basic power spectra for both DFT and LPC computed power spectra and Mel-scale LSFs outperform linear frequency-scale LSFs. A comparison of DFT power spectra and log power spectra extracted from two concatenated units judged to contain a discontinuity is presented in Fig. 4.5. The degree of mismatch is illustrated by the error vectors produced by subtracting the right unit feature vector from the left unit feature vector. It can be observed that mismatch for power spectra is concentrated in the spectral peaks, this is not the case for log power spectra in which the mismatch can be seen across the entire frequency range. This suggests that the perceived degree of mismatch is relative

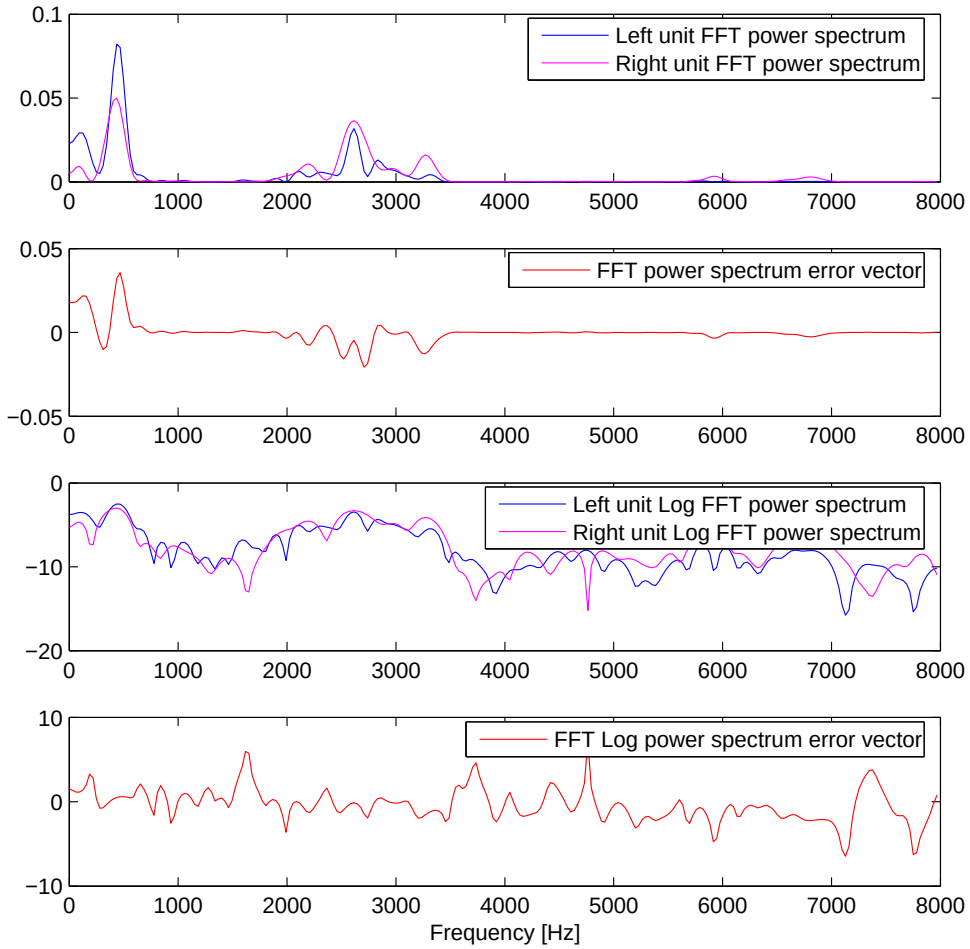

Figure 4.5: Comparing the difference between log DFT power spectra and DFT power spectra as a measure of spectral difference.

to the strength of the underlying component. The log amplitude operation mimics the instantaneous amplitude compression of the human auditory system which is a key stage in human auditory perception that facilitates the human ear in covering a large dynamic range. This may suggest that the non-linear dynamic range compression process in the auditory system is important for detecting spectral discontinuities.

LSFs are found to increase in performance when the Mel-scale is employed. The impact of an LSF mismatch of 500 Hz at a low frequency and for a high frequency is illustrated in Fig. 4.6. The LSF mismatch in the high frequency range will contribute relatively less than the equivalent mismatch in a lower frequency range due to the compression effect of the transformation. These points suggest that a feature set for detecting discontinuities

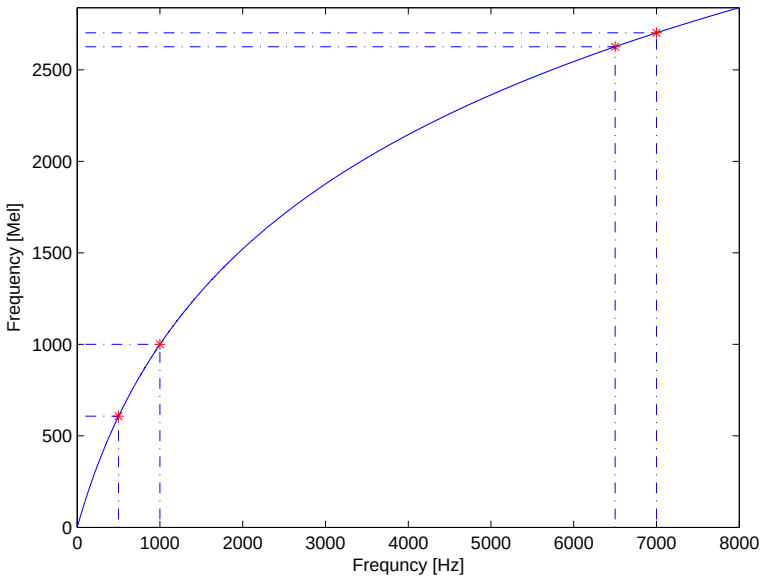

Figure 4.6: Illustrating the effect of using the Mel-scale for an equivalent frequency mismatch for both high and low frequency ranges.

should apply perceptual modelling in the spectral estimation stage and contain further processing to render the parameters into a suitable form.

4.4.2 Phonetic breakdown of results

The AUC values for each of the vowels containing joins in the test database are illustrated in Figures 4.7, 4.8 and 4.9. Fig 4.7 corresponds with the AUC values for MFCCs for each vowel, Fig 4.8 corresponds with the AUC values for LSFs on a Mel-scale for each vowel and Fig 4.9 corresponds with the AUC values for LPC cepstrum for each vowel. The l_2 distance was used to compute the AUC values for all figures. From these three figures it is quite clear that the results follow similar trends for each feature set. The trends exhibited for all the feature sets tested are found to be relatively similar to those illustrated in Figures 4.7, 4.8 and 4.9. Similarly, the bar charts illustrating the AUC values for each vowel showed no significant change when a different distance was employed, e.g. Euclidean or *Cos* for a given set.

The vowel with the highest AUC value for all of the feature sets tested was consistently found to be /ao/, this is the vowel with the smallest variance in F_1 and F_2. With MFCCs the AUC value for /ao/ was 0.95. The second highest AUC value for most feature sets

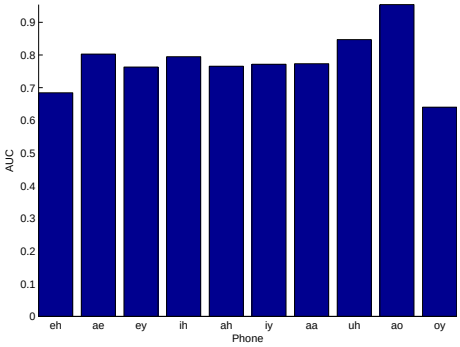

Figure 4.7: The AUC value for each vowel with MFCCs.

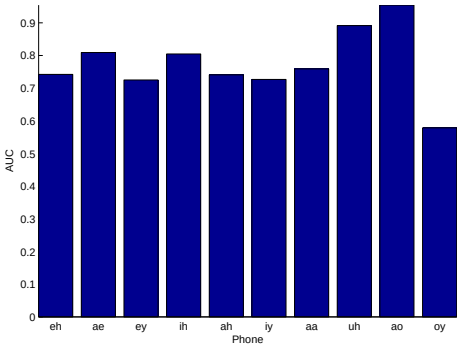

Figure 4.8: The AUC value for each vowel with LSFs on a Mel-scale.

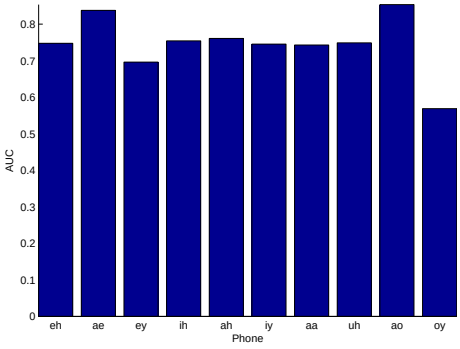

Figure 4.9: The AUC value for each vowel with LPC cepstrum.

was found to be /uh/. The vowel /uh/ was found to have the third lowest variance in $F1$, although the variance in $F2$ was found to be relatively high. The lowest AUC value for all feature sets was consistently found to be in the diphthong /oy/ and is significantly below the overall AUC value for the whole database. For example the AUC value for MFCCs across all vowels is 0.776. But for the vowel /oy/ the AUC value is 0.64. Diphthongs have been reported to be problematic in previous studies (Syrdal, 2001). No distinct pattern emerges for the other vowels, they score AUC values similar to the overall AUC value across all vowels for the given feature set.

These results indicate that each distance measure fails most significantly for the same vowel, the diphthong /oy/ and performs best for the same vowel /ao/. The success of the measures appears to be related to the amount of spectral variation in the vowel. Stationary vowels that exhibit little variation between utterances result in high detection rates, for example /ao/ has an AUC of 0.95 and vowels with significant spectral variation and dynamics result in low detection rates, for example /oy/ has a AUC value of 0.64. This breakdown of results, on a vowel by vowel basis, suggests that no advantage would be gained by combining feature sets as they fail and succeed for the same cases. Other studies have suggested that different measures are suitable for different vowels and have combined measures to increase the overall performance. The results also suggest that perhaps an alternative approach should be adopted to detect discontinuities in diphthongs or indeed any vowel that exhibits significant spectral change.

4.4.3 Time-frequency resolution

Window length was found to have a significant impact on the results. The size of the window chosen for feature extraction was found to be the largest source of variation in the results. The AUC is plotted against window length for the l_2 distance between selected feature sets in Fig. 4.10. This indicates the importance of selecting the appropriate window length. It can be observed that shorter window lengths are preferred, this demonstrates that temporal resolution is significantly important. For non-pitch synchronous windows the maximum AUC is achieved with a window length of 10 ms across all feature sets,

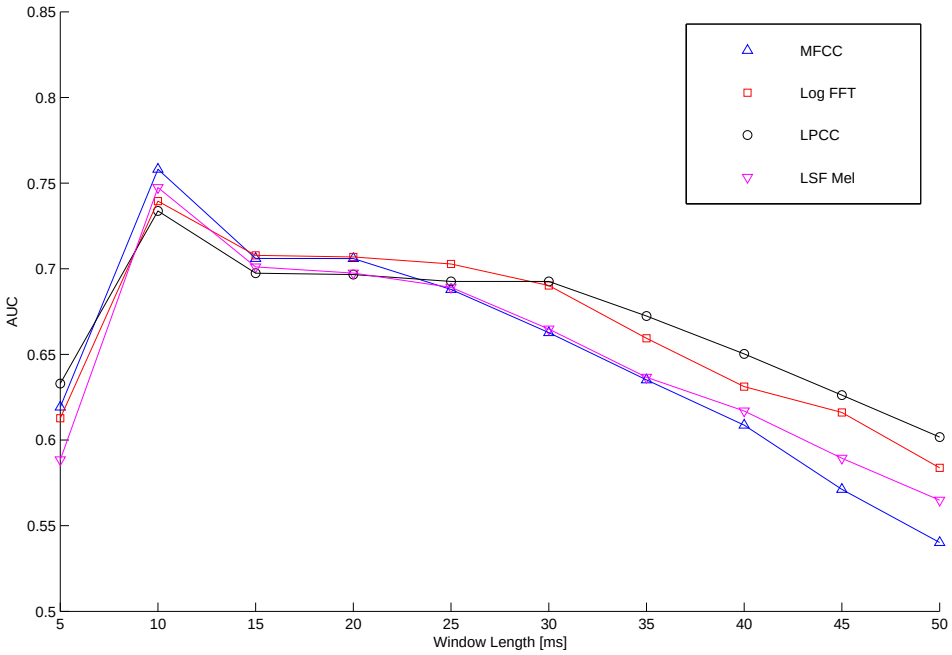

Figure 4.10: The AUC for varying window length for selected feature sets with the l_2 distance.

this corresponds approximately with the average pitch period for the database. The AUC value decreases rapidly when the window length is smaller than a pitch period in length. This is the window length at which the DFT can no longer resolve individual harmonic components and appears to be the lower threshold for frequency resolution at which point the AUC value decreases significantly. The optimum windowing strategy for all feature sets was found to be pitch synchronous windowing with a window length of one pitch period.

The choice of window length relates to the fundamental trade-off between time and frequency resolution. All of the feature sets considered are limited by this constraint in the same way, in that they are extracted from a fixed window length of data. In the previous section it was identified that diphthongs are most problematic as they contain significant spectral dynamics compared with monophthongs. If the spectral values change significantly within the time-span of the analysis window then the accuracy of the spectral estimate will be degraded. This is the case for diphthongs and as such the time resolution of the spectral estimation is particularly important.

4.5 Conclusion

In this chapter a number of standard spectral feature sets were tested for the task of detecting spectral discontinuities in the test database. The results were analysed in order to determine sources of variation in the results, the advantages and disadvantages of each feature extraction technique, to identify limitations of the features and to identify instances in which the measures perform particularly badly. A number of useful findings were noted.

- Perceptual modelling tends to improve the performance of a feature set. This is evident for amplitude compression and non-linear frequency scales.
- Feature sets such as MFCCs and LSFs with good properties for pattern recognition tasks produce better results.
- The choice of window length was found to be the most significant source of variation in the results. The change in performance that resulted from changing the window length was more significant than the difference in performance between feature sets.
- The candidate measures of spectral continuity performed best for the vowel with the least spectral variation across the database. Similarly, the lowest scores were for the diphthong /oy/ and vowels that exhibited the most spectral variation.
- The time frequency resolution trade-off for spectral estimation appears to be a limiting factor for all of the measures tested. The performance of the measures degrades significantly for vowels that exhibit spectral dynamics. These vowels require short windows to achieve a good time resolution of the time varying spectrum. Selecting a window shorter then a pitch period degrades the estimate as individual harmonics cannot be resolved.

The feature sets tested were found to have similar performance, with many of the measures statistically equivalent to an AUC measure of 0.75 when the standard error is considered. Perceptual modelling and feature sets suitable to pattern recognition were found to consistently, albeit marginally, perform better at the task of detecting disconti-

nuities. The results are largely in agreement with those reported in Stylianou and Syrdal (2001) and Wouters and Macon (1998).

It was found that the details of feature extraction, specifically the window size employed, had a significant impact on the results. Notably the choice of window size was found to have a more significant impact on the results than the choice of feature set or distance measure employed. The optimum choice of window was found to be a single pitch period window extracted pitch synchronously. This suggests a trade-off in the importance between time and frequency resolution when extracting features to predict spectral discontinuities.

A phonetic breakdown of the results revealed that most features scored similar AUC values when the results where computed on a vowel-by-vowel basis. For example, all features sets had the highest AUC value for the joins contained in the vowel /ao/ and the lowest AUC value for joins contained in the diphthong /oy/. This suggests that no particular advantage can be gained by combining feature sets to detect discontinuities. The results also suggest that feature sets that are suitable for monophthongs may not be as suitable for diphthongs. Diphthongs contain structured spectral dynamics with potentially significant variations in formant values, ideally this should be accounted for in the feature extraction. Time-frequency resolution is a particularly important factor to consider in a signal with a rapidly varying spectrum. The importance of window length and the low AUC values obtained for diphthongs strongly suggests that spectral dynamics and time-frequency resolution for spectral estimates are issues that need to be accounted for in the task of detecting discontinuities. Time-frequency resolution and spectral dynamic features are explored in future chapters.

Chapter 5

Alternative Spectral Features

Studies to date investigating spectral join costs have largely focused on the use of standard spectral representations which have found success in other areas of speech processing. Most of these measures are derived from the magnitude spectrum of a fixed length window of speech. This chapter, building on the results of previous chapters, aims to go beyond the standard spectral representations and address some of their fundamental limitations. This is approached in two ways: firstly by investigating the phase spectrum as a source of discontinuity and secondly through an investigation of wavelet-based spectral measures. No previous study has examined the relevance of phase spectra or the potential of wavelets in capturing spectral continuity in concatenative synthesis. Standard spectral representations of speech are computed from the magnitude spectrum. The investigation of the phase spectrum aims to explore the role of phase as a measure of spectral continuity. The investigation of wavelets as a spectral measure builds directly on results from Chapter 4 in which it was identified that time-frequency resolution limited the performance of the measures and that the time varying spectrum in diphthongs had a negative impact on the performance of the standard measures. Wavelet analysis enables a framework for spectral estimation in which the time-frequency resolution of the spectral estimates varies with frequency.

This chapter has two distinct components. The first component contains the details of the investigation of phase based measures of spectral continuity and the second compo-

nent contains the details of the study of the wavelet transform as a measure of spectral continuity.

5.1 Phase spectra

In this section phase-based distance measures are investigated as measures of spectral continuity. The Fourier spectrum, $X(\omega)$ (equation 5.1), can be expressed in terms of its magnitude, $|X\omega)|$ (equation 5.2), and phase spectrum, $\angle X(\omega)$ (equation 5.3). Traditionally in speech processing tasks the phase component of the spectrum is discarded and the magnitude component of the spectrum is retained and employed for further processing.

$$X(\omega) = |X\omega)|\, e^{j\angle X(\omega)} \tag{5.1}$$

$$|X\omega)| = (X_r(\omega)^2 + X_i(\omega)^2)^{1/2} \tag{5.2}$$

$$\angle X(\omega) = tan^{-1}\left(\frac{X_r}{X_i}\right) \tag{5.3}$$

Most distance measures reported to date are computed from the magnitude spectrum and discard the phase spectrum during feature extraction, furthermore no other studies have been reported that specifically incorporate the phase spectrum in a join cost. For example many spectral representations explicitly extract the magnitude spectrum, such as MFCCs. More indirectly, LPC representations remove the phase spectrum as they are computed from the auto-correlation or covariance sequence, the computation of which is a lossy process and removes phase information of the original speech signal. Although the LPC model derived does contain a phase spectrum it is not the phase spectrum of the original speech signal.

Many problems exist with respect to reliably extracting the phase spectrum in a manner that is robust to noise and suitable for pattern recognition. In this dissertation the phase spectrum was investigated in the form of the group delay function, $\tau(\omega)$, which is defined as the negative of the derivative of the phase spectrum with respect to frequency, given

by equation 5.4.

$$\tau(\omega) = -\frac{d}{d\omega}\{\angle X(\omega)\} \tag{5.4}$$

The raw phase spectrum is unsuitable to use as a distance measure or indeed for pattern recognition tasks in general due to the modular nature of phase, equation 5.5. As a result the phase spectrum for a given signal is not uniquely defined.

$$\angle X(\omega)|_{\omega=\alpha} = \theta_\alpha + n\ 2\pi \tag{5.5}$$

θ_α represents the phase angle limited between 0 and 2π and n is any integer value. The problem of multiple phase values is side-stepped by adopting the group delay function which can be unambiguously defined and is directly dependent on the phase spectrum (Yegnanarayana and Murthy, 1992).

5.1.1 Perception of phase

Historically the magnitude spectrum has been considered to be the most perceptually important component of the spectrum and as such most feature sets used in practice are derived from the magnitude spectrum. Phase up until recently has been largely ignored. A number of reasons contribute to the dominance of magnitude-based features over phase-based features; perceptually the phase spectrum is not fully understood and is difficult to process in comparison with the magnitude spectrum. Historically, phase was believed to play little or no role in speech perception. In Ohm (1843) it is reported that the human auditory system is essentially phase-deaf and only requires the magnitude spectrum to successfully process and interpret speech. A similar result was found by von Helmholtz (1875) who experimented with changing the phase spectrum for stimuli whilst holding the magnitude spectra constant and found that the stimuli contained no audible difference when the phase spectra changed.

Many subsequent studies have found that the phase spectrum is perceptually important in the human auditory system (Patterson, 1987) and in particular makes a significant contribution to the naturalness and quality of speech. The neural encoding stage within the

auditory system is reported to encode speech information directly from timing information in a way that reflects formant structure (Young and Sachs, 1979). The temporal envelope of the signal within an auditory channel determines when an auditory nerve fires and is dependent on the phase of the individual signal components within that auditory channel (Lindemann and Kates, 1999). Although these studies report useful models and offer results that promote a deeper understanding of phase perception they do not offer a practical model of phase perception that is readily usable in the context of speech processing.

Applications that require high quality speech to be synthesised have consistently identified the phase spectrum as a key limitation in achieving high quality natural sounding speech. Many phase models have been proposed in the speech coding literature. This has been particularly prevalent in sinusoidal transform coding (McAulay and Quatieri, 1992) which employs the sinusoidal model (McAulay and Quatieri, 1986) to compress the speech waveform. Varying degrees of speech quality can be achieved depending on the complexity of the phase model employed (McAulay and Quatieri, 1995; Almeida and Tribolet, 1983).

Although previous studies in the speech coding literature have indicated the importance of phase for speech quality and naturalness more recent studies have explicitly demonstrated the perceptual importance of phase, (Paliwal and Alsteris, 2005; Alsteris and Paliwal, 2006; Shi et al., 2005). The experiments conducted in these studies demonstrate that speech can be successfully interpreted by humans with only the phase spectrum used to reconstruct the speech signal, with all magnitude information discarded. This clearly indicates that the phase spectrum plays a significant role in speech perception and can be used in the identification of phonemes.

5.1.2 Objectives of the phase investigation

In this study the objective is to identify the role of the phase spectrum in concatenative speech synthesis, to identify if phase difference is a source of discontinuity and to determine a suitable feature set to measure the degree of phase mismatch. The objectives are:

- To explore spectral features derived from the phase spectrum in spectral join costs.

- To develop a feature extraction procedure suitable for phase-derived measures.
- To quantify and compare the performance of a selection of phase-derived measures for the task of detecting discontinuities in the test database.
- To identify the contribution of phase-based measures when used in conjunction with magnitude measures.

5.2 The group delay function and feature extraction

As stated earlier, the raw phase spectrum is unsuitable as a feature set and in this thesis the group delay function, equation 5.6, is used to represent the phase spectrum.

$$\tau(\omega) = -\frac{d}{d\omega}\{\angle X(\omega)\} \tag{5.6}$$

One of the key properties of the group delay function in comparison with the magnitude-based measures is that spectral peaks are considerably narrower when compared with magnitude spectra, enabling closely spaced formants to be resolved. Figure 5.1 compares a spectrogram computed from the DFT magnitude spectrum with one computed from the group delay function for a segment of speech. The same formant tracks are clearly evident in both spectrograms although they each have distinct characteristics.

Accurately estimating the group delay function is sensitive to noise, window shape and window length (Bozkurt et al., 2004). These parameters can be related to the zeros of the z-transform of the signal (Yegnanarayana and Murthy, 1992; Bozkurt et al., 2004). Periodicity in the signal corresponds with zeros located close to the unit circle, as a zero on the unit circle corresponds with a pure oscillation at the corresponding angular frequency. The proximity of zeros to the unit circle significantly increases the sensitivity of the group delay function to noise (Yegnanarayana and Murthy, 1992; Bozkurt et al., 2004). Thus the reliability of an estimate is significantly lowered when it is in the proximity of a zero and is undefined when the sampled frequency coincides exactly with a zero on the unit circle (Yegnanarayana and Murthy, 1992; Bozkurt et al., 2004). Bozkurt et al. have also reported

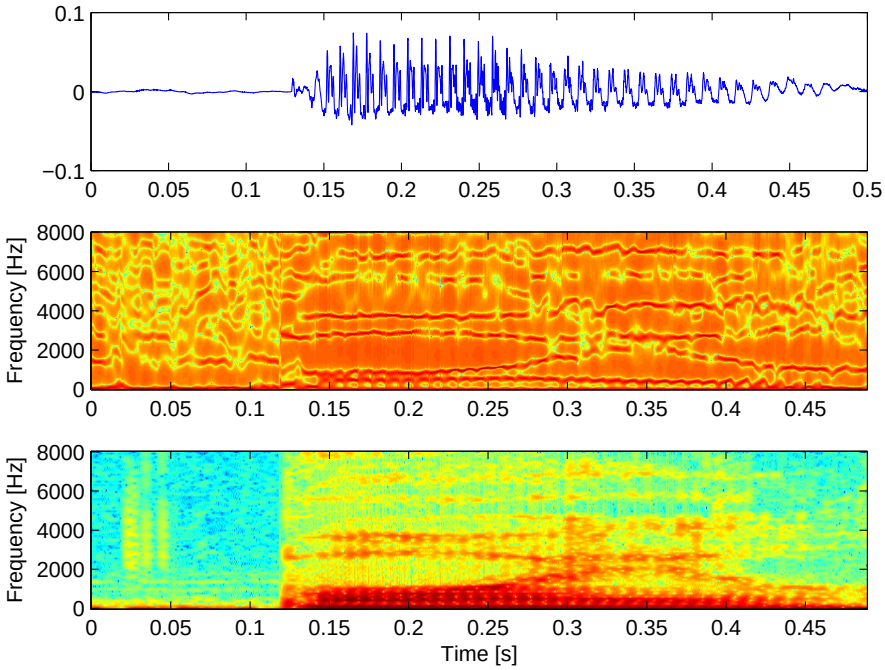

Figure 5.1: Comparison of group delay function (middle) and magnitude-based spectrograms (bottom) for a segment of speech (top).

that short window lengths and certain window shapes are also found to be more successful in moving zeros away from the unit circle which results in more reliable estimates of the group delay function.

For the task of detecting discontinuities in the test database Hamming windows were found to produce the best overall results. Pitch synchronous windowing with an analysis window of one pitch period in length was found to be the optimum windowing strategy, as with the standard spectral measures in the previous chapter. No pre-emphasis filter was employed for estimating the group delay function as its purpose is to pre-emphasise the peaks of the magnitude spectrum and this is not appropriate for estimating the group delay function. A number of methods for computing the $\tau(\omega)$ were investigated:

- Group delay function computed from DFT spectra (GDF DFT).
- Group delay function computed from LPC spectra (GDF LPC).
- Computation using the group delay function formula (GDF) (Oppenheim and Schafer, 1975; Yegnanarayana and Murthy, 1992).

- The modified group delay function (MODGDF) (Murthy and Gadde, 2003).

5.2.1 DFT group delay function

The DFT based computation of the group delay function employed a 512-point FFT. The phase spectrum was computed from the FFT coefficients and subsequently unwrapped. The unwrapped phase spectrum was numerically differentiated by computing the difference between successive phase values.

5.2.2 LPC group delay function

The GDF LPC was computed employing an LPC analysis using the autocorrelation technique. The LPC model was of order 16. The phase spectrum of the LPC model was computed, unwrapped and differentiated to produce the estimate of the group delay function. Phase unwrapping refers to the process of producing a continuous function from a set of phase values in a closed interval, such as from $-\pi$ to π.

5.2.3 Direct group delay function formula

The GDF was implemented as in Yegnanarayana and Murthy (1992) using the formula, equation 5.7.

$$\tau(\omega) = \frac{X_r(\omega)Y_r(\omega) + X_i(\omega)Y_i(\omega)}{|X(\omega)|^2} \tag{5.7}$$

$X(\omega)$ and $Y(\omega)$ denote the Fourier transform of $x(n)$ and $nx(n)$ respectively and the subscripts r and i denote real and imaginary components of the Fourier coefficients. This formula allows the group delay function to be computed without requiring phase unwrapping. The values of the group delay function are uniquely defined for a given signal, and as such, problems associated with phase ambiguity are removed.

5.2.4 Modified group delay function

The modified group delay function estimate (MODGDF) was implemented as in Murthy and Gadde (2003). The computation of the MODGDF is based on the formula for the

GDF 5.7, although a number of modifications are employed with the objective of moving the zeros of the z-transform further from the unit circle and in turn improving the spectral estimate (Murthy and Gadde, 2003). The equations to implement the MODGDF are as follows:

$$\tilde{\tau}_{\gamma}(\omega) = \frac{X_r(\omega)Y_r(\omega) + X_i(\omega)Y_i(\omega)}{S(\omega)^{2\gamma}} \tag{5.8}$$

$$\tilde{\tau}_{\alpha,\gamma}(\omega) = \frac{\tilde{\tau}_{\gamma}(\omega)}{|\ \tilde{\tau}_{\gamma}(\omega)|}\,|\tilde{\tau}_{\gamma}(\omega)|^{\alpha} \tag{5.9}$$

In equation 5.8 $S(\omega)$ is a cepstrally smoothed version of $|X\,(\omega)|$. The initial group delay estimate, $\tilde{\tau}_{\gamma}(\omega)$, depends on the parameter γ that is used to smooth the estimate. This initial estimate,$\tilde{\tau}_{\gamma}(\omega)$ is used in equation 5.9 to compute the MODGDF. Another parameter, α is introduced in equation 5.9 to further reduce the spikey nature of the group delay function estimate. In this thesis the values for α and γ were set as 0.4 and 0.9 respectively. These parameters were selected based on values reported to produce reliable estimates of the modified group delay function in Murthy and Gadde (2003).

5.3 Results - group delay function

This section contains the discontinuity detection results with the proposed group delay function estimates. The results represent the AUC value computed from the ROC curves and are presented for each phase-based feature set and distance measure combination in Table 5.1. The LPC-based GDF was found to have the highest AUC value when used in conjunction with the absolute distance as indicated in Table 5.1, with an AUC value of 0.7612. Both the MODGDF and GDF DFT were also found to perform well when used in conjunction with the l_1 and l_2 distances, with AUC values in the 0.73 to 0.75 range. This indicates that phase-based measures do correlate with human perception of discontinuity, although the level of correlation is less then that of many of the standard measures tested in Chapter 4. A comparison of the ROC curves generated for the phase-based measures alongside the MFCC-based measure is illustrated in Fig. 5.2. All of the

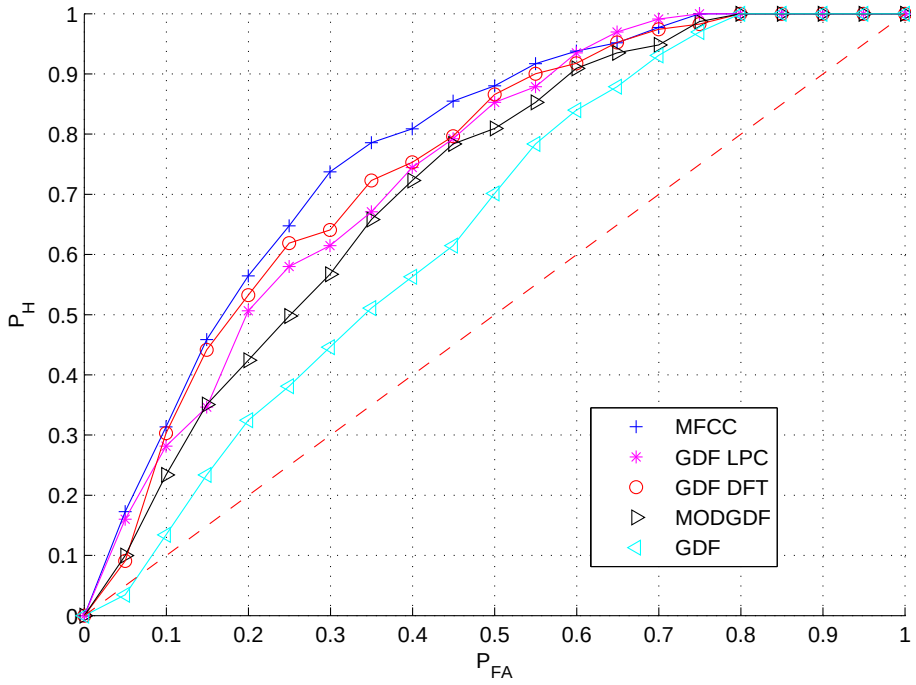

Figure 5.2: Comparion of the ROC curves for each of the phase based measures and the standard MFCC measure.

ROC curves generated from phase based measures employ the l_1 distance, the MFCC ROC curve was generated using the l_2 distance. The LPC-based GDF with the l_1 distance is the only phase based measure to lie within the standard error bands of the MFCC AUC value of 0.776.

Features	l_1	l_2	Cos
GDF LPC	0.7612	0.7305	0.732
GDF DFT	0.75	0.7474	0.6651
MODGDF	0.734	0.7334	0.7169
GDF	0.5902	0.6074	0.584

Table 5.1: Results for GDF-based measures computed from the phase spectrum, the table entries indicate the AUC.

5.3.1 Analysis of results

The results for the phase-based feature sets were analysed with respect to significant sources of variation. Window length and the type of window are known to have a significant influence on group delay estimation. The results indicating the AUC value for each different

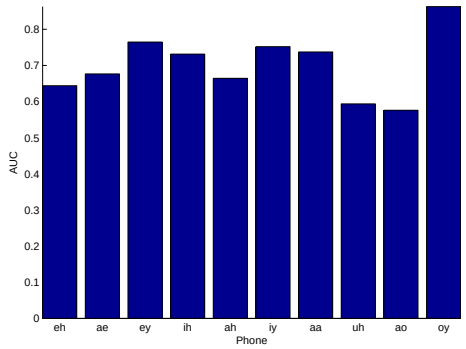

Figure 5.3: The AUC value for each vowel with the DFT-based GDF.

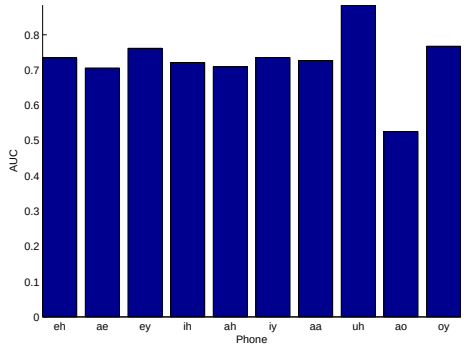

Figure 5.4: The AUC value for each vowel with the LPC-based GDF.

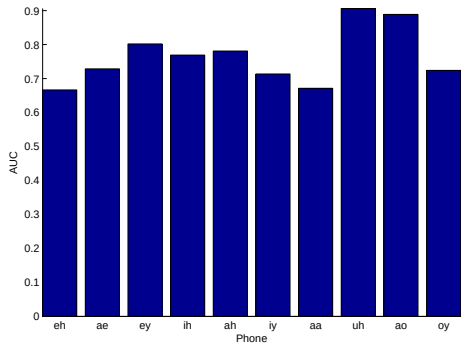

Figure 5.5: The AUC value for each vowel with MODGDF.

Features	Hanning	Hamming	Blackman	Rectangle
GDF LPC	0.7242	0.7612	0.7266	0.7415
MODGDF	0.732	0.734	0.7134	0.714
GDF DFT	0.672	0.75	0.6808	0.7414
GDF	0.6392	0.5902	0.6513	0.6427

Table 5.2: Results for GDF based measures computed using different window functions, the table entries indicate the AUC computed using the absolute distance.

window type considered are given in Table 5.2. The choice of window shape can change the performance of a measure such that it is statistically a better measure with respect to the standard error for AUC values. The highest AUC value obtained was with the Hamming window. The Hamming window scored the highest AUC value for all of the feature sets considered with the exception of the GDF computed from equation 5.7. The variation in the AUC value for varying window length was found to be similar to the results found for the standard spectral measures, with a peak at approximately 10 ms. Pitch synchronous analysis with a window length of one pitch period was found to yield the highest AUC values.

The AUC values for each vowel are illustrated in Fig. 5.3, Fig. 5.4 and Fig. 5.5 for DFT GDF, LPC GDF and the MODGDF respectively. The most notable observation when comparing the AUC values for each individual phone with the standard spectral measures is that the relative performance of the group delay measures with the phone /oy/ is higher then for the standard measures. The standard measures all scored the lowest detection rates for /oy/. For the both the GDF and the GDF DFT, the AUC value for /oy/ is higher then for all other phones tested. The group delay features, with the exception of the LPC GDF, outperform standard measures for the diphthong /oy/ with AUC values of over 0.72. The GDF DFT scored a particularly high AUC value of 0.8746 for the phone /oy/. The LPC GDF scored an AUC value of 0.6087 which is more consistent with the AUC values computed for /oy/ with the standard magnitude based representations. The group delay function of the LPC GDF represents the group delay function of the LPC model as opposed to the signal components and is different to the other group delay measures. The LPC GDF offers no advantage when considered as a join cost that may be used in conjunction

with standard spectral measures as it exhibits similar detection rates across all phones in the database. Of the remaining group delay function measures the GDF DFT measure has the highest AUC value and has a higher detection rate for diphthongs in the test database compared with standard spectral measures.

A study was conducted to test a combined measures of phase and standard spectral features. Combining the measures was found to produce no notable gain in performance. The method of combining measures by simply concatenating feature vectors and computing the Euclidean distance between the combined feature vectors limits the extent to which complementary discriminating information is usefully exploited in the combined measure. The idea of combining measures with a more sophisticated framework is discussed further in Chapter 7.

5.4 Wavelets

In order to overcome the limitations of the standard measures which adopt an analysis procedure with a fixed window length the wavelet transform was adopted. Wavelet analysis allows the time-frequency resolution to vary with respect to frequency. This allows spectral estimation to vary with a time-frequency resolution adapted to each frequency band.

For a given wavelet mother function, ψ, the wavelet transform at scale a and position u, equation 5.10, is defined in equation 5.11 (Mallat, 1998), were ψ^* denotes the complex conjugate of ψ.

$$\psi_{u,a}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-u}{a}\right) \tag{5.10}$$

$$Wx(u,a) = \int_{-\infty}^{+\infty} x(t)\psi^*_{u,a}(t)dt \tag{5.11}$$

Many possible wavelet functions, ψ, exist (Daubechies, 1992). A number of wavelet functions were tested as potential spectral join costs using the test database. In this thesis a number of wavelets are considered: the Daubechies family of wavelets, the Symlet family of wavelets and the Coiflet family. Also considered were a number of individual wavelet types; Gaussian, Mexican-hat, Morlet and Meyer wavelets.

The wavelet transform has been suggested as a means to model the frequency selectivity of the auditory system (Yang et al., 1992; Quatieri, 2002). Wavelets can be constructed that vary in bandwidth with respect to frequency to mimic the frequency response characteristics of the auditory system (Salimpour and Abolhassani, 2006). In this section wavelets are investigated as a measure of spectral continuity and evaluated for their ability to detect discontinuities in the test database. The objectives of the investigation of the wavelet transform are as follows:

- To quantify the performance of a selection of wavelet based spectral representations in spectral join costs.
- To investigate the performance of different families of wavelets in join costs and identify which wavelets provide the best measures of spectral continuity.
- To compare the performance of wavelet-based measures with standard spectral measures.
- To identify the potential for combining of wavelet-based measures of spectral continuity in conjunction with standard spectral measures.

5.4.1 Wavelet feature extraction

A feature vector of 50 wavelet coefficients was used to represent each unit, these coefficients correspond with an analysis centred on the pitch pulse. Fig. 5.6 illustrates a wavelet based scalogram and a conventional spectrogram across a join within a concatenated test word judged as discontinuous in the perceptual experiment. The periodic nature of the scalogram illustrates the importance of employing feature vectors from the same relative position within a pitch period as the features vary significantly within a pitch period.

5.5 Results - wavelets

The results for the detection of discontinuities with a number of wavelet-based spectral estimates are presented in Tables 5.3, 5.4 , 5.5, and 5.6. The results show the AUC

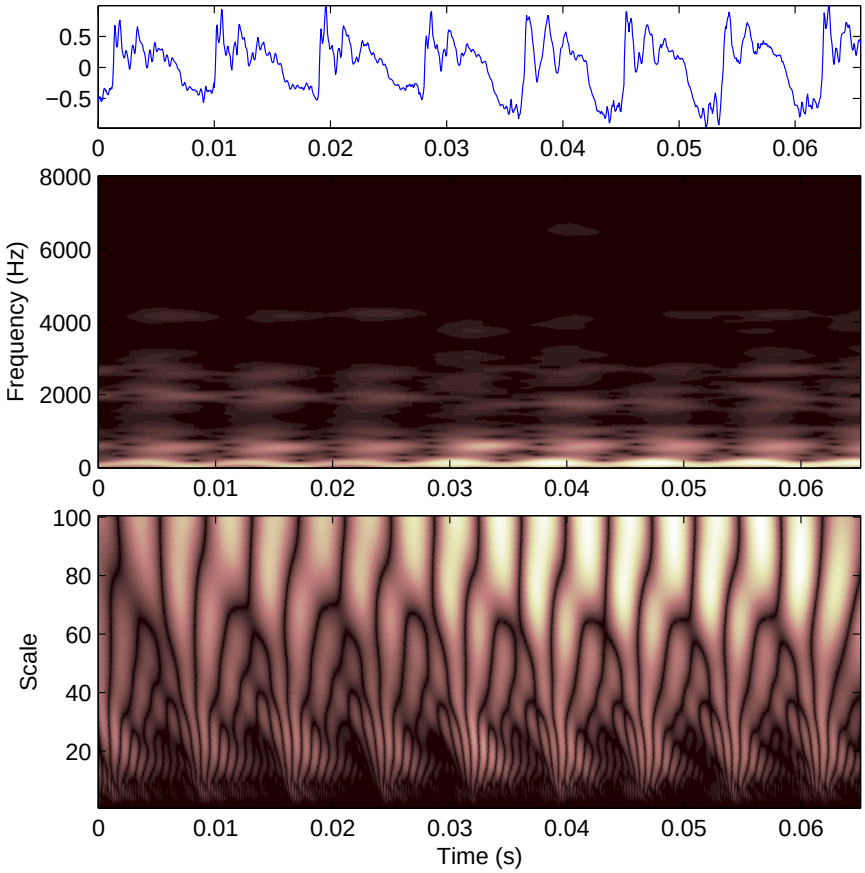

Figure 5.6: Time domain signal, spectrogram and scalogram across a join respectively from top to bottom. With a frame size of 10 ms for the spectrogram and a frame shift of one sample for both spectrogram and scalogram.

values computed for the specific wavelet type in conjunction with both the l_2 distance and the *Cos* distance. The l_1 distance was omitted from the presentation of results as it varies little from the results of the l_2 measure for all wavelets tested. The AUC values for the Daubechies, Symlets and Coiflet families are contained in Tables 5.3, 5.4 and 5.5 respectively. The results for the Mexican-hat, Meyer, Gaussian and Morlet wavelet are contained in Table 5.6.

A number of patterns emerge from the results.

- The *Cos* distance consistently outperforms the l_2 distance when used in conjunction with a wavelet-based spectral estimate.

Daubechies Wavelets	l_2	Cos
db1	0.7394	0.7576
db2	0.7719	0.8037
db3	0.77643	0.8069
db4	0.7628	0.7842
db8	0.7841	0.8046

Table 5.3: Results for measures employing Daubechies wavelets, the table entries indicate the AUC.

Symlet Wavelets	l_2	Cos
sym1	0.7394	0.7576
sym2	0.7719	0.8037
sym3	0.77643	0.8069
sym4	0.7537	0.7879
sym8	0.7608	0.8015

Table 5.4: Results for measures employing Symlet wavelets, the table entries indicate the AUC.

Coiflet Wavelets	l_2	Cos
coif1	0.7513	0.7987
coif2	0.7584	0.8022
coif3	0.7602	0.8017
coif4	0.7615	0.8024
coif5	0.7625	0.8063

Table 5.5: Results for measures employing Coiflet wavelets, the table entries indicate the AUC.

Wavelet	l_2	Cos
Meyer	0.7513	0.7987
Morlet	0.7678	0.8065
Gaussian	0.7548	0.7314
Mexican hat	0.7506	0.7455

Table 5.6: Results for spectral measures employing a selection of common wavelet functions, the table entries indicate the AUC.

- Many of the candidate wavelet measures outperform all standard measures tested in Chapter 4.
- The results obtained across all tested wavelets are similar and occupy a narrow numerical range.
- The first member of each wavelet family does not perform as well as the wavelets of order 2 or higher.

Many of the wavelet based measures tested outperform the best ranking standard spectral measures. The highest AUC value obtained employing standard spectral measures was 0.776 using MFCCs and the Euclidean distance. The highest AUC value for wavelets was 0.8069 (standard error = 0.0181) using the Daubechies 3 (db3) wavelet in conjunction with the *Cos* distance. The Daubechies 3 measure is statistically significantly better then the MFCC measure. Many of the other wavelets scored similar AUC values when used in conjunction with the *Cos* distance and lie within the standard error bands, making them statistically equivalent. The AUC values for most of the wavelets tested occupy a relatively narrow numerical range when used with the *Cos* distance, indicating that the success of the method is relatively independent of the choice of wavelet function. The majority of wavelets tested score AUC values of approximately 0.8, although this is not the case for certain wavelets. For example Daubechies 1 (db1), symlet 1 (sym1) the Gaussian and Mexican hat wavelets.

Fig. 5.7 compares ROC curves for standard MFCC, wavelet-based measures and a measure based on MFCCs extracted with a 40 ms analysis window. The MFCC measures were both computed with the l_2 distance, the wavelet measure was computed with the *Cos* distance.

5.5.1 Analysis of results

The results obtained for wavelet-based measures indicated that a number of these measures outperformed the standard spectral representations at detecting discontinuities in the test

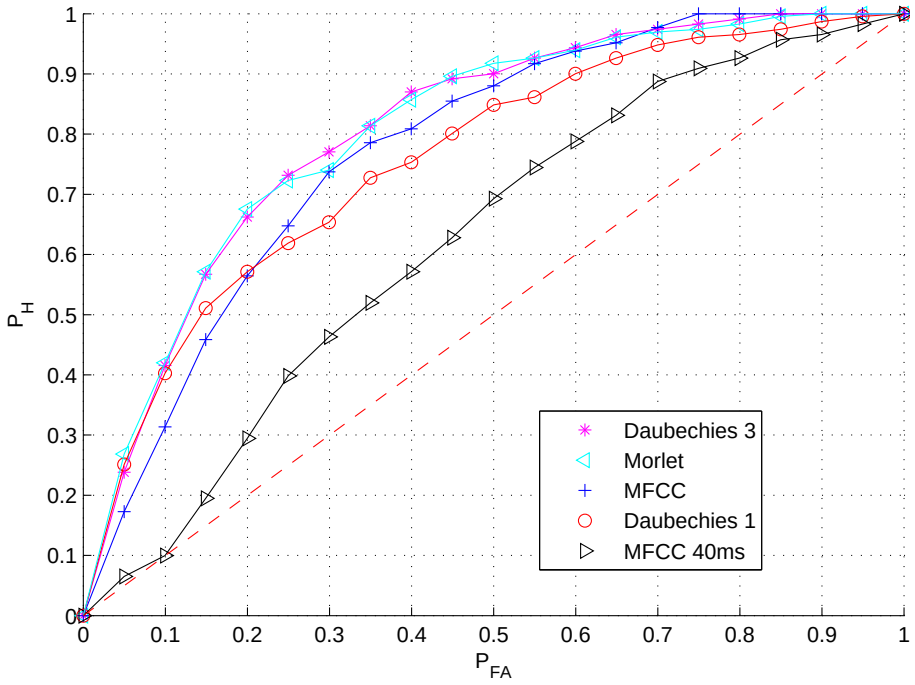

Figure 5.7: Comparison of the ROC curves from wavelet based measures and the standard MFCC based measures, with optimum windowing and with a 40 ms window.

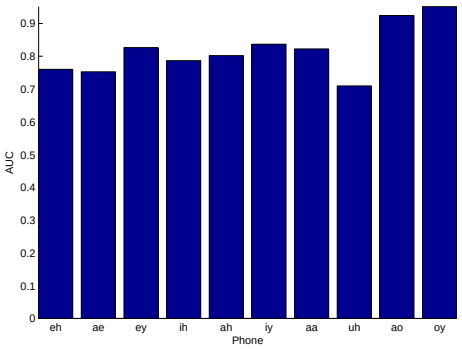

Figure 5.8: The AUC value for each individual phone class with the db8 wavelet with the *Cos* distance.

database. In this section the objective is to gain a deeper understanding of why the wavelet based measures outperform standard spectral measures.

Figures 5.8 and 5.9 illustrates a breakdown of the AUC values computed on a phone-by-phone basis for the test database for the db8 and Meyer wavelets respectively. When this set of AUC values is compared to those generated from the standard measures a key trend emerges: the standard measures consistently have the lowest AUC value for the diphthong /oy/, whereas with the wavelet based measures this is consistently the highest AUC value

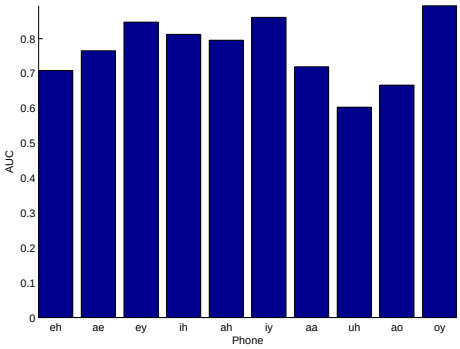

Figure 5.9: The AUC value for each individual phone class with the Meyer wavelet with the *Cos* distance.

obtained. The AUC value for the phone /oy/ with MFCCs is 0.6405, for wavelets based measures the AUC value is consistently over 0.9 and is always the phone with the highest AUC value. For other phones wavelets do not necessarily outperform the standard measures and the difference between the respective AUC values between wavelets and the standard measures is often quite small and not indicative of any particular pattern. All wavelets have the lowest AUC value for the phone /uh/. It is not clear why /uh/ consistently produces a low AUC value for wavelet-based measures.

Diphthongs can contain components that vary rapidly with time. When a signal is time-varying the time-frequency resolution can be a limiting factor in making a reliable spectral estimate of the signal. The results in Chapter 4 found that standard spectral measures performed significantly worse in vowels with non-stationary spectra. Wavelets are designed to overcome this limitation to some extent by allowing a multi-resolution analysis and are better equipped to estimate the spectral components of time varying signals, such as those found in diphthongs. The results presented in this chapter are consistent with this argument and many of the wavelet-based measures are found to outperform standard measure for diphthongs, but not necessarily for other phones with relatively stationary spectra.

5.6 Summary and conclusions

In this chapter phase and wavelet-based spectral measures were tested as measure of spectral continuity. Both were found to correlate with perceptual results. Furthermore many of the wavelet-based measures were found to outperform all standard spectral measures.

A number of phase-based spectral measures, in the form of the group delay function, were tested. The feature extraction parameters were investigated to determine the most suitable method to extract the group delay function for use in a spectral join cost. A pitch synchronous Hamming window was found to produce the best results using an LPC-derived group delay function. Analysis of the results indicated that the phase measures outperformed standard spectral measures for the diphthong /oy/ with the exception of the LPC GDF. It was not unexpected that the LPC GDF performed differently than the other group delay function measures tested as it represents the group delay function of the LPC spectrum as distinct from the original speech. Furthermore, phase information is lost in the computation of the LPC model. As a result, of the measures tested the LPC GDF is unlikely to offer any advantage when combined with standard spectral measures as it has similar detection rates in all phones to the standard spectral measures. The remaining group delay measures could offer an advantage as they have an increased performance over standard measures for diphthongs in the test database.

Motivated by the results for standard spectral measures that suggested the trade-off between time and frequency resolution was limiting performance, wavelet based spectral measures were investigated. Wavelet-based measures were found to correlate with human perceptual results and were found to outperform all of the standard spectral measures in many cases. The *Cos* distance was found to produce the best results with wavelet-based spectral measures. Wavelets were found to significantly outperform standard spectral measures for the diphthong /oy/. For most other phones the performance of wavelet measures and standard spectral measures were similar. Wavelets consistently scored low AUC values for the phone /uh/, it is unclear why they should consistently score low AUC values for this phone.

From the results presented in this chapter it can be concluded that both phase and

wavelet derived measures correlate with human perception of discontinuity. The results suggest that both measures can outperform standard measures in the case of predicting discontinuities in diphthongs.

Chapter 6

Spectral Dynamics as a Source of Discontinuity

In this chapter spectral dynamic behaviour is investigated as a potential source of discontinuity in concatenated speech. Spectral dynamics refers to changes in the spectrum of the signal with time. Spectral dynamic behaviour plays an important role in the human auditory system and its role in speech perception has been extensively studied. Many psychoacoustic and neuro-physiological phenomena have been identified from perceptual experiments relating to spectral dynamics in speech. Spectral dynamic behaviour is also important in relation to speech production as it relates to the underlying co-articulation between successive phonemes in an utterance and contains information about the neighbouring spectra adjacent to the unit boundary. Despite the known physical and perceptual importance of spectral dynamics, no dedicated study has been reported that investigates the role of spectral dynamics in concatenative speech synthesis. The objective of this chapter is thus to further understand the role of spectral dynamics in concatenative speech synthesis.

In this chapter spectral dynamic measures are investigated as candidate join costs to measure spectral continuity. A flexible and robust strategy to extract spectral dynamic features is presented. Each spectral dynamic measure is investigated to determine the optimum parameters for feature extraction, with respect to detecting discontinuities. The

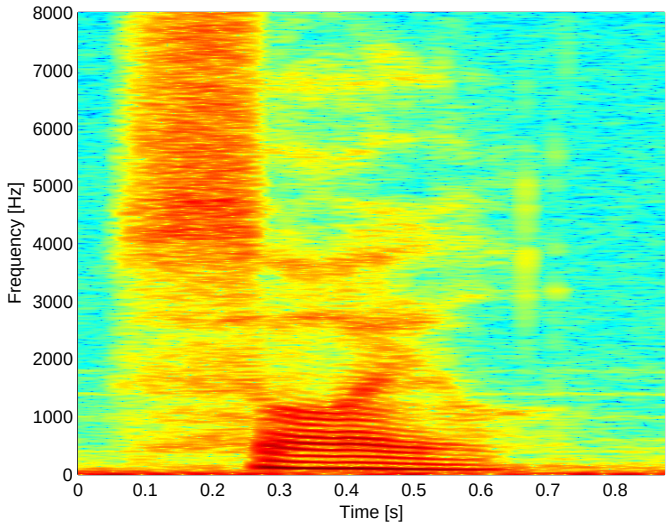

Figure 6.1: Spectrogram of the diphthong contained in the word 'soil'. Note the formant movements in the centre of the voiced region.

results for detecting discontinuities for each of the proposed spectral dynamic measures is presented. The results are analysed to determine the contribution of spectral dynamics as an independent source of discontinuity.

6.0.1 Spectral dynamics - overview and literature review

When two speech units are concatenated the resulting speech may contain an audible discontinuity. It has been identified that units exhibiting spectral dynamics near the join are more likely to contain discontinuities. Joins contained in diphthongs, which contain significantly higher levels of spectral dynamics than monophthongs, have been identified as having an increased likelihood of containing discontinuities (Syrdal, 2001). In Chapter 3 the diphthong /oy/ was found to contain the highest number of discontinuities in comparison to all other vowels in the test database. Fig. 6.1 illustrates the spectral dynamics in the diphthong /oy/ contained in the word 'soil'. Including spectral dynamic information in the join cost accounts for spectral movement on either side of the join. Such behaviour may result from co-articulation with neighbouring phones and is beyond the scope of local static spectral costs at unit boundaries. Static refers to features extracted from a single block of windowed speech.

It is understood that a number of different sources give rise to discontinuities in concatenated speech as reflected in the subcomponents of join costs used in unit selection. Which includes spectral, energy and $F0$ sub-costs and in some cases delta features. From the studies reported to date it is unclear if spectral dynamic mismatch plays a significant role in the perception of discontinuities. Furthermore it is unclear how to effectively incorporate spectral dynamic measures into join costs for unit selection. In Wouters and Macon (1998) and Vepa and King (2006) static spectral measures were combined with corresponding delta coefficients to represent spectral change in the overall measure. The addition of delta coefficients was found to contribute minor improvements, of the order of 1-2% for the study of Wouters and Macon, depending on the feature set. Vepa and King found that adding delta features sometimes decreased performance. The performance achieved by including delta features was inconsistent and relatively small when an improvement was achieved. This is in contrast to the successful use of delta features in other applications such as ASR (Rabiner and Juang, 1993) and HMM-based speech synthesis (Tokuda et al., 1995).

The objective in this chapter is to identify the role of spectral dynamic behaviour in the perception of discontinuities and to investigate spectral dynamic measures for use in defining perceptually salient join costs. A breakdown of specific objectives for this chapter follows:

- To quantify the correlation between spectral dynamic measures and human perception of discontinuity.
- To develop a technique for spectral dynamic feature extraction tailored specifically to represent spectral join costs.
- To analyse the role of spectral dynamic measures in conjunction with standard spectral measures.

6.0.2 Perception of spectral dynamics

The importance of spectral dynamics in speech perception has been well documented in the literature (Kluender et al., 2003; Moore, 2004; Bregman, 1990). The area has been studied in the context of psychoacoustic experiments (Furui, 1986a; Lindblom and Studdert-Kennedy, 1967), physiological models of the auditory system (Meddis et al., 1990) and has been successfully applied in many speech processing applications. The auditory system amplifies spectral bands that exhibit significant spectral change; a number of models have been proposed to mimic this behaviour both from physiological (Meddis et al., 1990) and psychoacoustic perspectives (Dau and Puschel, 1996). Many stages in the auditory system, from the hair cells upwards, respond to spectral dynamics. It is estimated that one third of cortical neurons can only be stimulated by time varying tones or complex transients (Yost, 2007).

Spectral dynamics are important for many auditory tasks. For example in sound separation, when the components of a sound are varied coherently it has the effect of accentuating the sound relative to other sounds present (Moore, 2004; Bregman, 1990). Sounds with incoherent changes are typically perceived as separate sounds. In Furui (1986a) it was reported that 'perceptual critical points' exist in a vowel coinciding with the points of maximal spectral transition. The concept of 'perceptual critical points' is particularly relevant in considering where a unit boundary should be located in concatenative synthesis as it identifies points at which the auditory system can predict future sounds based on spectral dynamics in the preceding sound. In concatenative speech synthesis, units of speech are joined together in the hope that they will be perceived as a single stream of continuous speech. If the spectral dynamic behaviour on either side of a join does not match then this may cause the auditory system to identify the join as an acoustic cue to a new auditory event. Similarly, spectral dynamics on the left hand side of a join might cause the auditory system to anticipate a sound that is inconsistent with the actual sound on the right hand side of the join.

Spectral dynamic information has been incorporated successfully into many applications in speech processing including ASR (Furui, 1990, 1986b), speech enhance-

ment (Quatieri and Dunn, 2002), speaker recognition (Soong and Rosenberg, 1988) and speech coding (Knagenheim and Kleijn, 1995). Understanding the role of spectral dynamics in speech perception provides a meaningful path to successfully integrating spectral dynamic information for a given application.

6.0.3 Approach

In order to investigate the role of spectral dynamics in concatenative speech synthesis and to meet the objectives of the chapter the following approach was adopted.

- To determine a suitable strategy for extracting spectral dynamic features - a new approach tailored to the problem of detecting discontinuities should be developed.
- To explore if the spectral dynamic measures correlate with human perception of discontinuity - each candidate measure should be tested for its ability to detect discontinuities in the test database.
- To identify the relative contribution of spectral dynamics when combined with standard spectral measures - a correlation analysis should be carried out to determine the degree of correlation between static and dynamic measures.

6.1 Spectral dynamic feature extraction

The temporal variations of both resonant frequencies across the frequency axis and amplitudes within a particular spectral band were both investigated. A three dimensional plot depicting the spectrogram of a diphthong is given in Fig. 6.2, in which the variations along the frequency and amplitude axes can be observed with respect to time. To fully describe the spectral dynamics requires estimates of the partial derivatives of this surface along each axis with respect to time. To determine the spectral dynamic measures an estimate of spectral dynamics is computed at the edge of each unit and the Euclidean distance is computed between the spectral dynamic features representing each unit.

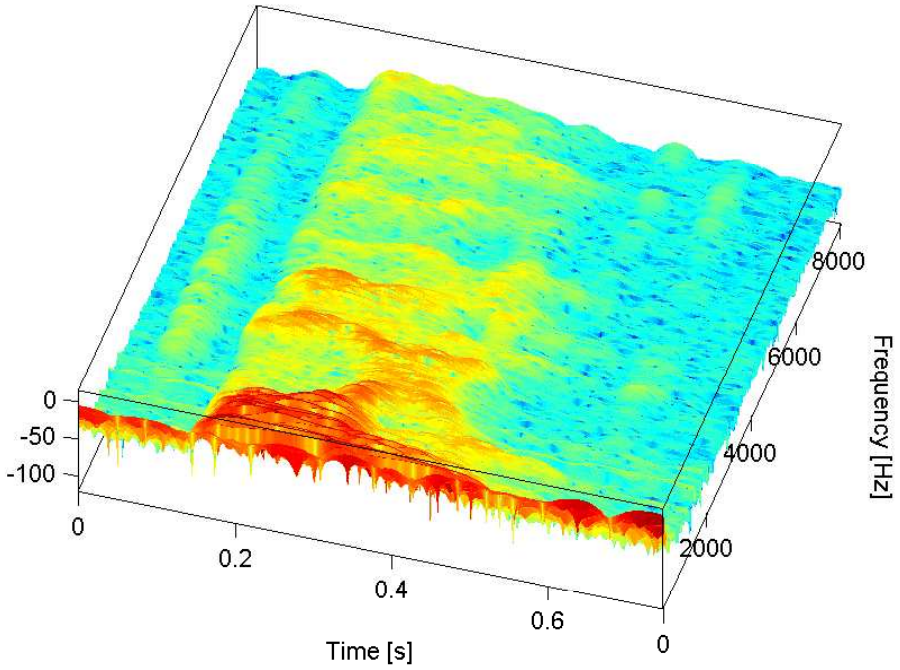

Figure 6.2: Three dimensional spectrogram of a diphthong illustrating the variation in frequency and amplitude of the spectrum components with respect to time.

6.1.1 Feature sets and feature extraction parameters

The amplitude-based features were MFCCs and the frequency-based features were LSFs. MFCCs were chosen to represent amplitude changes over time and LSFs to represent frequency changes. Other standard feature sets were tested with similar results and are not presented.

- LSFs were computed using the Burg algorithm to compute a 16^{th} order LPC model.
- MFCCs were computed as in Rabiner and Juang (1993) using a total of 19 coefficients; the first cepstral coefficient was discarded.

For both feature sets the raw speech was pre-emphasised with the filter $H(z) = 1 - 0.95z^{-1}$ and the parameters were computed with a 50 ms Hanning window centered on a pitch mark. The features were computed on each pitch pulse throughout the voiced region for each of the words in the inventory, of which there is a total of 300. This

results in parameter trajectories through the vowel centre. The spectral dynamic features were obtained by calculating the derivatives of the trajectories at the unit boundaries.

6.1.2 Trajectory modelling

Each parameter trajectory, computed through the voiced region in the vowel centre of the natural speech recordings, was modelled using a polynomial. The polynomial coefficients were computed that best fit the parameter trajectories with respect to the mean squared error. The order of the polynomial to model the trajectory was found to be important. As illustrated in Fig. 6.3, a low order polynomial will not sufficiently capture the variation in the data and over-modelling with an excessively high order polynomial will introduce variations in the trajectory that do not relate to the underlying data. Polynomials of order 4 were found to produce consistently good trajectory models. The number of data points included in the model was also found to be important. Empirically it was found that 9 data points gave an effective polynomial approximation. This corresponds with employing 9 pitch periods in the trajectory model. Fitting too many data points, covering a longer time span, was found to have a negative impact on results due to over modelling in regions distant from the join. Fitting too few points did not effectively represent the spectral changes about the join. The polynomial models computed for data points over 5, 9 and 15 pitch periods are illustrated in Fig. 6.4, each of which produce a different model and as such a different estimate of the spectral dynamic behaviour at the unit boundary. In particular note the slope difference at the centre data point where the derivatives are to be computed.

The first and second derivatives with respect to time were computed at the unit boundaries from the estimated polynomials and were employed to represent the spectral dynamics of the corresponding units. As a baseline for comparison, delta coefficients were also computed. The delta coefficients were computed using pitch synchronous analysis. A window length of one pitch period was employed to extract the features in the two pitch periods preceding the unit edge. The delta coefficients were generated by computing the difference between the feature vectors. Pitch synchronous feature extraction with a window length

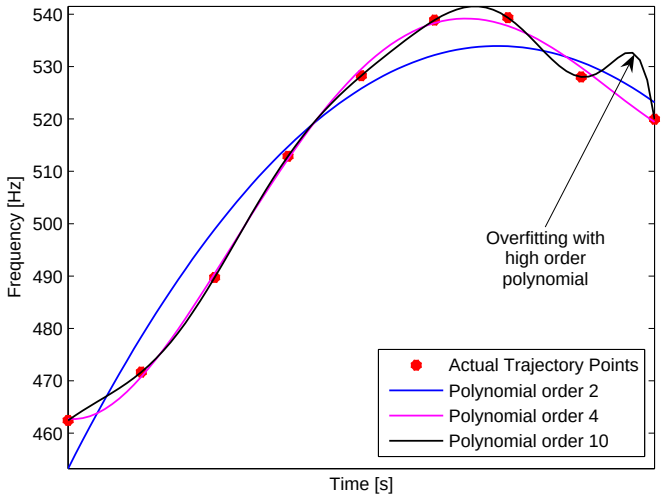

Figure 6.3: Plot of spectral feature (LSF) varying with respect to time, with the estimated polynomial models to fit the data. This illustrates the impact of the polynomial order in fitting a polynomial model to the data points.

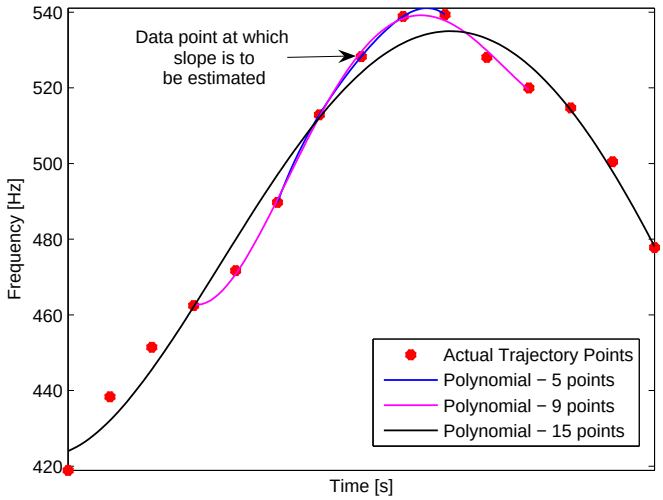

Figure 6.4: Plot of spectral feature (LSF) varying with respect to time, with the estimated polynomial models to fit the data. This illustrates the impact of the number of data points used to compute the polynomial model.

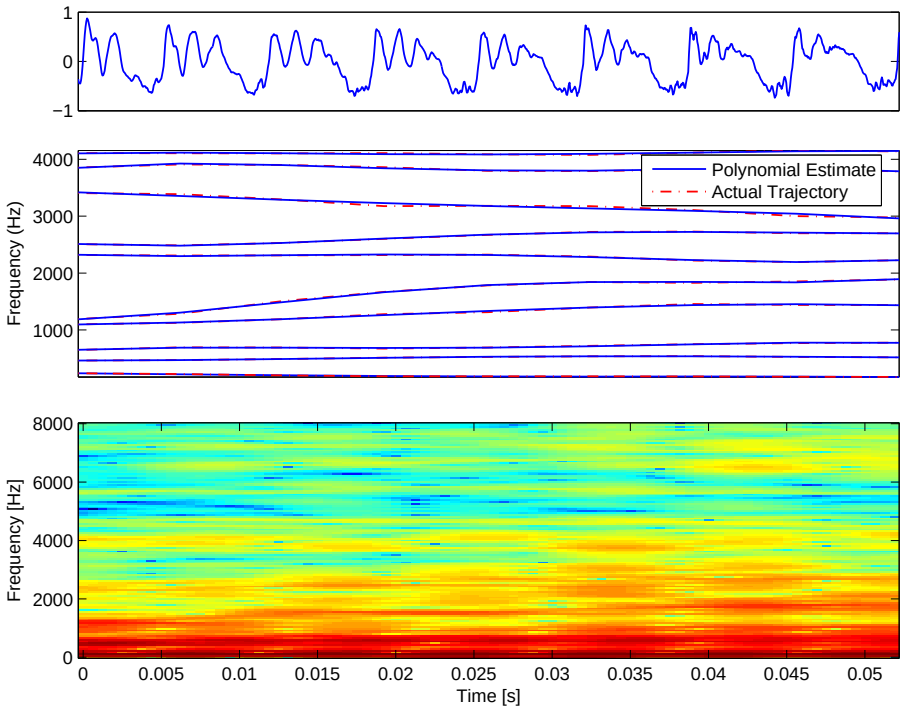

Figure 6.5: Illustrating the LSF trajectories in the vowel centre of the word 'went'; the speech waveform, the computed LSF parameters, the polynomial fit and the spectrogram are shown.

of one pitch period was found to be the optimum strategy for the task of detecting discontinuities with static spectral measures on the test database. Note that no attempt was made to select parameters to improve the performance of the delta coefficients.

6.2 Results

Each of the spectral dynamic measures were tested for the task of detecting discontinuities in the test database. In this section the results are presented for each candidate measure of spectral dynamics. The results are represented by the AUC value computed for each measure. The results for the spectral dynamic measures are presented in Table 6.1. The dynamic measures evaluated from the polynomial based derivatives are denoted by $d\mathbf{x}/dt$ and $d^2\mathbf{x}/dt^2$, the static measures by $\mathbf{x}$ and the delta features by $\Delta\mathbf{x}$. The Euclidean distance was used to quantify the degree of mismatch between feature vectors. The choice

of metric did not significantly influence the results.

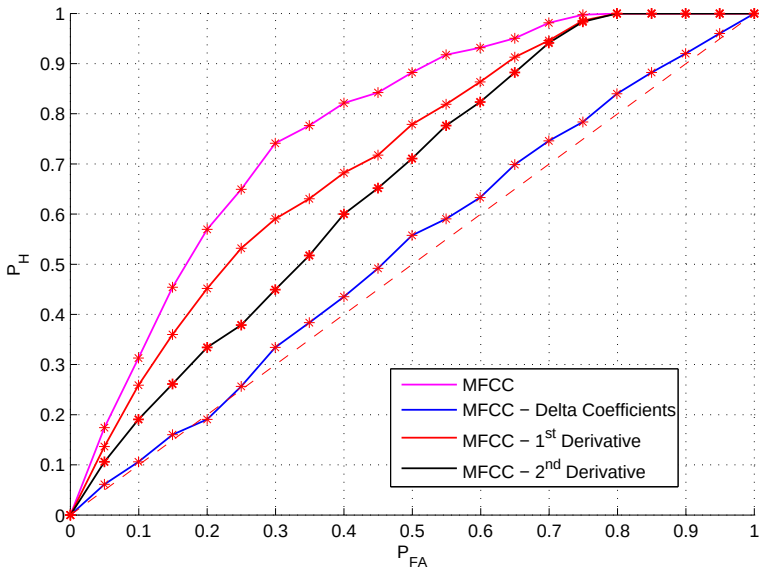

Figure 6.6: Comparing the ROC curves for MFCC measures from standard features, delta features and from the 1^{st} and 2^{nd} derivative trajectory based measures .

Features	$\mathbf{x}$	$\Delta\mathbf{x}$	$d\mathbf{x}/dt$	$d^2\mathbf{x}/dt^2$
MFCC	0.77458	0.52688	0.70860	0.65568
LSF	0.74238	0.50997	0.70801	0.67225

Table 6.1: Comparison of results for each candidate measure of spectral dynamics, the table entries indicate the AUC value for each of the measures.

The highest AUC value for spectral dynamic measures was obtained from the first derivative of the MFCC trajectories with an AUC value of 0.70860. The AUC values obtained for the delta coefficients was just above the value of pure chance, which corresponds with an AUC value of 0.5. The low detection rates for delta coefficients may be due to sensitivity to noise in numerical differentiation or perhaps that the measure reflects spectral change on a shorter time scale, due to shorter window lengths, which may not be relevant for the detection of discontinuities. The spectral dynamic measures based on the second derivatives were found to correlate with human perception, the results of which are also contained in Table 6.1. The results suggest that dynamic measures correlate with human perceptual results provided the features are appropriately extracted. The difference in using either LSFs or MFCCs for the dynamic measures had little impact on the results and

they are statistically equivalent for each type of dynamic measure with respect to the standard error for the AUC values presented. The same patterns emerge for both parameter sets. LSFs performed better than MFCCs for the measures based on the second derivative although they are still statistically equivalent. ROC curves comparing the performance of the dynamic measures and standard MFCCs are illustrated in Fig. 6.6.

It was found that adopting a longer window yielded better results with the proposed spectral dynamic measures, this is counter to static measures for which a single pitch period was found to yield the best results. Longer window lengths gave rise to less variation between successive spectral estimates and in turn gave rise to smoother trajectories, which were found to yield spectral dynamic measures with higher AUC values. It is significant to note that different feature extraction parameters are required for static and dynamic measures.

6.2.1 Combining static and dynamic features

In this section the results are presented for combined measures of continuity, consisting of both static and dynamic spectral features. The static and dynamic features are combined by concatenating the static and dynamic feature vectors. The distance of the combined feature vectors was computed as the Euclidean distance between the concatenated feature vectors. The results of the combined measures are presented in Table 6.2.

$$\mathbf{x} = \left[\mathbf{x}_{static}^{T},\ \mathbf{x}_{dynamic}^{T}\right]^{T} \tag{6.1}$$

The results of the combined measures are approximately equal to the results produced for the standard spectral measures for both MFCCs and LSFs. Combining the spectral dynamic measure computed from the second derivative was found to result in a decrease in performance when compared with the AUC value of static spectral measures, although it is a small decrease only evident in the third decimal place. This suggests that no advantage is gained by combining static and dynamic measures in this manner for the test database. The combined measures are statistically equivalent to the static measure and are well within the standard error bands of the original static measures. The reason that

no significant gain in performance is observed could be due to the manner in which the measures are combined or because each of the measures do not contain complementary discriminating information. This is discussed further in section 6.2.2. These results are similar to those reported in Wouters and Macon (1998) and Vepa and King (2006)

Features	MFCC	LSF
$[\mathbf{x},\ d\mathbf{x}/dt]$	0.77540	0.74046
$[\mathbf{x},\ d^2\mathbf{x}/dt^2]$	0.77399	0.73964
$[\mathbf{x},\ d\mathbf{x}/dt,\ d^2\mathbf{x}/dt^2]$	0.77409	0.73858

Table 6.2: Comparison of results for combining static and dynamic spectral measures, the table entries indicate the AUC values for each of the measures.

6.2.2 Analysis of results

In this section the results are analysed to gain further understanding of the advantages and limits of the proposed spectral dynamic measures as potential join costs. Window length and feature extraction parameters have already been identified as sources of variation in the results and in this section a phonetic breakdown of the results and a correlation analysis of spectral dynamic and static measures is presented.

Phonetic breakdown of results

The AUC values of the spectral dynamic measures computed from the first derivative of the features trajectories are illustrated in Fig. 6.7 and Fig. 6.8 for LSFs and MFCCs respectively for each phone in the test database. The results for each phone are similar for both MFCCs and LSFs, although LSFs significantly outperfrom MFCCs for the phone /ao/. From these plots it can be noted that the dynamic measures, similar to standard spectral measures, score relatively low AUC values for the diphthong /oy/. Apart from this it is difficult to extract any significant pattern from the data. The overall profile of the AUC values on a phone by phone basis is relatively similar to that found for standard spectral measures and does not contain any distinctive anomalies. The results suggest that phones containing spectral dynamics are not necessarily better served by a measure based

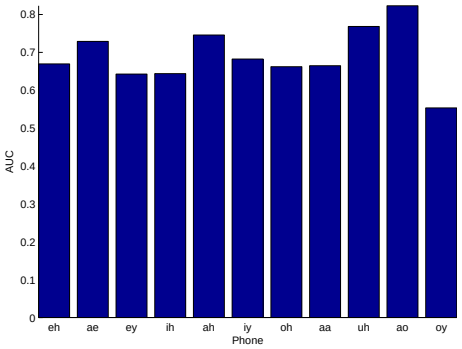

Figure 6.7: Bar chart indicating the AUC value computed for dynamic LSF (first derivative) for each phone containing joins in the test database.

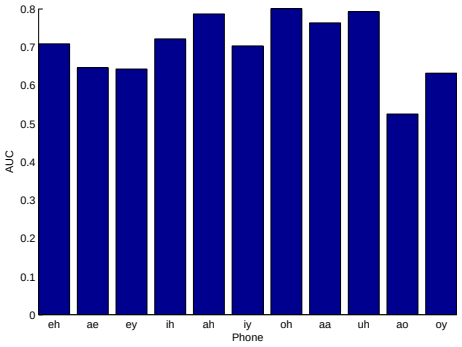

Figure 6.8: Bar chart indicating the AUC value computed for dynamic MFCC (first derivative) for each phone containing joins in the test database.

on spectral dynamics over a standard static spectral measure. In particular the results indicate that spectral dynamic measures do not offer any advantage over standard static spectral measures for diphthongs.

Correlation analysis

Motivated by the disappointing increase in results from combining spectral dynamic with standard measures, an analysis was conducted to investigate if a correlation exists between static and dynamic spectral mismatch in the test database. If static and dynamic mismatches are correlated then they are likely to occur simultaneously. It is unlikely that a significant increase in performance will result from combining these measures in such a scenario.

The degree of correlation between static and dynamic measures was quantified using

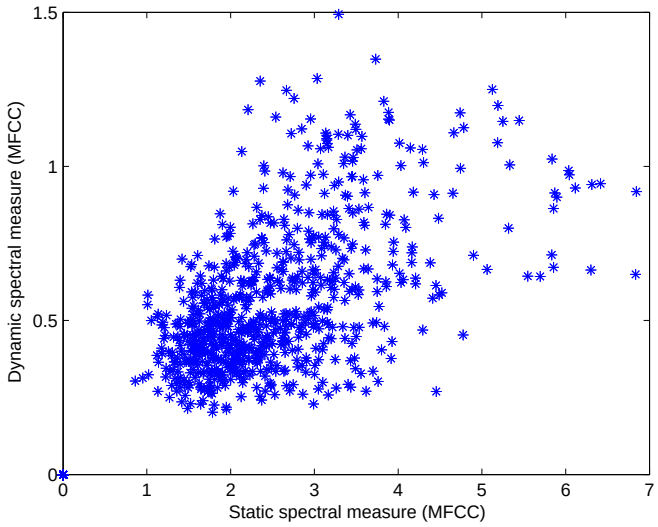

Figure 6.9: Scatter plot of the distance computed for MFCCs versus the spectral dynamic distance (first derivative) computed from MFCCs for each test word in the database.

the correlation coefficient, equation 6.2.

$$r_{xy} = \frac{\sum_{n=1}^{N}(x - m_x)(y - m_y)}{(N-1)\sigma_x\sigma_y} \tag{6.2}$$

The dynamic and corresponding static distances are represented by the vectors x and y, where m_x, m_y, σ_x and σ_y represent the mean and standard deviations of the distance vectors.

The correlation coefficients computed between static and dynamic spectral measures are presented in Table 6.3. The table entries indicate the correlation coefficients computed between the static and dynamic measures for both MFCCs and LSFs using measures of dynamics computed from both the first and second derivatives. The results indicate that there is significant correlation between static and dynamic spectral mismatch for the

Features	MFCC	LSF
$d\mathbf{x}/dt$	0.7748	0.7345
$d^2\mathbf{x}/dt^2$	0.7522	0.7272

Table 6.3: Correlation coefficients between static and dynamic spectral measures for both LSFs and MFCCs.

test database, with correlation coefficients in the range of 0.72 to 0.77. A scatter plot of static spectral distance versus dynamic spectral distance is illustrated in Fig. 6.9 for MFCCs. An approximate linear relationship can be observed from the plot. The correlation between the measures in the test database indicates that they are unlikely to produce significantly improved results when combined. This may explain why the improvement in results obtained from combining static and dynamic spectral measures is small relative to their individual performance.

6.3 Conclusion

In this chapter the role of spectral dynamics in concatenative synthesis was investigated. A procedure to compute spectral dynamic features was presented. The technique involved fitting polynomial models to feature trajectories. The details of feature extraction were found to significantly influence the results. In particular, the order of the polynomial model, the number of points in the trajectory to be modelled and the size of the analysis window used to extract the initial features to compute the trajectory model were all found to play a role. Using the polynomial model of the trajectories the derivatives could be estimated at the unit boundaries for each trajectory for a given feature set.

The first and second derivatives were tested as measures of spectral continuity for both LSFs and MFCCs. Dynamic measures computed from the first derivative showed a more significant correlation with the perceptual results than measures from the second derivative. To demonstrate the significance of carefully selecting the feature extraction technique and parameters, a dynamic measure was also tested based on a simple computation of delta features. The detection rate of the delta features for both MFCCs and LSFs was at the level of pure chance with AUC values of approximately 0.5, indicating that they do not correlate with human perceptual results of continuity. The degree of correlation with human perception was found to be sensitive to basic parameters used in feature extraction, the results indicated that smooth trajectories that do not contain the fine detail of temporal variations are more suitable for the task at hand. This is in contrast to results for static

spectral measures for which temporal resolution was found to have a significant impact for the detection of discontinuities.

Combining spectral dynamic measures with standard static spectral measures was not found to produce a significant gain in performance, similar to the results reported in Wouters and Macon (1998) and Vepa and King (2006). An analysis of the results was carried out in order to further understand the relationship between static and dynamic measures. Spectral dynamic distance measures were found to correlate with standard distance measures for the test database. This indicates that a large spectral dynamic mismatch is likely to occur simultaneously with a mismatch in the static spectrum. If this was to occur both measures would be likely to detect the same discontinuities in the test database. A breakdown of the detection rates for spectral dynamic measures indicated that the profile of the AUC values across all phones in the database was similar to that for standard spectral measures. Most notably the AUC value for spectral dynamic measures was consistently low for the diphthong /oy/ compared to AUC result for other phones. This was also the case for standard spectral measures. This indicates that spectral dynamic measures are not necessarily advantageous in joins contained in diphthongs.

Spectral dynamic measures were found to correlate with human perception of discontinuity although they did not provide a gain in performance when combined with standard static spectral measures. The details of feature extraction were found to significantly impact on the ability of the measures to predict discontinuities in the test database.

Chapter 7

Feature Transformations

In unit selection systems the spectral join cost is computed by extracting spectral features from speech frames adjacent to the unit boundaries and calculating the Euclidean distance between the features. In this computation the level of spectral mismatch between corresponding features is treated equally for all features. Perceptually it is unlikely that all features are equally significant. Mismatch below a certain threshold is likely to be perceptually irrelevant and should be discarded with no contribution to the overall distance measure. Certain spectral bands may be of more significance, for example mismatch coinciding with the location of a formant would be expected to be of more perceptual importance than mismatch in other regions of the spectrum. It has been reported that an abrupt increase in an acoustic component is more perceptually significant than a sudden drop in amplitude (Summerfield et al., 1986). This indicates that mismatch due to the introduction of a new component should be weighted more heavily than mismatch due to a drop in component energy. The central auditory system contains many complex components to process transients and detect changes in signal components that are likely to be active in the detection of discontinuities. A system that can detect discontinuities in concatenated speech from the extracted spectral features should have sufficient complexity to model this behaviour.

The objectives of this chapter are to investigate feature transformation algorithms to improve the performance of join costs in detecting discontinuities in concatenated speech.

This general objective can be broken down further into the following objectives for the chapter:

- To develop a framework that facilitates the use of feature transformation algorithms tailored for join costs.
- To investigate the use of linear feature transformations.
- To investigate the use of complex non-linear neural network based transformations on the features in join costs.
- To investigate the use of feature transformations to automatically combine features from different sub-costs into a single measure of continuity.

In this chapter a framework to represent a join cost as a feature vector that facilitates feature transformations is presented. Linear and non-linear feature transformations are applied to the task of improving the rate of detection of discontinuities, specifically principal component analysis (PCA) and neural networks. The results for applying the transformations to a set of spectral features to predict discontinuities in the test database is presented. PCA is investigated as a dimensionality reduction technique to be used as a preprocessing stage prior to applying the neural network. Neural networks are investigated as a means to automatically combine different feature sets into a single measure. The detection rates of the raw features and the transformed features are presented to illustrate the impact of the transform on the the detection of discontinuities.

7.1 Feature space framework

In order to apply feature transformations that fully exploit the discriminating information within each feature, it is necessary to define a suitable vector to represent a join between two concatenated units. Existing methods employ a distance to represent the continuity of a join and feature vectors to represent individual units of speech. In this thesis the error vector is used to represent a join, hereby referred to as the join vector, which is computed by subtracting the left and right unit feature vectors, $\mathbf{x}_{left}$ and $\mathbf{x}_{right}$. The left unit refers

to the unit on the left hand side of a join and the right unit refers to the unit on the right hand side of a join. Each feature in the join vector represents the degree of mismatch between the corresponding features in the left and right units.

$$\mathbf{x}_{join} = \mathbf{x}_{left} - \mathbf{x}_{right} \tag{7.1}$$

Different strategies to construct join vectors were initially investigated. The strategies were motivated by the standard, l_p norms and the symmetric Kullback-Leibler. Strategies to develop the join vectors were motivated by generalising the standard distance measure approaches reported in the literature on join costs (Klabbers and Veldhuis, 1998; Stylianou and Syrdal, 2001; Vepa et al., 2002; Wouters and Macon, 1998). The approach based on the l_p norm proved to be the most suitable as it produced better results in preliminary investigations (Kirkpatrick et al., 2006). This join vector generalises the standard distance measure approach for the l_p norms. Classification of the join vectors at this stage without applying a feature transformation, for join vectors constructed using equation 7.1, corresponds exactly with calculating the l_p distance between the original left and right feature vectors, for a given p, equation 7.2.

$$l_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{N} |x(i) - y(i)|^p \right)^{1/p} \tag{7.2}$$

The ideal join corresponds with the origin in the feature space and the quality of the join can be quantified as the distance of the join vector from the origin. Thus joins can be classified as continuous or discontinuous with respect to distance from the origin. When classification is based on a distance from the origin, the subsequent choice of norm for the feature space establishes the geometry of the classifier. The classifiers corresponding to the l_1, l_2, l_4 and l_∞ norms are illustrated in Fig. 7.1. This illustrates how the standard distance measures can be interpreted in the feature space. At this stage computing the l_2 distance is equivalent to computing the Euclidean distance between the original feature vectors. The advantage of this framework is that it enables the application of feature transformations with the potential to improve performance in detecting discontinuities.

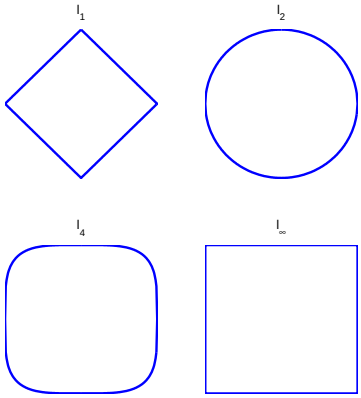

Figure 7.1: Classifier shape in 2 dimensions employing l_1, l_2, l_4 and l_∞ norms.

With the join vector representation it is possible to apply a transformation, $\mathbf{A}$, on the join vector before computing the final measure of mismatch, equation 7.4.

$$\mathbf{X} = \mathbf{A}(\mathbf{x}_{join}) \tag{7.3}$$

With this approach standard techniques can be applied to increase the separability of join vectors representing continuous and discontinuous joins. The application of a linear feature transformation is equivalent to stretching or contracting along individual axes and rotating the classifiers from Fig. 7.1, with a possible reduction in dimensionality. The final measure of mismatch, D, can be computed from the transformed vector, $\mathbf{X}$.

$$\mathbf{D} = |\mathbf{X}| \tag{7.4}$$

Two techniques were investigated: PCA and a neural network-based approach. For the neural network-based approach PCA was used as a preprocessing stage for dimensionality reduction of the input data. Ideally a feature transformation will remove redundant information and weight perceptually important information resulting in improved discrimination between continuous and discontinuous joins.

7.2 Feature transforms

Many studies have been conducted in the automatic speech recognition literature investigating the use of feature transformations to improve the discriminating qualities of the features (Somervuo et al., 2003; Somervuo, 2003). In this study PCA (Jolliffe, 1986) and neural networks (Bishop, 1995; Haykin, 1994) are applied to transform features representing the spectral join vectors. Supervised learning techniques require data to train the transform. This requires splitting the database into a training set and a testing set. Neural networks are supervised and must be trained on suitable data. PCA is an unsupervised transform and does not require training data. The objective is to investigate the ability of these techniques to enhance existing measures for objective detection of discontinuities.

7.2.1 Principal component analysis

Principal component analysis is an unsupervised learning technique and does not require splitting the database into separate training and testing datasets. In this thesis PCA is investigated in two roles: firstly for its ability as a linear unsupervised learning technique to improve the detection of discontinuities and secondly as a dimensionality reduction technique to remove redundant information preceding the application of neural networks. The removal of redundant information with PCA often leads to an improvement in performance in many pattern recognition tasks (Jolliffe, 1986).

In the implementation of PCA the data is centered in the feature space about the origin by subtracting the mean vector computed over the complete database. The data is also normalised with respect to variance such that the standard deviations are equal to one. The normalised data is transformed using PCA, this produces transformed join vectors whose components are uncorrelated and ordered according to the magnitude of their variance.

7.2.2 Neural networks

With the application of neural networks a mapping can be learned from data provided from subjective listening tests relating continuous and discontinuous joins with input feature

vectors representing a join. The appropriate relationship between the features and the detection of a discontinuity is data-driven and does not require advanced knowledge of auditory processing.

For each of the feature sets considered and for a number of possible combinations of feature sets a neural network was trained from the training set of the database and subsequently tested on the testing set. PCA is applied as a preprocessing step to reduce the dimensionality of the input vectors before training the networks. PCA is applied to retain components that contribute more than 0.1% to the variance in the data set. When the join vector is passed through the neural network the final distance measure is output. To train the neural networks join vectors corresponding with discontinuities are assigned an output value of 1 and continuous joins are assigned an output value of 0. The training set consisted of 50% of the database and the testing set consisted of the other 50%. Joins containing high $F0$ and energy differentials were retained in the database. A preliminary investigation found that results obtained were similar with or without joins exhibiting large $F0$ and energy differentials. $F0$ and energy can be easily appended to the feature vectors and is discussed in 7.3.3. The testing and training subsets were divided such that an equal number of discontiuities occurred in each subset. Similarly each subset had equal coverage across all phones represented in the databse.

A number of classes of neural networks were initially investigated as candidate feature transformations including;

- General regression neural networks (GRNN) (Specht, 1991; Wasserman, 1993).
- Probabilistic neural networks (PNN) (Specht, 1990; Bishop, 1995).
- Feedforward neural networks (FFNN) (Haykin, 1994; Bishop, 1995).
- Radial basis neural networks (RBN) (Haykin, 1994; Bishop, 1995).
- Exact radial basis neural networks (ERBN) (Haykin, 1994).

Initial test results indicated that FFNNs, PNNs, RBNs and ERBNs were found to be less consistent than GRNNs. GRNNs were found to be the most suitable for the task as

they consistently produced high AUC values. Gains in AUC values were robust to parameter variations and changing feature sets with GRNNs unlike the other neural networks considered. Radial basis and exact radial basis networks were found to produce relatively high AUC values in some cases although the results were sensitive to parameter changes and feature sets. The results produced by feedforward networks were significantly lower then for GRNNs and were inconsistent. The algorithm employed to train feedforward networks contains parameters that are randomly initialised. The selection of these parameters was found to influence the resulting performance with different AUC values produced each time a feedforward networks was trained. The choice of the initial parameters can result in problems with the network parameters getting trapped in local minima (Haykin, 1994), which is a general problem with iteratively trained neural networks. GRNNs do not suffer from the same problem. The results presented in this chapter are for GRNN networks.

GRNNs are often used in applications for function approximation when the form of the underlying relationship is unknown. They are similar to radial basis networks and contain an input layer, a radial basis layer and a linear output layer. In training a GRNN a value must be selected for the spread parameter for the neurons in the radial basis layer. The spread parameter determines the width of the distribution of the (Gaussian) activation functions within each neuron in the second layer (Patterson, 1996). The spread parameter can be employed to fine tune the performance of a network.

7.2.3 Combining feature sets

To combine feature sets, each corresponding join vector is concatenated and subsequently the transformation is applied.

$$\mathbf{x} = \left[\mathbf{x}^T_{features1},\ \mathbf{x}^T_{features2}\right]^T \tag{7.5}$$

When combining feature sets it is important to effectively utilise the discriminating information contributed by each feature vector. The individual feature vectors may be on considerably different scales and there may be significant correlation between the measures

introducing redundancy. The transformations are tested for the ability to automatically combine the information represented in the individual feature vectors. This can be applied to combine different spectral measures, for example MFCCs with LSFs or alternatively to combine different sub-costs such as energy, $F0$ and spectral measures into a single measure. An ideal transform will remove redundant information and retain information to facilitate discriminating between continuous and discontinuous joins. To successfully combine measures the transformations must retain the discriminating information unique to each constituent component. In this chapter results are presented based on combining standard spectral measures with dynamic spectral measures, combining standard spectral measures with non-spectral measures such as energy and $F0$. In addition all possible combinations are considered to determine the overall best performing measure.

7.3 Results

In this section the results are presented for both PCA and neural network transformations for a number of feature sets. The results represent the AUC value which indicates the rate of detection of discontinuities in the test database. Results are presented for both PCA and neural networks for the following features sets: MFCC, LSF, Log PS, Morlet wavelet and GDF LPC. Results are presented for spectral dynamic measures (LSF and MFCC) and various combinations of feature sets with the neural network transformation. Note that the testing set for neural networks is a subset of the database. This should be considered when making comparisons with AUC values reported earlier in the thesis. To create a reference point for comparison, the AUC value of the standard Euclidean distance for a given feature set is computed and presented alongside the results employing the transformation approach. As the number of discontinuities in the testing set is now half the number compared to AUC values computed in previous chapters the standard error values are different for the AUC values presented in this chapter. The standard error versus AUC value corresponding with the results presented in this chapter is given in Fig. 7.2.

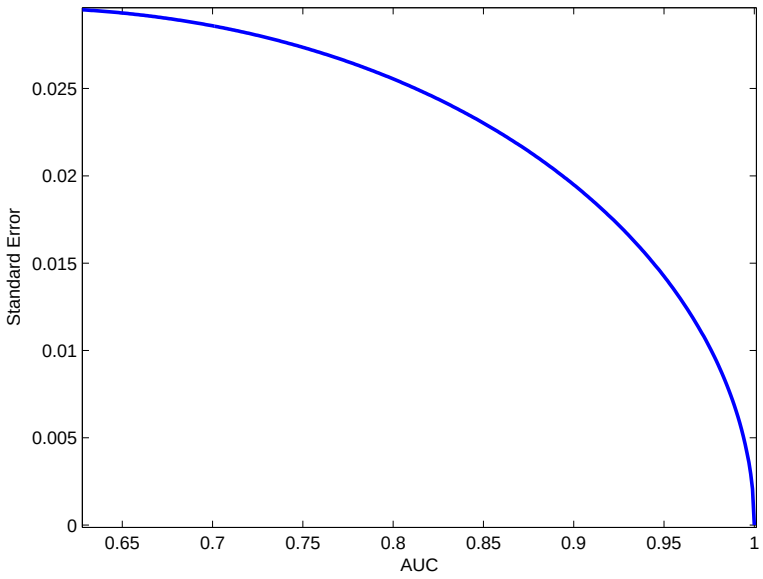

Figure 7.2: Plot of the AUC versus standard error for the testing protion of the database employed with feature transformation based measures.

Features	$\mathbf{x}$	$PCA[\mathbf{x}]$	Dimension
Morlet	0.8059	0.8123	3 (50)
Log PS	0.7673	0.7892	32 (256)
MFCC	0.7565	0.7714	3 (19)
LSF	0.7468	0.7113	16 (16)
GDF LPC	0.7521	0.7097	8 (256)

Table 7.1: Comparison of results with and without the application of PCA for each feature set, the table entries indicate the AUC value and the dimension of the feature vector after PCA is applied.

7.3.1 PCA

The results comparing the AUC values computed before and after the application of PCA are presented in Table 7.1. The results include the AUC value from the Euclidean distance and after the application of PCA to the join vectors. The final column indicates the dimension of the join vector after PCA with the original dimension in brackets. These results were computed across the test subset of the database. Similar results were computed for the complete database. For consistency in the presentation of results in the section on neural networks the results are presented for the test subset only.

For MFCCs, Log PS and Morlet wavelets, the application of PCA was found to improve

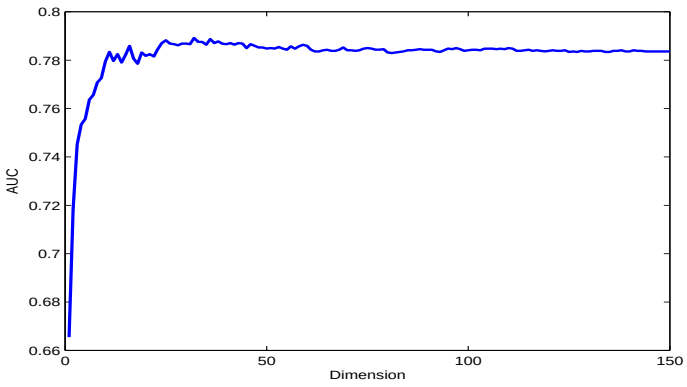

Figure 7.3: Plot of the AUC versus output dimension when applying PCA to the Log PS join vectors.

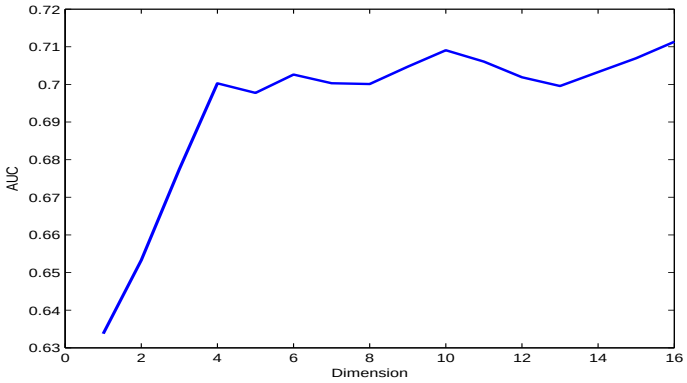

Figure 7.4: Plot of the AUC versus output dimension when applying PCA to the LSF join vectors.

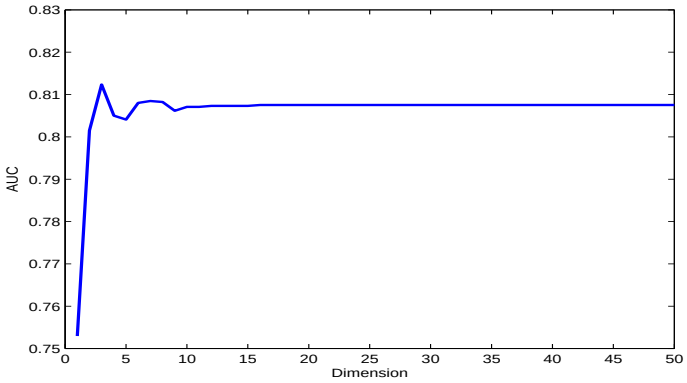

Figure 7.5: Plot of the AUC versus output dimension when applying PCA to the Morlet wavelet join vectors.

the rate of detection of discontinuities. The improvement is of the order of 1-2%. This is not a statistically significant improvement with respect to the standard error. For LSFs and the GDF LPC, PCA was found to result in a decrease in the AUC value. This decrease in performance is due to the preprocessing stage in which the mean is subtracted. The decrease is not observed when the preprocessing stage is omitted. The dimension of the transformed vectors was chosen to maximise the AUC value for each of the feature sets. Figures 7.3, 7.4, 7.5 illustrates the resulting AUC values as the dimension of the PCA transformed vector is varied for join vectors constructed from Log PS, LSFs and the Morlet wavelet. Each figure has a similar characteristic, with a rapid rise in the AUC values as the dimension increses followed by a plateau. Fig. 7.4 is different to the other plots as the plateau is not as distinct and the maximum value corresponds with the maximum dimension. This was found to be due to the preprocessing stage in which the mean is subtracted. When this preprocessing is omitted the plot for LSFs exhibits the same characteristics as the other feature sets considered with a peak at a low dimension and a more distinct plateau. This suggests that much of the discriminating information can be compressed into a small number of coefficients. In particular the wavelet and MFCC measures have a maximum AUC value at a dimension of 3.

The results indicate that PCA is limited as a method to improve the detection of discontinuities but is effective at compressing the discriminating information in the join vectors into a low dimensional representation. This suggests that PCA is suitable as a preprocessing stage for neural networks.

7.3.2 Neural networks

The results for each of the candidate feature sets before and after the application of the proposed neural network-based transformation are presented in Table 7.2. The results presented are for GRNN type networks with the spread parameter equal to 1. The neural network is trained on the training set and tested on a separate testing set. PCA is employed as a preprocessing stage to reduce the dimensionality of the input vectors. The first column of results, $\mathbf{x}$, represents the AUC value for the corresponding feature set computed using

the Euclidean distance. The second column, $NN[\mathbf{x}]$, represents the AUC value for the corresponding feature set computed from the measure output by the neural network.

Features	$\mathbf{x}$	$NN[\mathbf{x}]$
Log PS	0.7673	0.8744
MFCC	0.7565	0.8413
Morlet	0.8059	0.8307
LSF	0.7468	0.7955
GDF LPC	0.7521	0.7135

Table 7.2: Comparison of results before and after the application of the neural network for each feature set, the table entries indicate the AUC value.

The results contained in Table 7.2 indicate that the neural network-based approach enhances the performance considerably in most cases. The gains in performance achieved by the neural network are statistically significant, with respect to the standard error, in all cases except for the Morlet wavelet. This is most notable for Log PS in which the AUC value increases from 0.7673 with the standard distance measure approach to a value of 0.8744 with the proposed approach. Gains are also achieved for MFCCs, Morlet wavelets and LSFs. The only measure not to improve is the GDF LPC. The ROC curves comparing before and after the application of the neural networks are illustrated in Fig. 7.6 for MFCCs and Fig. 7.7 for Log PS.

Further improvements on these results can be achieved by tailoring the spread parameter for each individual neural network. The spread parameter was varied in steps of 0.25 and the maximum AUC value was recorded. The results are presented in Table 7.3 including the maximum AUC value and the corresponding spread parameter. Fine tuning the spread parameter results in an increase in performance for all feature sets, including the GDF LPC that showed a decrease in performance in Table 7.2 when the spread value was set to the default value of 1.

The results presented in Table 7.4 represent the performance of dynamic measures, as computed in Chapter 6, before and after the application of the neural network transformation. The results indicate that the neural network can enhance the performance of the spectral dynamic measures. The spectral dynamic measures computed from the first

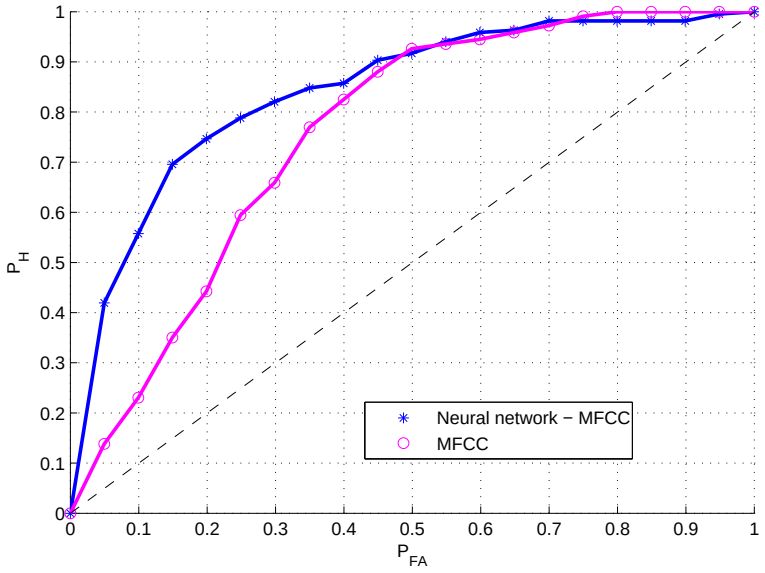

Figure 7.6: ROC curves for MFCCs before and after the feature transformation with the neural network.

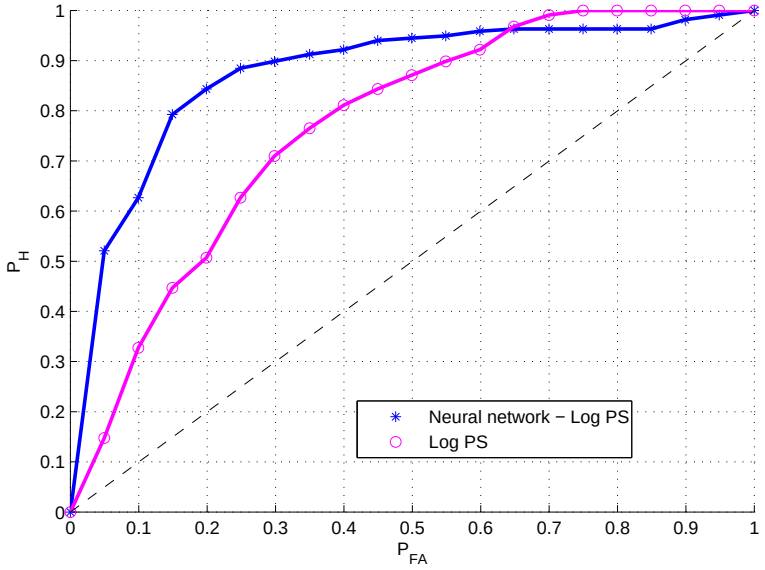

Figure 7.7: ROC curves for Log PS before and after the feature transformation with the neural network.

Features	Spread	$NN[\mathbf{x}]$
Log PS	4.5	0.8944
MFCC	1.5	0.8572
Wavelet Morlet	1.5	0.8521
LSF	1.25	0.7969
GDF LPC	5.25	0.773

Table 7.3: Results of the neural network transformation for each feature set with the spread parameter tuned for each individual feature set.

derivative of the LSF trajectories scored an AUC value of 0.8201 after the neural network transformation. This score is higher then any of the AUC values computed with static spectral measures in conjunction with the Euclidean distance (of which the best reported for the testing set is 0.8059 with the Morlet wavelet). Measures based on the second derivative also show an improvement in the AUC value with the application of the neural network. These results further validate the results of Chapter 7 indicating that spectral dynamic measures correlate with human perception of discontinuity.

Features	$d\mathbf{x}/dt$	NN$[d\mathbf{x}/dt]$	$d^2\mathbf{x}/dt^2$	NN$[d^2\mathbf{x}/dt^2]$
MFCC	0.7369	0.8017	0.6643	0.7274
LSF	0.7065	0.8201	0.6625	0.7518

Table 7.4: Comparison of results before and after the application of the neural network for each feature set, the table entries indicate the AUC value.

7.3.3 Combined measures

To combine the measures, each join vector is concatenated and subsequently PCA is applied to decorrelate the combined join vector and reduce the dimensionality. The output from PCA is used to train the neural network using the training subset of the database. The results here are for a GRNN neural network with the spread parameter set to 1. Results are presented for combining different standard spectral representations, combining standard spectral representations with dynamic measures and combining $F0$ and energy measures with standard representations.

The results computed from the standard distance measures with no transformation

based on the concatenated join vectors and those computed from the neural network-based measures are presented in Table 7.5. The neural network-based measure derived from the combination of MFCC and Log PS features has the highest AUC value of the combined measures at 0.8859. For each of the combined measures the neural network-based measure outperforms the corresponding standard measure. The results indicate that combining standard spectral measures can increase performance when using neural networks.

Features	$\mathbf{x}$	$NN[\mathbf{x}]$
MFCC	0.7565	0.8413
MFCC + LSF	0.7468	0.8581
MFCC + Log PS	0.7673	0.8859
MFCC + Morlet	0.8121	0.8735
MFCC + GDF LPC	0.7525	0.7135

Table 7.5: Comparison of results with and without the application of the neural network for possible combinations of spectral feature sets, the table entries indicate the AUC value.

The results for combining static and dynamic measures in conjunction with the neural network transformation are illustrated in Table 7.6. Combining the standard spectral measures with the dynamic measure computed from the first derivative, $[\mathbf{x}, d\mathbf{x}/dt]$, increases the AUC value for both feature sets. The addition of the spectral dynamic measure computed from the second derivative, $[\mathbf{x}, d^2\mathbf{x}/dt^2]$, also results in an increase in the AUC. The increase of performance from combining the dynamic measure from the first derivative greater. Combining all three of these measures,$[\mathbf{x}, d\mathbf{x}/dt, d^2\mathbf{x}/dt^2]$, also provided an increase. For MFCCs the combination of the second derivative measure, $d^2\mathbf{x}/dt^2$, with the static and first derivative measure reduced the AUC value.

Features	NN$[\mathbf{x}, d\mathbf{x}/dt]$	NN$[\mathbf{x}, d^2\mathbf{x}/dt^2]$	NN$[\mathbf{x}, d\mathbf{x}/dt, d^2\mathbf{x}/dt^2]$
MFCC	0.8880	0.8563	0.8673
LSF	0.8307	0.8142	0.8346

Table 7.6: Results after the application of the neural network for combined static and dynamic features, the table entries indicate the AUC value.

The results for combining standard spectral measures with $F0$ and energy measures are presented in Table 7.7. The results indicate that a marginal increase in the AUC value

can be achieved with the inclusion of these parameters.

Features	$\mathbf{x}$	$NN[\mathbf{x}]$
MFCC + F0	0.7142	0.8455
MFCC + E	0.7567	0.8468
MFCC + E + F0	0.7124	0.8562
LSF + F0	0.7491	0.8330
LSF + E	0.7470	0.8197
LSF + E + F0	0.7491	0.8408
Log PS + F0	0.7867	0.8774
Log PS + E	0.7673	0.8756
Log PS + E + F0	0.7807	0.8770

Table 7.7: Comparison of results with and without the application of the neural network for feature sets with and without $F0$ and energy, the table entries indicate the AUC value.

A large number of possibilities exist to combine measures, most of which produce very similar results. The highest AUC value achieved was for a combined measure consisting of:

- Morlet wavelet.
- MFCC.
- MFCC based dynamic measure computed from the first derivative.
- $F0$.
- Energy.
- The neural network had a spread parameter of 2.5.

The AUC value achieved for this measure was 0.9218. The ROC curve for this measure is illustrated in Fig. 7.8 alongside the benchmark MFCC measure (MFCC AUC = 0.7565).

7.3.4 Summary of results

From the results presented a number of key points can be noted:

- Neural networks produced a statistically significant improvement in the performance of standard features in detecting discontinuities. Log PS increased its performance

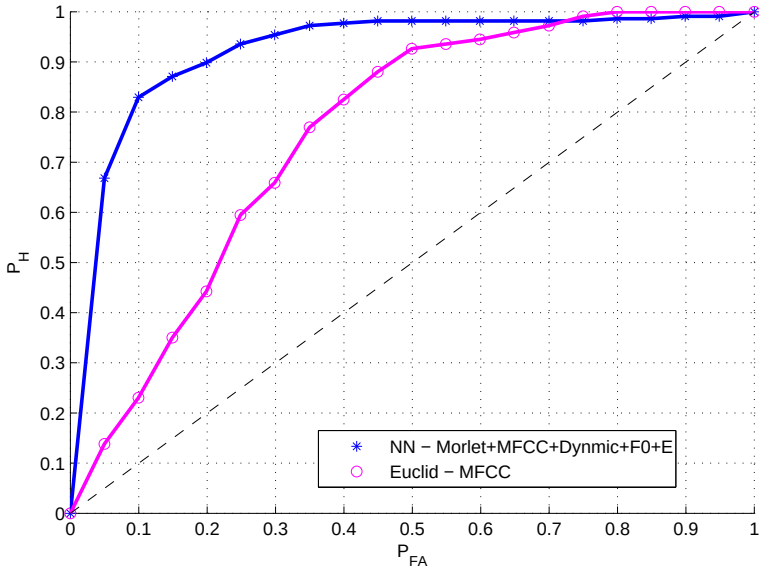

Figure 7.8: ROC curves comparing the performance of the best neural network based measure and the baseline MFCC measure.

from 0.7673 with Euclidean distance to 0.8944 with a neural network transformation in which the spread parameter was tuned to optimise performance.

- Not all neural networks provided useful results. GRNNs were found to be the most suitable of the candidate networks considered in this study.

- PCA is an effective means to compress data contained in join vectors. For example wavelet coefficients could be compressed from 50 to 3 and achieve a small increase in the AUC value (from 0.8059 to 0.8123).

- Neural networks can be used to produce a single combined measure from a number of distinct feature sets. The best performing measure was a combined measure with an AUC value of 0.9218.

7.4 Conclusion

A feature space framework was successfully developed that generalises the standard distance measure computation used on join costs to date. This framework is significantly more flexible and enables the application of complex feature transformation algorithms. Feature

transformations were investigated as a means of improving the detection of discontinuities in the test database. A linear transformation, PCA, and a neural network based technique were applied and tested. PCA was not significantly advantageous as a means to improve detection although it proved to be suitable as a preprocessing stage to compress join vectors prior to the application of neural networks. A neural network based approach using GRNN type neural networks was found to considerably improve the detection of discontinuities. A combined measure, employing many features, using the feature space framework and a neural network transformation produced the highest AUC value of all measures tested in this thesis. The results in this chapter support the theory, presented in Chapter 2, that discontinuity detection is related to neural processing in the auditory system. The neural network based approach does not require detailed knowledge of the underlying auditory processes as it is data-driven. The results suggest that this technique has sufficient complexity to model the auditory processes involved in the detection of discontinuities.

Chapter 8

Conclusions

The objective of this thesis was to develop a deeper understanding of human perception of discontinuity in concatenated speech and to use that knowledge to develop new measures of spectral continuity that correlate with human perception. The approach adopted to solve the problem was to identify the limits of existing measures and investigate new strategies guided by knowledge of the human auditory system.

The perceptual experiment

A perceptual experiment was conducted on a test database of concatenated words. The database contained a set of concatenated synthetic words each of which was judged to be continuous or discontinuous by human listeners in a formal perceptual experiment. Analysis of the database revealed that discontinuities were more likely to occur in vowels with relatively more formant variation and, as with previous studies, diphthongs were identified as having a high probability of containing a discontinuity. This database was used to correlate perceptual data with candidate measures of continuity and was used extensively throughout the thesis to correlate candidate measures of continuity with perceptual results.

Comparison of standard spectral measures

An investigation was conducted to quantify the degree of correlation between standard spectral representations used in speech processing and the perceptual results. The results

of this study revealed that although the measures correlated with human perception, none of the standard measures tested had a satisfactory correlation, and that significant scope for improvement existed. Analysis of feature extraction parameters revealed that the results for the standard measures were dependent on the window length used for feature extraction. The choice of window length was found to have a more significant impact on the results then the choice of feature set. The tests indicated that a pitch synchronous feature extraction procedure provided the best results. The results obtained for varying the window size indicated that a trade-off in the time-frequency resolution of the spectral estimates may be limiting the performance of the spectral measures. The limitations of standard spectral measures, with respect to time-frequency resolution, was further reflected in the results produced for individual vowels. Vowels with relatively stationary spectra produced the best results and diphthongs containing spectral dynamics consistently produced the lowest results.

Wavelets

To provide a solution to the limitation imposed by the time-frequency resolution of standard spectral features a measure based on the wavelet transform was proposed and evaluated. Wavelet transform measures were found to outperform standard spectral measures and demonstrated a higher degree of correlation with human perceptual results. The improvement in the results was found to be consistent for a number of different wavelet functions, with only a few exceptions, and is thought to be due to the time-frequency characteristics of the wavelet transform. Analysis of the results indicated that the increase in performance for wavelet based measures was particularly significant for diphthongs. Diphthongs contain more spectral dynamics than monophthong vowels and were found to produce the lowest results for standard measures compared with results for other vowels in the database. The trade-off between time and frequency resolution to produce a reliable spectral estimate is more challenging in diphthongs compared with monophthong vowels due to the non-stationary spectra. Wavelets did not always outperform standard spectral measures, such as MFCCs, for vowels with relatively stationary spectra. The results suggest that the

wavelet transform can outperform standard measures in the case of joins contained in non-stationary vowels but does not necessarily offer an advantage for vowels with stationary spectra at the unit boundary.

Phase-based spectral measures

Standard spectral features employed in speech processing are predominantly computed from the magnitude spectrum. It is known that information relating to the magnitude spectrum, phase spectrum and spectral dynamics are utilised in the auditory system in speech perception. Most previous studies have tested magnitude-based features only as measures of spectral continuity. In this thesis measures based on the phase spectrum and spectral dynamics were explored.

A measure of spectral continuity derived from the phase spectrum, in the form of the group delay function was proposed as a measure of spectral continuity and evaluated on the test database. The group delay function is the derivative of phase with respect to frequency and was chosen as a representation of phase as it does not suffer from the ambiguities of the raw phase spectrum. A number of methods to estimate the group delay function were investigated. Each of the measures tested were found to correlate with human perceptual results, although the level of correlation was less then that for standard spectral measures. A number of the group delay measures outperformed magnitude measures for diphthongs in the test database. The choice of parameters for feature extraction, such as window shape and length, was found to have a significant effect on the results. The group delay measures of continuity did not outperform the magnitude measures in general, although they did produce better results for diphthongs.

Spectral dynamics

Spectral dynamics were investigated as a new candidate measure of continuity. Spectral dynamics have been demonstrated as an important perceptual cue in speech and play an important role in the auditory system. A measure of spectral dynamics was developed by modelling the trajectories of standard spectral features with a polynomial and computing

the first and second derivatives of the polynomial at the join. The proposed measures were demonstrated to correlate with human perceptual results in the test database. The parameters for feature extraction were found to be important. In particular the window length and the parameters required to construct the polynomial model were found to significantly influence the results. Features extracted with unsuitable parameters were found to show no correlation with human perceptual results. For example MFCC delta coefficients computed using an analysis window of one pitch period were found to contain almost no discriminating information to indicate if a join is continuous or discontinuous. Combining spectral dynamic measures with standard magnitude measures did not produce any significant increase in performance. An analysis of the database revealed that spectral dynamic and magnitude measures are correlated in the test database. This explains to some extent why combining spectral dynamic and magnitude measures does not produce a significant increase in performance.

Feature transformations

The standard approach to quantify the degree of discontinuity between two speech units is to compute the Euclidean distance between the feature vectors representing each unit. The processing of transient sounds and the detection of events in the human auditory system is complex and involves many locations in the central auditory system. It is likely that similar processes are involved in the detection of discontinuities. The Euclidean distance does not have sufficient complexity to model such behaviour. As a solution, a novel feature space framework was proposed that uses a vector to represent a join. This framework enables the application of feature transformations that have been successful in other areas of speech processing. This approach allows research to focus on investigating how the mismatch at a join should be processed to correspond with human perception as opposed to feature extraction.

The application of PCA and neural networks were investigated as feature transforms and were tested for their ability to produce measures that correlated with the perceptual results. PCA was found to be an effective tool for reducing the dimensionality of vectors

whilst retaining discriminating information. PCA was not found to significantly improve the correlation with human perceptual results, although small gains were found in some cases. A neural network transform using GRNN type neural networks was found to significantly improve the correlation of standard feature sets with perceptual results. Measures using this technique consistently correlated with human perceptual results and produced correlations of over 90%. Standard measures (without transformation) were found to correlate at approximately 75% and wavelet measures at approximately 80%. Neural networks are a supervised learning technique and require training data. This is a disadvantage for the application of this approach in practice.

Summary

The key findings of the thesis are as follows;

- Time-frequency resolution is a limiting factor for standard spectral measures. Wavelet-based measures have superior time frequency resolution and were found to improve the correlation with perceptual results; the most notable increase in performance was for diphthongs.

- Measures derived from the phase spectrum, specifically the group delay function, were found to correlate with human perception of discontinuity in the test database. These measures did not perform as well as magnitude measures in general but were found to produce better results for diphthongs.

- Spectral dynamics measures were found to correlate with human perception of discontinuity. The details of feature extraction were found to significantly influence the degree of correlation. Combining spectral dynamic measures with standard measures provided no significant improvement in performance.

- A new feature space framework was proposed to allow feature transformations on vectors used to represent joins. This approach combined with a neural network based transformation produced a significant increase in the correlation with human perceptual results and provided the highest correlations reported in this thesis.

The findings in this thesis advance the understanding required to relate human perception of spectral discontinuity with objective measures of continuity computed directly from the speech signal.

8.1 Future work

Many of the results presented in this thesis would benefit from further study. A number of areas that could expand on the research presented in this thesis are outlined in this section.

8.1.1 Test database and perceptual experiment

One of the limitations of the results presented in this thesis relates to the amount and nature of the test data from the perceptual experiment. The test database is of a single male speaker and only considers joins in vowels. To support the conclusions presented it would be beneficial to further investigate the key findings with a larger and more varied test database. An ideal test database would have a number of both male and female speakers, each with varied vocal characteristics and speaking styles. This would create a framework to test and develop measures of spectral continuity that are more likely to generalise to unknown speakers. For example, to an arbitrary speaker in a unit selection database. This would also establish a framework to analyse the impact of speaker specific characteristics on each candidate measures of spectral continuity, for example $F0$, and thus providing a strategy to develop a measure that is speaker independent. Generalising the results to multiple speakers is particularly important considering the number of conflicting results reported in previous studies in the literature. Further research results could be generated by considering a larger variety of phones. This study only considers joins contained in vowels and does not consider joins in consonants. A larger quantity of each individual phone type may also make it easier to identify consistent patterns in the data.

8.1.2 Wavelets and feature extraction

A further investigation to clarify when it is advantageous to use wavelets over standard spectral measures would be beneficial in terms of further understanding perception of discontinuity and as a guide to develop more advance join costs for unit selection. Results in this study indicate that wavelets are advantageous when spectral dynamics are present, for example in diphthongs, although there is not sufficient evidence to say that this is generally the case. Such a study could investigate measurable signal characteristics, for instance spectral dynamics, to determine if a wavelet based measure or a standard spectral measure would be more suitable in a particular instance. Such a study would benefit from a larger more varied test database.

8.1.3 Neural networks

There are many open questions regarding the application of neural networks as a measure of spectral continuity. The results suggest that this approach can offer significant advantages over existing measures and a number of challenges exist to determine the limits and the applicability of this technique in general. A study with multiple speakers and a large and more varied dataset would be highly beneficial and could provide answers to some of the following questions with respect to neural networks.

- Can the neural network based approach provide similar gains in performance for different speakers?
- Will the neural network based approach provide similar gains for joins in consonants?
- How will the neural network extrapolate when used on sounds not represented in the training set?
- What is the appropriate and simplest strategy to train a neural network that is to be used in a unit selection synthesiser?
- What is the minimum amount of training data required to train a neural to reliably detect discontinuities?

- Is it possible to train a neural network from multiple speakers that will generalise to all speakers and not require training for individual speakers?

Successfully incorporating this approach into a unit selection system also presents a number of challenges and would benefit from further research. Specifically, demonstrating the effect of using the neural network based approach on the quality of speech generated by a unit selection synthesiser. Similar studies could be conducted to determine the potential of the neural network based measures to improve database pruning and enrichment, unit boundary training and optimal coupling of units.

References

Allen, J. B. and Rabiner, L. R. (1977). A unified theory of short-time spectrum analysis and synthesis. *Proc. IEEE*, 65:1158–1564.

Almeida, L. B. and Tribolet, J. M. (1983). Non-stationary modelling of voiced speech. *IEEE Trans. on Acoustics Speech and Signal Processing*, 31:664–678.

Alsteris, L. D. and Paliwal, K. K. (2006). Short-time phase spectrum in speech processing: A review and some experimental results. *Speech Communication*.

Alvarez, Y. V. and Huckvale, M. (2002). The reliability of the ITU-T P.85 standard for the evalustion of text-to-speech systems. In *Proc. ICSLP*, Denver, Colorado, USA.

Ashmore, J. and Gale, J. (2004). The cochlear amplifier. *Current Biology*, 14:403–404.

Bailly, G. (1999). Accurate estimation of sinusoidal parameters in an harmonic+noise model for speech synthesis. In *Proc. Eurospeech*, Budapest, Hungary.

Bellegarda, J. (2004). A novel discontinuity metric for unit selection speech text-to-speech synthesis. In *Proc. of the 5th ISCA speech synthesis workship*, Pittsburgh, USA.

Bellegarda, J. (2006). A global, boundary-centric framework for unit selection text-to-speech synthesis. *IEEE Trans. on Audio, Speech and Language Processing*, 14:990 – 997.

Bellegarda, J. (2007). Globally optimal training of unit boundaries in unit selection text-to-speech synthesis. *IEEE Trans. on Audio, Speech and Language Processing*, 15:957–965.

Bennett, C. L. (2005). Large scale evaluation of corpus-based synthesizers: Results and lessons from the blizzard challenge 2005. In *Proc. Interspeech*, Lisbon, Portugal.

Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., and Syrdal, A. (1999). The at&t next-gen ttssystem. In *Proc. of joint meeting of ASA, EAA, and DAGA.*, Berlin, Germany.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford.

Bjørkan, I., Svendsen, T., and Farner, S. (2005). Comparingspectral distance measures for join cost optimization in concatenative speech synthesis. In *Proc. Eurospeech*.

Black, A. and Taylor, P. . . (1997). Festival speech synthesis system: system documentation (1.1.1). Technical report, Human Communication Research Centre Technical Report HCRC/TR-83.

Black, A. and Taylor, P. (1994). CHATR: A generic speech synthesis system. In *Proc. COLING*, pages 983–986.

Black, A. W. and Tokuda, K. (2005). The blizzard challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. In *Proc. Interspeech*, Lisbon, Portugal.

Boersma, P. and Weenink, D. (2005). *Praat: doing phonetics by computer (Version 4.3.12)*. Institute of Phonetics Sciences of the University of Amsterdam. Retrieved May 19, 2005, from http://www.praat.org/.

Bozkurt, B., Doval, B., D'Alessandro, C., and Dutoit, T. (2004). Appropriate windowing for group delay analysis and roots of z-transform of speech signals. In *Proc. EUSIPCO*, Vienna, Austria.

Breen, A. (2000). Issues in the development of the next generation of concatenative speech synthesis systems. In *IEE seminar on State of the art in speech synthesis*.

Breen, A. P. and Jackson, P. (1998). Non-uniform unit selection and the similarity metric within BTs Laureate TTS system. In *Proc. SSW-3*, Jenolan Caves, Blue Mountains, Australia.

Breen, A. P. and Jackson, P. (1999). Issues in the design of an advanced unit selection method for natural sounding concatenative synthesis. In *Proc. 13 International Meeting of the Acoustical Society of America*, Berlin, Germany.

Breen, A.P. Page, J. (1994). Designing the next generation of text-to-speech systems. In *IEE Colloquium onTechniques for Speech Processing and their Application*, London, UK.

Bregman, A. S. (1990). *Auditory scene analysis, the percetual oganization of sound*. The MIT press.

Campbell, N. (1996). CHATR: A high definition speech resequencing system. In *Proc. of 3rd ASA/ASJ Joint Meeting*, pages 1223–1228, Hawaii.

Chappell, D. T. and Hansen, J. H. (2002). A comparison of spectral smoothing methods for segment concatenation based speech synthesis. *Speech Communication*, 36:343–374.

Chen, J. D. and Campbell, N. (1999). Objective distance measures for assessing concatenative speech synthesis. In *Proc. Eurospeech*, Budapest, Hungary.

Clark, R., Richmond, K., and King, S. (2004). Festival 2 - build your own general purpose unit selection speech synthesiser. In *Proc. 5th ISCA workshop on speech synthesis*, Pittsburgh, Pennsylvania, USA.

Conkie, A. and Isard, S. (1997). *Optimal coupling of diphones*, chapter 23, pages 293–304. Springer, NewYork.

Crowe, A. and Jack, M. (1987). Globally optimising formant tracker using generalised centroids. *Electronic Letters*, 23:10191020.

Dau, T. and Puschel, D. (1996). A qualitative model of the'effective' signal processing in the auditory system. *J. Acoust. Soc. Am.*, 99:3615–3622.

Daubechies, I. (1992). *10 Lectures on Wavelets*. SIAM, Philadelphia, PA.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representation for monosyllable word recognition in continuously spoken sentences. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 28:357–366.

de Cheveigné, A. and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111:1917–1930.

Delgutte, B. (2002). *The Handbook of Phonetic Sciences*, chapter Auditory Neural Processing, pages 507–538. Oxford.

Donovan, R. E. (2001). A new distance measure for costing spectral discontinuities in concatenative speech synthesisers. In *Proc. of the 4th ISCA workshop on speech synthesis*, Pitlochry, UK.

Donovan, R. E. and Eide, E. (1998). The IBM trainable speech synthesis system. In *Proc. ICSLP*, Sydney, Australia.

Duda, R. and Hart, R. E. (2001). *Pattern Classification*. John Wiley and Sons, 2 edition.

Dutoit, T. (2001). *An introduction to text-to-speech synthesis*. Kluwer acacdemic publishers.

Flanagan, J. L. (1960). Models for approximating basilar membrane displacement. *Bell systems technology journal*, 39:1163–1191.

Flanagan, J. L. (1972). *Speech analysis, synthesis and perception*. Springer-Verlag.

Forney, G. D. (1973). The Viterbi algorithm. *Proc. of the IEEE*, 61:268–278.

Founda, M., Tambouratzis, G., Chalamandaris, A., and Carayannis, G. (2001). Reducing spectral mismatches in concatenative speech synthesisvia systematic database enrichment. In *Proc. Eurospeech*, Aalborg, Denmark.

Furui, S. (1986a). On the role of spectral transitions for speech perception. *J. Acoust. Soc. Am.*, 80:1016–1025.

Furui, S. (1986b). Speaker independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. on acoust, speech and signal processing*, 34:52–59.

Furui, S. (1990). On the use of hierarchical spectral dynamics in speech recognition. In *Proc. ICASSP*, Albuquerque, New Mexico , USA.

Gobl, C. and NíChasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40:189–212.

Gold, B. and Morgan, N. (2000). *Speech and audio signal processing: processing and perception of speech and music*. John Wiley and Sons.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations (Third Edition)*. The John Hopkins University Press.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36.

Hayes, M. H. (1996). *Statistical digital signal signal processing and modelling*. John Wiley and Sons, Inc.

Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation*. Macmillan.

Hermansky, H. (1990). Perceptual Linear Prediction (PLP) analysis of speech. *J. Acoust. Soc. Am.*, 87:1738–1752.

House, A. S., Williams, C., Hecker, M. H. L., and Kryter, K. D. (1963). Psychoacoustic speech tests: A modified rhyme test (A). *J. Acoust. Soc. Am.*, 35:1899–1899.

Huang, X., Acero, A., and Hon, H. W. (2001). *Spoken language processing. A guide to theory, algorithm and system developement*. Prentice Hall.

Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. ICASSP*, Atlanta, Georgia, USA.

Itakura, F. (1975). Line spectrum representation of linear predictive coefficients of speech signals. *J. Acoust. Soc. Am.*, 57.

Jesteadt, W., Bacon, S., and Lehman, J. (1982). Forward masking as a function of frequency, masker level, and signal delay. *J. Acoust. Soc. Amer.*, 71:950962.

Jolliffe, I. (1986). *Principal Component Analysis*. Springer-Verlag.

Karaiskos, V., King, S., Clark, R. A. J., and Mayo, C. (2008). The blizzard challenge 2008. In *Proc. of the Blizzard Challenge 2008*, Brisbane, Australia.

Kim, D.-S. (2001). On the perceptually irrelevant phase information in sinusoidal representation of speech. *IEEE Trans. on speech and audio processing*, 9(8):900–905.

King, S. and Karaiskos, V. (2009). The blizzard challenge 2009. In *Proc. of the Blizzard Challenge 2009*, University of Edinburgh, Scotland.

Kirkpatrick, B., O'Brien, D., and Scaife, R. (2006). A comparison of spectral continuity measures as a join cost in concatenative speech synthesis. In *Proc. of the IET Irish Signals and Systems Conference (ISSC)*, Dublin, Ireland.

Klabbers, E., van Santen, J. P. H., and Kain, A. (2007). The contribution of various sources of spectral mismatch to audible discontinuities in a diphone database. *IEEE Trans. on audio speech and language processing*, 15:949 – 956.

Klabbers, E. and Veldhuis, R. (1998). On the reduction of concatenation artefacts in diphone synthesis. In *Proc. ICSLP*, volume 6, Sydney, Australia.

Klabbers, E. and Veldhuis, R. (2001). Reducing audible spectral discontinuities. *IEEE Trans. on speech and audio processing*, 9:39 – 51.

Klabbers, E. and Veldhuis, R. (2003). On the computation of the Kullback-Leibler distance measure for spectral distances. *IEEE Trans. on speech and audio processing*, 11:100 – 103.

Kluender, K. R., Coady, J. A., and Kiefte, M. (2003). Sensitivity to change in the perception of speech. *Speech Communication*, 41:59–69.

Knagenheim, H. P. and Kleijn, W. B. (1995). Spectral dynamics are more important than spectral distortion. In *Proc. ICASSP*, Detroit, Michigan, USA.

Kraus, N., McGee, T., Sharma, A., Carrell, T., and Nicol, T. (1992). Mismatch negativity event-related potential elicited by speech stimuli. *Ear and Hearing*, 13:158–64.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Ann. Math. Statist.*, 6:79–86.

Lindblom, B. E. F. and Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *J. Acoust. Soc. Am.*, 42:830–843.

Lindemann, E. and Kates, J. (1999). Phase relationships and amplitude envelopes in auditory perception. In *Proc. of the IEEE workshop on applications of signal processing to audio and acoustics*, New Paltz, New York.

Ling, Z. H., Lu, H., Hu, G. P., and Wang, L. R. D. R. H. (2008). The USTC sytem for Blizzard challenge 2008. In *Proc. of the Blizzard Challenge 2008*, Brisbane, Australia.

Ling, Z.-H., Richmond, K., Yamagishi, J., and Wang, R.-H. (2009). Integrating articulatory features into HMM-based parametric speech synthesis. *IEEE Trans. on Audio, Speech and Language Processing*, 17:1171–1185.

Lu, H., Ling, Z.-H., Lei, M., Wang, C.-C., Zhao, H.-H., Chen, L.-H., Hu, Y., Dai, L.-R., and Wang, R.-H. (2009). The ustc system for blizzard challenge 2009. In *Proc. of the Blizzard Challenge 2009*, University of Edinburgh, Scotland.

Lyon, R. F. and Mead, C. (1988). An analog electronic cochlea. *IEEE Trans. on Acoustics,Speech and Signal Processing*, 36:1119–1134.

Mallat, S. (1998). *A Wavelet Tour of Signal Processing*. Academic Press.

Markel, J. D. and Gray, A. H. (1976). *Linear Prediction of Speech*. Springer-Verlag, Berlin.

May, P., Tiitinen, H., Illmoniemi, R. J., Nyman, G., Taylor, J. G., and Naatanen, R. (1999). Frequency change detection in the human auditory cortex. *Journal of computational neuroscience*, 6:99–120.

McAulay, R. J. and Quatieri, T. F. (1986). Speech analysis/synthesis based on a sinusoidal represenation. *IEEE Trans. on speech and audio processing*, 34:744–754.

McAulay, R. J. and Quatieri, T. F. (1992). *Advances in speech processing*, chapter 6. Low rate speech coding based on the sinusoidal model. NewYork: Marcel Dekker.

McAulay, R. J. and Quatieri, T. F. (1995). *'Sinusoidal Coding', Speech Coding and Synthesis*, chapter 4. Elsevier.

Meddis, R. (1999). Computer models of the auditory periphery. *J. Acoust. Soc. Am.*, 105:963.

Meddis, R., Hewitt, M., and Shackleton, T. (1990). Implementation details of a computational model of the inner hair-cell/auditory-auditory nerve synapse. *J. Acoust. Soc. Am.*, 87:1813 – 1816.

Mendel, J. M. (1991). Tutorial on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications. *Proc. IEEE*, 79:278–305.

Moon, T. K. and Stirling, W. K. (2000). *Mathematical Methods and Algorithms for Signal Processing*. Prentice Hall.

Moore, B. C. J. (2004). *An introduction to the psychology of hearing*. Elsevier, 5 edition.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453467.

Murthy, H. A. and Gadde, V. R. R. (2003). The modified group delay function and its application to phoneme recognition. In *Proc. ICASSP*, Hong Kong.

Naatanen, R. (1995). The mismatch negativity: a powerful tool for cognitive neuroscience. *Ear Hear*, 16:6–18.

Ohm, G. S. (1843). Uber die definition des tones, nebst daran geknupfter theorie der sirene und ahnlicher tonbildender vorrichtungen. *Ann. Phys. Chem.*, 59:513–565.

Oppenheim, A. V. and Schafer, R. W. (1975). *Digital Siganl Processing*. Prentice Hall, Englewood Cliffs, NJ.

Page, J. H. and Breen, A. P. (1998). The Laureate text to speech system architecture and applications. *Speech Technology for Telecommunications*.

Paliwal, K. and Alsteris, L. (2005). On the usefulness of STFT phase spectrum in human listening tests. *Speech Communication*, 45:153170.

Pantazis, Y. and Stylianou, Y. (2007). On the detection of discontinuities in concatenative speech synthesis. *SpringerLecture Notes on Computer Science*, pages 89–100.

Pantazis, Y., Stylianou, Y., and Klabbers, E. (2005). Discontinuity detection in concatenated speech synthesis based on nonlinear speech analysis. In *Proc. Interspeech*, Lisbon, Portugal.

Patterson, D. (1996). *Artificial Neural Networks*. Prentice Hall.

Patterson, R. D. (1987). A pulse ribbon model of monaural phase sensitivity. *J. Acoust. Soc. Am.*, 82:1560–1586.

Patterson, R. D., Allerhand, M., and Giguere, C. (1995). Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform. *J. Acoust. Soc. Am.*, 98:1890–1894.

Pickles, A. (1988). *An introduction to auditory physiology*. Academic Press, 2nd edition.

Purwins, H., Blankertz, B., and Obermayer, K. (2000). Computing auditory perception. *Organised Sound*, 5(3):159–71.

Quatieri, T. F. (2002). *Discrete-time speech signal processing - principles and practice*. Prentice Hall.

Quatieri, T. F. and Dunn, R. B. (2002). Speech enhancement based on auditory spectral change. In *Proc. ICASSP*, Orlando, Florida, USA.

Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of speech recognition*. PTR Prentice Hall.

Sagisaka, Y., Kaiki, N., and Iwahashi, N. (1992). ATR - v-talk speech synthesis system. In *Proc. ICSLP*, Alberta, Canada.

Salimpour, Y. and Abolhassani, M. D. (2006). Auditory wavelet transform based on auditory wavelet families. In *Proc. 28th IEEE EMBS International Conference*, New York, USA.

Schroeder, M. (1981). Direct (nonrecursive) relations between cepstrum and predictor coefficients. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 2:297–301.

Shi, G., Shannechi, M. M., and Aarabi, P. (2005). On the importance of phase in human speech recognition. *IEEE Trans. on audio, speech and language processing.*

Skoglund, J. and Kleijn, B. (2000). On time-frequency masking in voiced speech. *IEEE Trans. on speech and audio processing*, 8(4).

Smith, R. L. (1977). Short-term adaptation in single auditory nerve fibers. *J. Neurophys.*, 40:1098–1111.

Somervuo, P. (2003). Experiments with linear and nonlinear feature transformtions in hmm based phone recognition. In *Proc. ICASSP*, Hong Kong.

Somervuo, P., Chen, B., and Zhu, Q. (2003). Feature transformations and combinations for improving ASR performance. In *Proc. EUROSPEECH*, Geneva, Switzerland.

Soong, F. and Juang, B.-H. (1984). Line spectrum pair (LSP) and speech data compression. In *Proc. IEEE ICASSP*, San Diego, USA.

Soong, F. and Rosenberg, A. (1988). On the use of instantaneous and transitional spectral information in specaker recognition. *IEEE Trans. on Audio, Speech and Signal Processing*, 36:871–879.

Specht, D. F. (1990). Probabilistic neural networks and the polynomial adaline as complementary techniques for classification. *IEEE Trans. Neural Networks*, 1:111–121.

Specht, D. F. (1991). A general regression neural network. *IEEE Trans. on Neural Networks*, 2.

Strope, B. and Alwan, A. (1997). A model of dynamic auditory perception and its application to robust word recognition. *IEEE Trans. Speech Audio Processing*, 5:451–464.

Stylianou, Y. (1999). Assessment and correction of voice quality variabilities in large speech databases for concatenative speech synthesis. In *Proc. ICASSP*, Phoenix, Arizona, USA.

Stylianou, Y. (2001). Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. on Speech and Audio Processing*, 9:21 – 29.

Stylianou, Y. and Syrdal, A. K. (2001). Perceptual and objective detection of discontinuities in concatenative speech synthesis. In *Proc. ICASSP*, Salt Lake City, USA.

Summerfield, Q., Sidwell, A., and Nelson, T. (1986). Auditory enhancement of changes in spectral amplitude. *J. Acoust. Soc. Am.*, 81:700 – 708.

Syrdal, A. K. (2001). Phonetic effects on listener detection of vowel concatenation. In *Proc. EUROSPEECH*, Aalborg, Denmark.

Toda, T., Black, A., and Tokuda, K. (2005). Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In *Proc. ICASSP*, Philadelphia, USA.

Tokuda, K., Kobayashi, T., and Imai, S. (1995). Speech parameter generation from HMM using dynamic features. In *Proc. ICASSP*, Detriot, USA.

Tsuzaki, M. and Kawai, H. (2002). Feature extraction for unit selection in concatenative speech synthesis: Comparison between AIM, LPC and MFCC. In *Proc. ICSLP*, Denver, USA.

Vepa, J. and King, S. (2003). Kalman-filter based join cost for unit-selection speech synthesis. In *Proc. EUROSPEECH*, Geneva, Switzerland.

Vepa, J. and King, S. (2004a). Subjective evaluation of join cost functions used in unit selection speech synthesis. In *Proc. ICSLP*, Jeju, Korea.

Vepa, J. and King, S. (2004b). *Text to Speech Synthesis: New Paradigms and Advances*, chapter 3 - Join cost for unit selection speech synthesis. Prentice Hall.

Vepa, J. and King, S. (2006). Subjective evaluation of join cost and smoothing methods for unit selection speech synthesis. *IEEE Trans. on Audio, Speech and Language Processing*, 14:1763 – 1771.

Vepa, J., King, S., and Taylor, P. (2002). Objective distance measures for spectral discontinuities in concatenative speech synthesis. In *Proc. ICSLP*, Denver, USA.

von Helmholtz, H. L. F. (1875). On the sensations of tone. *(English Translation by A. J. Ellis, Longmans Green and Co. London, 1912).*

von Békésy, G. (1960). *Experiments in hearing.* McGraw-Hill.

Wasserman, P. D. (1993). *Advanced Methods in Neural Computing.* Van Nostrand Reinhold.

Wouters, J. and Macon, M. (1998). Perceptual evaluation of distance measures for concatenative speech synthesis. In *Proc. ICSLP*, volume 6, Sydney, Australia.

Yang, X., Wang, K., and Shamma, S. A. (1992). Auditory representations of acoustic signals. *IEEE Trans. on Information Theory*, 38:824–839.

Yegnanarayana, B., d'Alessandro, C., and Darsinos, V. (1998). An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Trans. on Speech and Audio Processing*, 6:1–11.

Yegnanarayana, B. and Murthy, H. A. (1992). Significance of group delay functions in spectrum estimation. *IEEE Trans. on Signal Processing*, 40:2281 – 2289.

Yost, W. A. (2007). *Fundamentals of hearing; an introduction.* Elsevier Academic Press, 5th edition.

Young, E. D. (2007). Neural representation of spectral and temporal information in speech. *Phil. Trans. R. Soc.*, 363:923–945.

Young, E. D. and Sachs, M. B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers. *J. Acoust. Soc. Am.*, 66:1381–1403.

Zen, H. and Toda, T. (2005). An overview of Nitech HMM-based speech synthesis system for Blizzard challenge 2005. In *Proc. Interspeech*, Lisbon, Portugal.

Zen, H., Tokuda, K., and Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*.